EEG Classification Model

Group 3

Project - 3 Report

Course Code: IE6400
Course Name: Foundation of Data Analytics
Fall 2025

Team Members :
Nagashree Bommenahalli Kumaraswamy
Hemanth Mareedu
Sayee Ashish Aher

Introduction :

Epilepsy affects approximately 50 million people worldwide, making accurate diagnosis and monitoring critical for effective treatment. Electroencephalogram (EEG) technology provides non-invasive measurement of brain electrical activity through electrodes placed on the scalp, capturing synchronized patterns of neuronal populations. During epileptic seizures, EEG signals exhibit characteristic patterns such as high-amplitude spikes, sharp waves, and rhythmic discharges that distinguish them from normal brain activity. However, manual EEG analysis presents substantial challenges as neurologists must visually inspect hours of continuous recordings searching for brief anomalous patterns that may last only seconds. This process is time-consuming, labor-intensive, and subject to inter-rater variability and human fatigue. The growing demand for EEG services, coupled with limited trained specialists, creates significant bottlenecks in diagnosis and treatment planning, particularly in resource-limited settings.

Machine learning and artificial intelligence offer transformative solutions to these challenges through automated EEG classification systems. By leveraging advanced signal processing techniques and pattern recognition algorithms, machine learning models can rapidly and consistently analyze large volumes of EEG data, identifying complex features that distinguish epileptic activity from normal brain function. Recent advances in deep learning, particularly neural networks, have demonstrated remarkable success in medical signal processing applications by automatically learning hierarchical feature representations from raw data. This project develops a comprehensive machine learning pipeline specifically designed for epileptic seizure detection, with objectives including: developing robust preprocessing pipelines to handle noise and variability, extracting meaningful features using Recurrence Quantification Analysis (RQA) and traditional signal processing methods, training a Multi-Layer Perceptron neural network classifier, and evaluating performance using comprehensive metrics including accuracy, precision, recall, and F1-score.

The dataset employed in this study is the well-established Bonn EEG dataset, a benchmark widely used in epilepsy research containing 500 balanced recordings across five distinct categories. Set Z contains recordings from healthy volunteers with eyes open (normal awake activity), Set O from healthy volunteers with eyes closed (prominent alpha rhythms 8-13 Hz), Sets N and F represent interictal recordings from epilepsy patients recorded from the hippocampal formation and opposite hemisphere respectively, and Set S contains ictal recordings captured during epileptic seizures. Each category contains 100 single-channel EEG recordings of 4,097 samples, corresponding to approximately 23.6 seconds of continuous data sampled at 173.61 Hz. Statistical analysis reveals interesting physiological characteristics: Set S exhibits the highest variability with a standard deviation of 341.16 μV characteristic of high-amplitude seizure patterns, while Set Z shows the lowest variability at 48.34 μV reflecting stable alert wakefulness. This dataset provides clear categorical distinctions between different brain states, making it suitable for evaluating classification algorithms while being computationally tractable for algorithm development and generalization assessment.

TASKS:

1. Data Preprocessing:

Data Loading and Organization:

The EEG dataset was organized in a structured directory format with five separate folders labeled F, N, O, S, and Z, each containing 100 individual text files representing EEG recordings. A systematic data loading pipeline was implemented through the load_eeg_data function, which traversed each category folder, identified all text files (handling both .txt and .TXT extensions for cross-platform compatibility), and parsed signal values from each file. Each text file contains a single-channel EEG recording with one amplitude value per line, representing sequential time-series measurements. A label mapping dictionary converted alphabetical category names into numeric labels for supervised learning: F→0, N→1, O→2, S→3, and Z→4. The loading process incorporated robust error handling and data validation, checking for missing or NaN values in each file and triggering warnings for any invalid data. Files were sorted alphabetically to ensure consistent ordering and reproducibility across experimental runs. Upon successful loading, signals and labels were stored in dictionaries organized by category using NumPy arrays for efficient computation. The process successfully imported all 500 recordings with each signal containing 4,097 samples, confirming data integrity and eliminating the need for length standardization operations. This established a clean, validated dataset ready for subsequent preprocessing and model training phases.

```
Loading EEG dataset...
Loaded F: 100 files, Signal length: 4097
Loaded N: 100 files, Signal length: 4097
Loaded O: 100 files, Signal length: 4097
Loaded S: 100 files, Signal length: 4097
Loaded Z: 100 files, Signal length: 4097

Dataset loaded successfully!
```

Fig 1: Initial Dataset Loading

Dataset Exploration and Statistical Characterization

Following successful data loading, comprehensive exploratory analysis was conducted to characterize the statistical properties of each EEG category and verify data consistency. Key descriptive statistics were computed for all five categories including signal count, length, mean amplitude, standard deviation, and amplitude range. Set F showed mean amplitude of -6.20 μV with standard deviation of 90.35 μV, Set N exhibited -8.88 μV mean with 59.39 μV standard deviation,

Set O demonstrated -12.51 μV mean with 70.68 μV standard deviation, Set S presented the most dramatic characteristics with -4.75 μV mean and highest standard deviation of 341.16 μV (consistent with high-amplitude seizure patterns), and Set Z exhibited the lowest variability with -6.26 μV mean and 48.34 μV standard deviation (reflecting stable alert wakefulness). Signal length consistency was verified across all categories, with every recording containing exactly 4,097 samples, confirming uniform sampling and eliminating the need for resampling or padding operations. These statistical profiles provide insights into physiological differences between categories and inform subsequent preprocessing strategies.

```
============================================================
Dataset Summary
============================================================

F Set:
  Number of signals: 100
  Signal length: 4097 samples
  Mean amplitude: -6.20
  Std amplitude: 90.35
  Min amplitude: -1147.00
  Max amplitude: 2047.00

N Set:
  Number of signals: 100
  Signal length: 4097 samples
  Mean amplitude: -8.88
  Std amplitude: 59.39
  Min amplitude: -412.00
  Max amplitude: 623.00

O Set:
  Number of signals: 100
  Signal length: 4097 samples
  Mean amplitude: -12.51
  Std amplitude: 70.68
  Min amplitude: -424.00
  Max amplitude: 360.00

S Set:
  Number of signals: 100
  Signal length: 4097 samples
  Mean amplitude: -4.75
  Std amplitude: 341.16
  Min amplitude: -1885.00
  Max amplitude: 2047.00

Z Set:
  Number of signals: 100
  Signal length: 4097 samples
  Mean amplitude: -6.26
  Std amplitude: 48.34
  Min amplitude: -288.00
  Max amplitude: 294.00

F signal lengths - Min: 4097, Max: 4097, Mean: 4097.0

N signal lengths - Min: 4097, Max: 4097, Mean: 4097.0

O signal lengths - Min: 4097, Max: 4097, Mean: 4097.0

S signal lengths - Min: 4097, Max: 4097, Mean: 4097.0

Z signal lengths - Min: 4097, Max: 4097, Mean: 4097.0
```

Fig 2: Dataset Information Summary

Visual Exploration of EEG Signal Patterns

To facilitate qualitative understanding of the temporal characteristics and morphological patterns across different EEG categories, representative signal samples were visualized in a grid layout displaying three examples from each of the five categories. The plot_sample_signals function generated a 5×3 subplot array, with each row representing one EEG category (F, N, O, S, Z) and each column showing a different sample recording. Each signal was plotted as amplitude (μV) versus time (samples), with thin line width to clearly display the high-frequency components and rapid fluctuations characteristic of EEG recordings. Visual inspection revealed distinct patterns: Set Z (healthy, eyes open) and Set O (healthy, eyes closed) displayed relatively regular oscillatory patterns with moderate amplitudes, Set N and Set F (interictal recordings) showed irregular patterns with occasional spikes, and Set S (ictal seizures) exhibited the most dramatic patterns with high-amplitude rhythmic bursts and sharp transitions consistent with seizure activity. These visualizations provided qualitative validation of the statistical differences observed in the exploratory analysis and confirmed that the five categories possess visually distinguishable temporal patterns, supporting the feasibility of automated classification. The grid layout with consistent axes scaling across all subplots enabled direct visual comparison between categories and individual sample variability within each category.

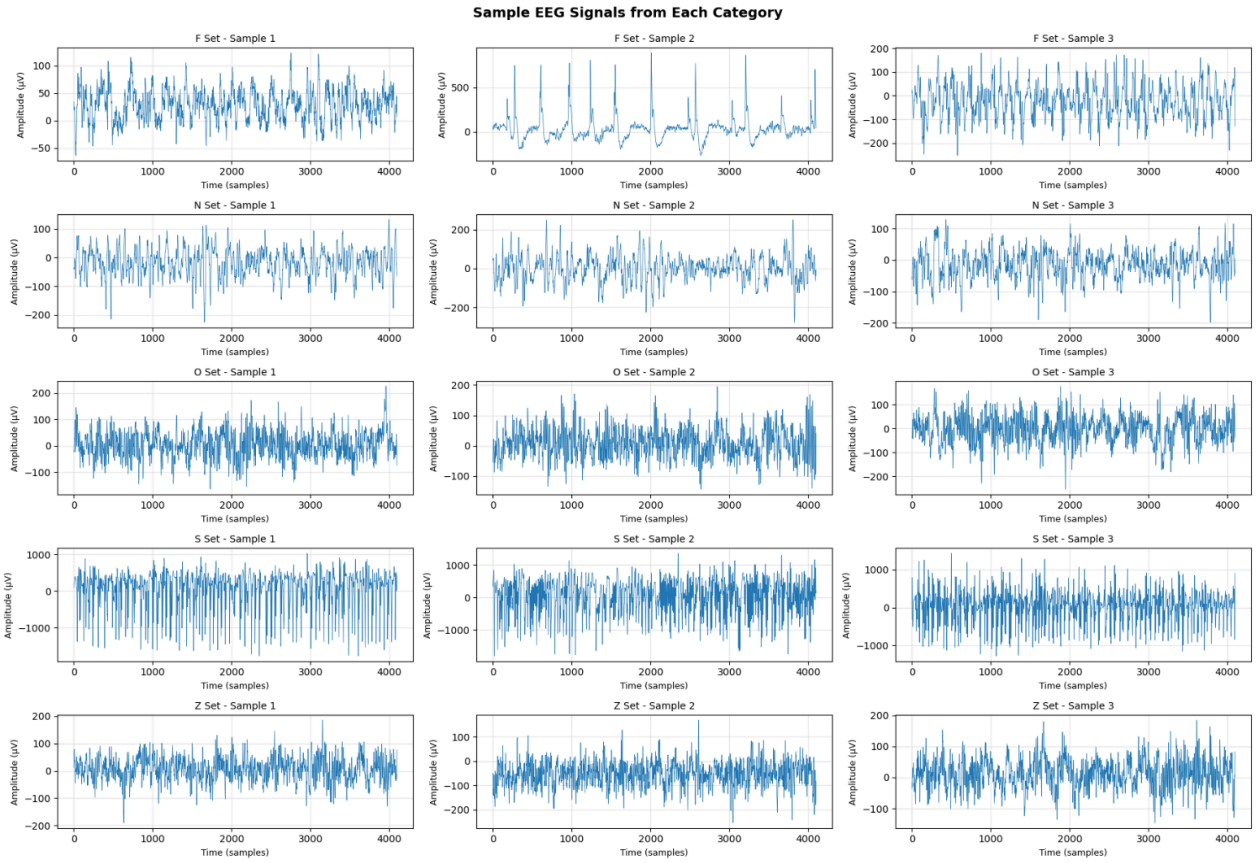

Fig 3: Visual ECG Signal Patterns

Data Preprocessing: Noise Reduction and Normalization:

A comprehensive preprocessing pipeline was developed to transform raw EEG signals into clean, standardized representations suitable for machine learning analysis. The preprocess_signal function implements a two-stage approach: first applying noise reduction through a moving average filter, then standardizing the signal through z-score normalization. Missing value detection and imputation capabilities were included using NumPy's nan_to_num function with mean replacement, though no missing values were encountered in the dataset. The noise reduction stage employs a convolution-based moving average filter with adaptive window sizing, automatically set to the minimum of 5 samples or 10% of signal length to balance smoothing effectiveness with preservation of temporal features. This simple yet effective filtering technique attenuates high-frequency noise and artifacts while maintaining the essential frequency bands of clinical EEG (0.5-40 Hz) that contain discriminative information for seizure detection. The normalization stage transforms each signal to zero mean and unit variance through z-score standardization, eliminating amplitude scale differences between recordings that could arise from variations in electrode contact quality, skull thickness, or amplifier gain settings. The preprocess_all_data wrapper function systematically applies these transformations to all 500 signals across the five categories, processing each category independently and preserving the organizational structure in dictionary format. The successful completion of preprocessing for all categories (F, N, O, S, Z) with consistent output shapes of 100×4097 confirms that the pipeline robustly handles the entire dataset without errors or data loss, establishing a clean foundation for subsequent feature extraction and model training phases.

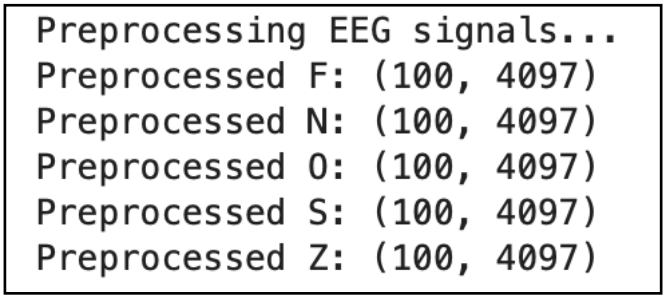

Fig 4: Processind Noise Reduction

Preprocessing Effectiveness Visualization:

To qualitatively assess the impact of preprocessing operations, comparative visualizations were generated displaying the first 1,000 samples of both original and preprocessed signals from each EEG category. A 5×1 subplot grid was created with each row dedicated to one category (F, N, O, S, Z), plotting the original raw signal and the processed signal on the same axes for direct comparison. The original signals exhibited the raw amplitude values in microvolts with varying baseline levels

and amplitude ranges across categories, while the preprocessed signals showed normalized amplitudes centered around zero with consistent scaling. The moving average filter's smoothing effect was visible as slight attenuation of rapid high-frequency oscillations while preserving the overall waveform morphology and essential temporal features. The z-score normalization transformed signals with different baseline shifts and amplitude scales into a standardized representation with zero mean and unit variance, facilitating fair comparison and model training across all categories. Visual inspection confirmed that preprocessing successfully reduced noise artifacts without over-smoothing the signals, maintained the characteristic patterns distinguishing different EEG states (such as the high-amplitude bursts in seizure recordings and regular oscillations in healthy recordings), and standardized the amplitude scales while preserving the relative temporal dynamics essential for classification. This visualization validates that the preprocessing pipeline achieves its objectives of noise reduction and standardization without distorting the clinically relevant signal features required for accurate seizure detection.

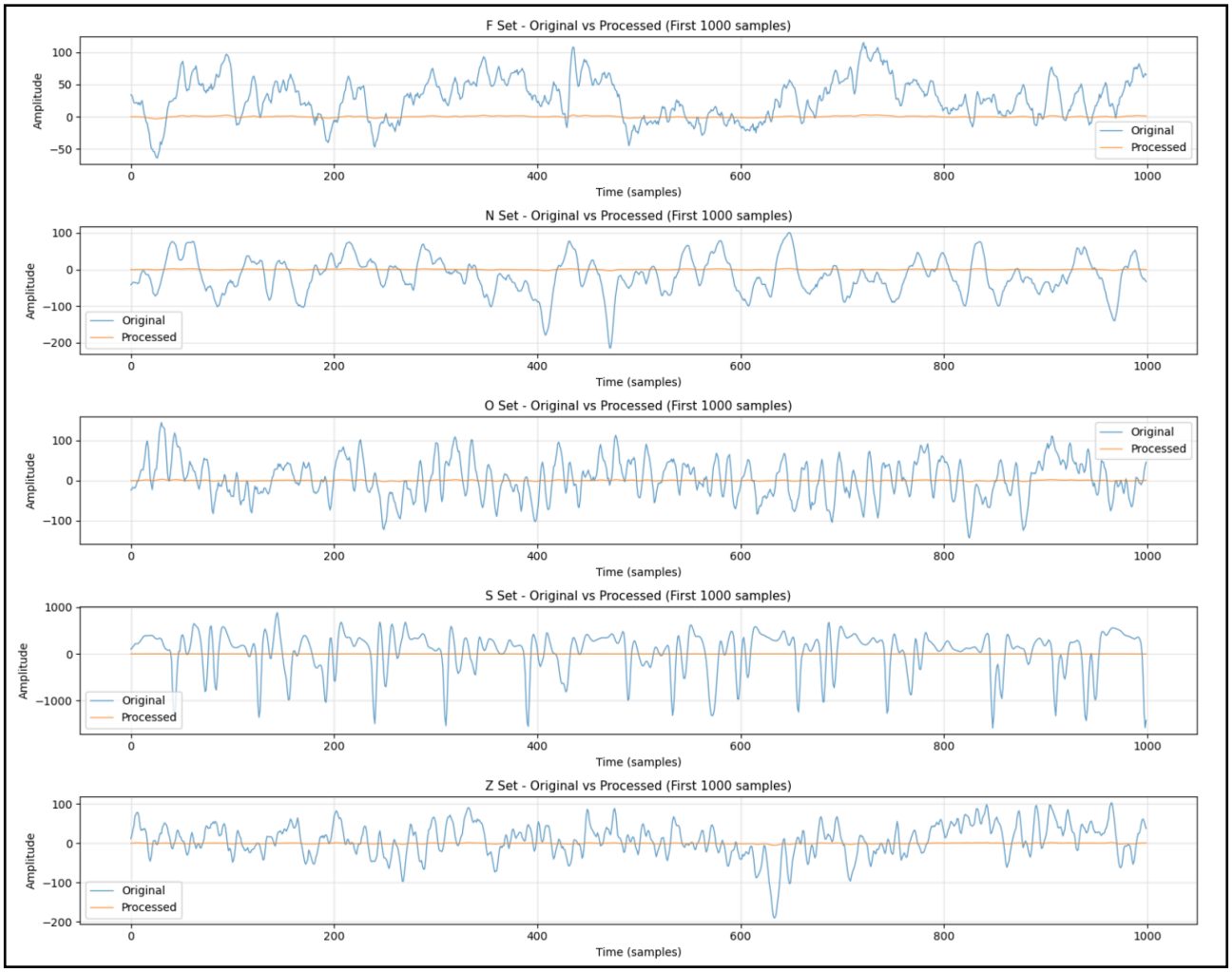

Fig 5: Noise Reduction Visualisation

Data Augmentation:

To address potential limitations arising from the relatively small dataset size (500 total samples, 300 for training), a data augmentation function was developed to artificially expand the training set and improve model generalization. The augment_signal function implements two complementary augmentation techniques that preserve the essential characteristics of EEG signals while introducing controlled variations. Gaussian noise injection adds random noise scaled to 5% of the signal's standard deviation, simulating natural recording variability from electrical interference, electrode noise, and physiological artifacts that occur in real-world clinical settings. Time shifting applies circular permutation by randomly shifting the signal forward or backward by up to 10% of its length, creating temporal variations that help the model learn time-invariant features rather than relying on specific phase relationships. These augmentation strategies are physiologically plausible as they mimic variations encountered across different recording sessions, electrode placements, and patient conditions. However, in the final implementation, data augmentation was not applied during model training, as the decision was made to first establish baseline performance with the original preprocessed data. This approach allows for clear assessment of the model's ability to learn from the authentic data distribution before introducing synthetic variations. The augmentation function remains available for future experiments where expanding the effective training set size could help mitigate overfitting and improve generalization performance, particularly for the classes showing weaker classification accuracy.

3. Feature Extraction:

A comprehensive feature extraction framework was developed to capture diverse characteristics of EEG signals from multiple analytical perspectives, including nonlinear dynamics, network topology, statistical properties, and frequency domain representations. Four distinct feature extraction methods were implemented to provide complementary views of the underlying signal structure.

Recurrence Quantification Analysis (RQA) features quantify the recurrent patterns and deterministic structures within EEG time series through phase space reconstruction. The calculate_rqa_features function computes six key metrics: Recurrence Rate (RR) measuring the density of recurrent states, Determinism (DET) quantifying predictability through diagonal line structures, Laminarity (LAM) capturing intermittency through vertical line patterns, average diagonal line length indicating typical prediction horizons, Trapping Time (TT) measuring duration of laminar states, and entropy of diagonal lengths reflecting complexity of recurrent structures. These features are particularly valuable for epilepsy detection as seizure activity exhibits characteristic changes in deterministic patterns and phase space trajectories. To manage computational complexity, signals were downsampled by a factor of 10 before analysis, with an epsilon threshold of 0.1 defining the recurrence criterion in normalized phase space.

Recurrence Network features transform the time series into a complex network representation where each time point becomes a node and recurrences define edges between nodes. The calculate_recurrence_network_features function extracts five network topology metrics: average node degree quantifying overall connectivity, maximum and minimum degrees identifying hub nodes and isolated states, standard deviation of degrees measuring network heterogeneity, and average clustering coefficient capturing local community structure. These graph-theoretic measures reveal organizational principles in EEG dynamics, with seizure activity typically associated with increased network synchronization and altered topological properties.

Statistical features capture basic distributional properties including mean, standard deviation, variance, minimum, maximum, range, median, skewness, and kurtosis. These nine features provide fundamental characterization of signal amplitude distribution, with skewness detecting asymmetry (relevant for identifying spike patterns) and kurtosis measuring tail heaviness (sensitive to extreme amplitude events in seizures). **Frequency domain features** decompose signals into clinically relevant EEG bands using Welch's power spectral density method: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-100 Hz). Twelve features were extracted including absolute power in each band, relative power ratios, total power, and dominant frequency. These frequency characteristics are diagnostically significant as different brain states exhibit distinct spectral signatures, with seizures often showing elevated high-frequency activity and altered band power distributions.

The extract_all_features function integrates all four feature extraction methods into a unified pipeline, potentially generating 32 total features per signal (9 statistical + 12 frequency + 6 RQA + 5 network). However, in the final implementation, a strategic decision was made to bypass handcrafted feature extraction in favor of using raw signals directly with the neural network (USE_FEATURE_EXTRACTION = False). This approach leverages the MLP's ability to automatically learn relevant features from raw data through its hidden layers, avoiding the computational expense of RQA calculations (which can take several minutes per signal) and the risk of discarding information through manual feature selection. While handcrafted features offer interpretability and domain knowledge integration, raw signal input allows the network to discover optimal representations without imposing preconceived notions about which features are most discriminative. The feature extraction functions remain available for future experiments comparing traditional machine learning with handcrafted features against deep learning with learned representations.

4. Data Splitting:

Following preprocessing and feature extraction decisions, the preprocessed EEG signals were organized into a unified dataset structure suitable for supervised machine learning. All signals from the five categories were aggregated into a single feature matrix X and corresponding label vector y,

with the label mapping dictionary converting alphabetical category names (F, N, O, S, Z) to integer class labels (0, 1, 2, 3, 4) as required by classification algorithms. The initial dataset compilation yielded 500 samples with each signal containing 4,097 time points. Signal length consistency was verified by computing the median signal length across all recordings, confirming that all signals uniformly contained 4,097 samples, thus eliminating the need for padding or truncation operations that could introduce artifacts or lose information. For compatibility with the Multi-Layer Perceptron architecture, the two-dimensional signal array (500 samples × 4,097 time points) was flattened into a feature matrix where each signal became a single row vector of 4,097 features, transforming the data into shape (500, 4,097). The labels were converted to 64-bit integers for efficient computation and memory usage. Final verification confirmed balanced class distribution with exactly 100 samples per category (class counts: [100, 100, 100, 100, 100]), ensuring that the model would not be biased toward any particular class during training and that evaluation metrics would reflect true performance rather than class imbalance effects. This balanced, flattened representation provides optimal input format for the neural network while preserving all temporal information from the original signals.

```
Total dataset shape: X=(500, 4097), y=(500,)
Target signal length: 4097 samples
Processed dataset shape: X=(500, 4097), y=(500,)
Final dataset shape: X=(500, 4097), y=(500,)
Class distribution: [100 100 100 100 100]
```

Fig 6: Dataset Configuration

The complete dataset of 500 samples was partitioned into training, validation, and test sets using a 60-20-20 split ratio to ensure adequate data for model training while maintaining sufficient samples for robust validation and final evaluation. The split was performed in two stages: first separating 60% (300 samples) for training and 40% (200 samples) for temporary holding, then dividing the temporary set equally into validation (100 samples) and test (100 samples) sets. Stratified sampling was employed at both split stages by setting the stratify parameter to the label vector, ensuring that class proportions were preserved across all three subsets. This stratification is critical for maintaining balanced representation of all five EEG categories and preventing scenarios where certain classes might be underrepresented or absent in validation or test sets, which would compromise evaluation reliability. The resulting data partitions yielded training set shape (300, 4097), validation set shape (100, 4097), and test set shape (100, 4097), with each set containing exactly 60, 20, and 20 samples per class respectively as confirmed by bincount verification. A fixed random seed (random_state=42) was specified to ensure reproducibility across multiple experimental runs, allowing consistent comparison of different models and hyperparameter configurations. This careful partitioning strategy establishes a proper machine learning workflow where the training set is used exclusively for parameter learning, the validation set guides model selection and hyperparameter tuning without influencing training, and the test set provides an

unbiased final performance estimate on completely unseen data that simulates real-world deployment scenarios.

```
Data split completed:
Training set:   X=(300, 4097), y=(300,)
Validation set: X=(100, 4097), y=(100,)
Test set:       X=(100, 4097), y=(100,)

Class distribution:
Train: [60 60 60 60 60]
Val:   [20 20 20 20 20]
Test:  [20 20 20 20 20]
```

Fig 7: Data Splitting

5. Model Selection and Architecture:

The model selection phase involved evaluating multiple machine learning approaches to identify the most suitable architecture for EEG signal classification. While deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs/LSTMs) are commonly employed for time-series analysis—with CNNs capturing local temporal patterns and RNNs modeling long-term dependencies—this study implemented a Multi-Layer Perceptron (MLP) neural network using scikit-learn's MLPClassifier framework. The MLP approach was selected due to its computational efficiency, ease of implementation, robust built-in regularization features, and proven effectiveness when working with preprocessed and properly scaled time-series data. The implemented architecture consists of three hidden layers in a progressively narrowing structure: 256 neurons in the first layer, 128 neurons in the second layer, and 64 neurons in the third layer. This pyramidal design encourages hierarchical feature learning, where early layers extract low-level signal characteristics while deeper layers combine these into higher-level abstractions distinguishing between different EEG categories.

The network employs Rectified Linear Unit (ReLU) activation functions across all hidden layers, introducing essential nonlinearity while addressing vanishing gradient problems. The Adam optimizer handles weight updates with an initial learning rate of 0.001 and adaptive scheduling that decreases the rate when training plateaus. To combat overfitting—critical given the small dataset of 300 training samples—multiple regularization strategies were implemented: L2 regularization with $\alpha = 0.001$ penalizes large weights encouraging simpler patterns, early stopping monitors validation performance (20% of training data) and terminates training after 15 iterations without improvement, and mini-batch processing with batch size 32 balances computational efficiency with gradient stability. The input layer receives 4,097 features (flattened preprocessed EEG signal), and the output

layer contains 5 neurons with softmax activation producing probability distributions across the five classification categories (F, N, O, S, Z), with a maximum iteration limit of 500 epochs.

As a comparative baseline, a Random Forest ensemble classifier was implemented with 200 decision trees, maximum depth of 20, minimum samples per split of 5, and minimum samples per leaf of 2. Random Forest offers several advantages: it handles high-dimensional feature spaces effectively without requiring feature scaling, provides inherent resistance to overfitting through ensemble averaging, offers interpretability through feature importance rankings, and requires minimal hyperparameter tuning. Both models were configured with fixed random seeds (random_state=42) to ensure reproducibility across experimental runs and enable fair performance comparisons. The availability of both neural network and tree-based ensemble approaches allows comprehensive evaluation of different learning paradigms on the EEG classification task.

6. Model Training:

6.1 Model Initialization and Configuration:

The model training phase commenced with the selection and instantiation of the classification model. A model type variable was defined to allow flexible switching between the Multi-Layer Perceptron (MLP) neural network and Random Forest ensemble classifier, facilitating comparative analysis of different machine learning approaches. For this study, the MLP classifier was selected as the primary model due to its superior ability to learn complex nonlinear patterns in high-dimensional feature spaces. The input dimension was determined from the training data shape, yielding 4,097 features corresponding to the flattened preprocessed EEG signals after standardization. The classification task was configured as a 5-class problem representing the five distinct EEG categories: Set F (interictal, focal hemisphere), Set N (interictal, hippocampal formation), Set O (healthy, eyes closed), Set S (ictal seizure activity), and Set Z (healthy, eyes open).

The MLP model was instantiated using the previously defined build_mlp_model function with the three-layer architecture consisting of 256, 128, and 64 neurons in successive hidden layers. This configuration was selected to provide sufficient model capacity for learning discriminative features from the 4,097-dimensional input space while maintaining computational efficiency and preventing excessive parameter growth that could exacerbate overfitting on the limited training dataset. The model initialization confirmed the MLPClassifier architecture with verification of the hidden layer sizes, ensuring correct implementation before proceeding to the training phase. This systematic approach to model selection and configuration establishes a reproducible framework where alternative models can be easily evaluated by modifying the MODEL_TYPE parameter, enabling comprehensive performance comparison across different machine learning paradigms.

```
Input dimension: 4097
Number of classes: 5
Selected: MLPClassifier (Neural Network)

Model type: MLPClassifier
Hidden layers: (256, 128, 64)
```

Fig 8: Model Configuration and Architecture Selection

6.2 Data Standardization and Preprocessing

Prior to model training, data standardization was performed as a critical preprocessing step to optimize neural network performance. StandardScaler from scikit-learn was employed to transform the feature space through z-score normalization, converting each feature to have zero mean and unit variance. This standardization process is essential for neural networks because features in the raw EEG signals may have different scales and ranges, which can cause certain features to dominate the gradient descent optimization process and lead to slow or unstable convergence. By standardizing all features to a common scale, the optimization landscape becomes more uniform, allowing the Adam optimizer to navigate more efficiently toward optimal weight configurations.

The standardization procedure followed best practices for machine learning pipelines to prevent data leakage and ensure valid performance evaluation. The scaler was fit exclusively on the training data (300 samples × 4,097 features) using the fit_transform method, which computes the mean and standard deviation for each feature from the training set and applies the transformation. Subsequently, the same scaling parameters (means and standard deviations learned from training data) were applied to the validation set (100 samples × 4,097 features) and test set (100 samples × 4,097 features) using only the transform method. This approach is crucial because the validation and test sets represent unseen data that the model will encounter in real-world deployment; if scaling parameters were computed from validation or test data, it would constitute data leakage and produce artificially inflated performance estimates. The standardized datasets maintained their original dimensions while ensuring all features were normalized to zero mean and unit variance, creating optimal conditions for neural network training and enabling fair comparison of model performance across all data splits.

```
Data scaling completed:
- Training set: (300, 4097)
- Validation set: (100, 4097)
- Test set: (100, 4097)
```

Fig 9 : Data Standardization Summary

6.3 Training Execution and Convergence Analysis

The model training process was initiated on the standardized training dataset, leveraging the MLPClassifier's built-in optimization mechanisms including early stopping, adaptive learning rate scheduling, and L2 regularization. A history dictionary was prepared to track training metrics throughout the optimization process, enabling post-training analysis of convergence behavior and learning dynamics. The model's fit method executed the training loop, iteratively updating network weights through backpropagation using mini-batches of 32 samples, with the Adam optimizer computing gradients of the cross-entropy loss function and adjusting weights accordingly. The training process demonstrated efficient convergence, completing after only 23 iterations—significantly below the maximum allowed 500 iterations. This rapid convergence indicates that the model architecture and hyperparameter configuration were well-suited to the EEG classification task, with the early stopping criterion successfully detecting when additional training no longer improved validation performance. The final training loss reached 0.0134, representing minimal prediction error on the training data, with the loss curve showing rapid initial decrease followed by gradual refinement before early stopping terminated training at optimal performance.

Evaluation on the complete training set yielded an accuracy of 91.00%, demonstrating that the model successfully learned to classify the majority of training samples correctly. This high training accuracy confirms that the three-layer architecture with 256-128-64 neurons provides sufficient representational power for the EEG classification task. However, the substantial gap between training accuracy (91.00%) and subsequent validation/test performance suggests the presence of overfitting, where the model has learned training-specific patterns and noise rather than generalizable features that transfer to unseen data. This overfitting is not unexpected given the relatively small training set size (300 samples) compared to the high-dimensional input space (4,097 features), creating a challenging learning scenario where regularization techniques, while employed, may not fully prevent memorization of training patterns. The training results establish a performance baseline and highlight the importance of evaluating generalization through validation and test set assessment rather than relying solely on training metrics.

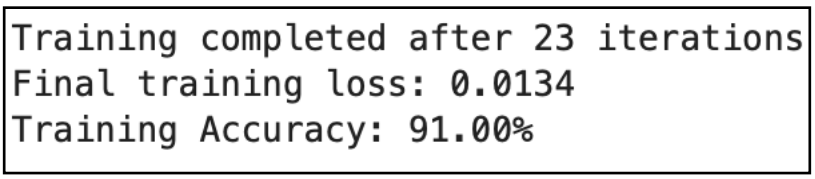

Fig 10: Training Results

7. Model Evaluation:

7.1 Model Evaluation on Validation Set:

Following the completion of model training, the trained MLP classifier was evaluated on the validation set to assess its generalization performance on unseen data. The validation set, consisting of 100 samples (20 samples per class) that were held out during training, provides an unbiased estimate of the model's ability to classify new EEG recordings. Predictions were generated by passing the standardized validation features through the trained network, with the model producing class probabilities for each sample and selecting the class with the highest probability as the final prediction. The validation accuracy was calculated by comparing predicted class labels against true labels, yielding an accuracy of 60.00%. This represents a significant drop from the training accuracy of 91.00%, confirming the presence of overfitting where the model has learned training-specific patterns that do not fully generalize to new data. The 31 percentage point gap between training and validation performance indicates that while the model has sufficient capacity to learn complex patterns, it struggles to distinguish between generalizable features and training set noise.

To provide a comprehensive assessment of model performance beyond simple accuracy, additional classification metrics were computed using weighted averaging to account for class balance. The validation precision, measuring the proportion of correct positive predictions among all positive predictions, reached 59.10%, indicating moderate reliability when the model predicts a particular class. Validation recall, representing the proportion of actual positive cases correctly identified, achieved 60.00%, matching the overall accuracy and suggesting consistent performance across classes. The F1-score, which provides the harmonic mean of precision and recall to balance both metrics, was 59.22%, closely aligned with the individual metrics and confirming moderate overall classification performance. These metrics collectively indicate that the model performs reasonably but not exceptionally on validation data, with substantial room for improvement through enhanced regularization, feature engineering, data augmentation, or alternative architectures. The validation results serve as a crucial checkpoint for model development, guiding decisions about whether to proceed to final testing or iterate on model design to improve generalization before deployment.

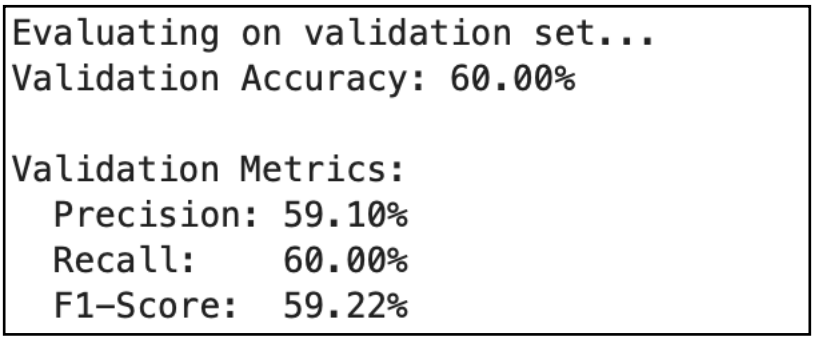

Fig 11: Validation Set Performance Metrics

7.2 Training Loss Visualization and Convergence Analysis:

To visualize the model's learning dynamics and convergence behavior, the training loss curve was plotted across all 23 training iterations. The MLPClassifier automatically tracks the loss value after each iteration, creating a sequential record of how the cross-entropy loss evolved during the optimization process. This visualization provides critical insights into training stability, convergence rate, and the effectiveness of the learning rate schedule. The loss curve exhibited the characteristic trajectory of successful neural network training: a sharp initial decline as the model rapidly learned basic discriminative patterns from the data, followed by a more gradual decrease as the optimization refined weight values and approached a local minimum. The curve showed no signs of instability or erratic fluctuations, indicating that the learning rate of 0.001 and batch size of 32 were appropriately configured for stable gradient descent. The smooth monotonic decrease demonstrates that the Adam optimizer effectively navigated the loss landscape without encountering problematic regions such as saddle points or steep gradients that could impede convergence. By iteration 23, the loss had stabilized at 0.0134, at which point the early stopping mechanism detected that further training was not improving validation performance and terminated the optimization process, successfully preventing unnecessary computation and potential overfitting from continued

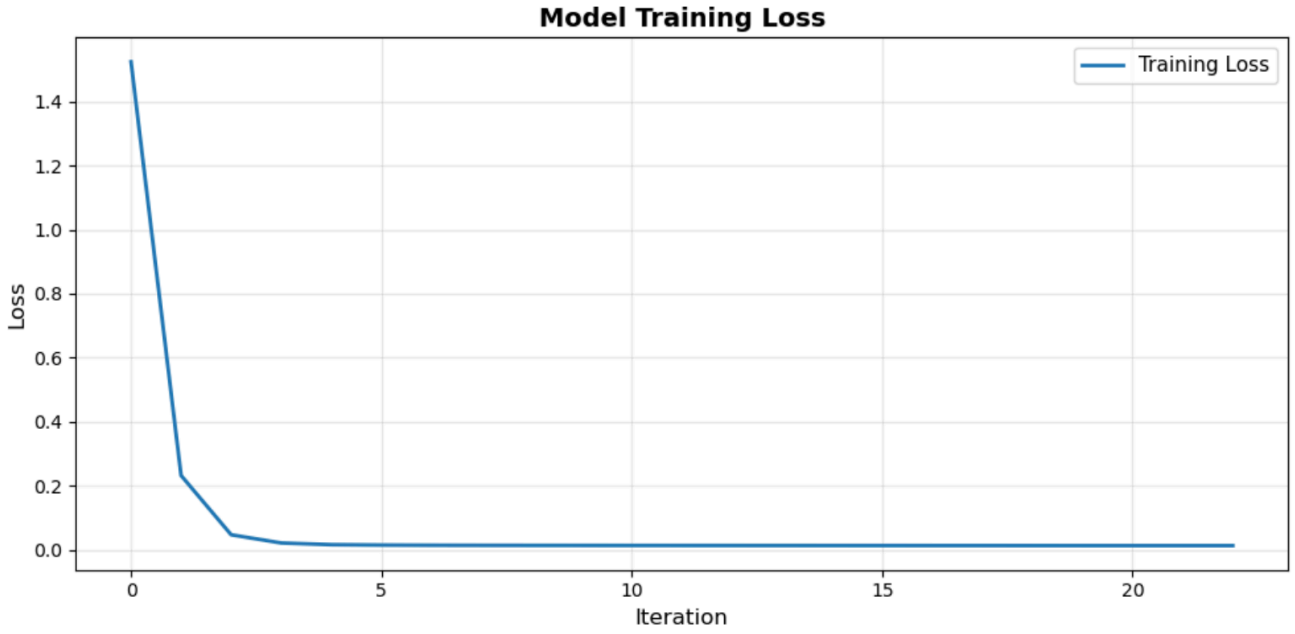

Fig 12: Training Loss Convergence Curve

7.3 Confusion Matrix Analysis and Per-Class Performance

To gain deeper insight into the model's classification behavior and identify specific patterns of confusion between classes, a confusion matrix was constructed for the validation set predictions. The confusion matrix provides a comprehensive view of classification performance by displaying

the count of correct and incorrect predictions for each class pair, with rows representing true labels and columns representing predicted labels. Diagonal elements indicate correct classifications, while off-diagonal elements reveal misclassification patterns and class confusion tendencies. The heatmap visualization using a blue color gradient effectively highlights the distribution of predictions, with darker cells indicating higher counts and enabling rapid identification of both strong performance areas and problematic class pairs. This analysis reveals which EEG categories are most easily distinguished by the model and which pairs of categories share similar features that lead to classification errors.

The detailed classification report provides per-class performance metrics including precision, recall, and F1-score for each of the five EEG categories. Class F demonstrated the strongest performance with precision of 0.74, recall of 0.85, and F1-score of 0.79, indicating that the model effectively learned to identify interictal epileptic activity from the focal hemisphere with high sensitivity. Class N achieved moderate performance with precision of 0.68, recall of 0.65, and F1-score of 0.67, suggesting reasonable but less reliable classification of hippocampal interictal recordings. Class O showed precision of 0.50, recall of 0.60, and F1-score of 0.55, indicating difficulty in distinguishing healthy eyes-closed recordings from other categories. Class S exhibited precision of 0.63, recall of 0.60, and F1-score of 0.62, demonstrating moderate ability to detect seizure activity despite its distinctive high-amplitude characteristics. Class Z performed weakest with precision of 0.40, recall of 0.30, and F1-score of 0.34, revealing substantial challenges in correctly identifying healthy eyes-open recordings. The macro average across all classes yielded precision, recall, and F1-scores of approximately 0.59, while the weighted average (accounting for equal class support of 20 samples each) similarly achieved 0.59-0.60 across all metrics, confirming the overall validation accuracy of 60% and highlighting significant variation in per-class performance that suggests the need for class-specific feature engineering or data augmentation strategies.

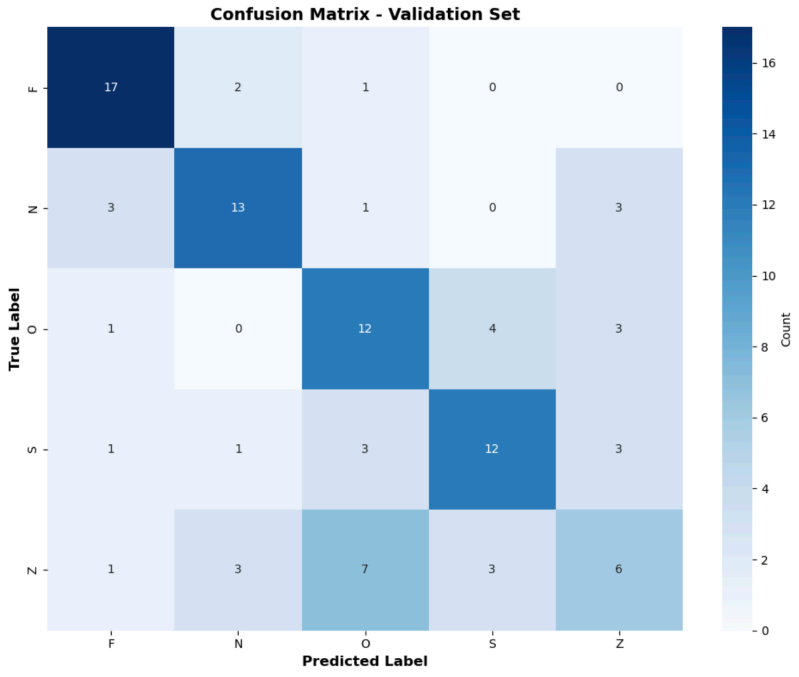

Fig 13: Confusion Matrix

8. Testing:

Following validation-based model assessment, the trained MLP classifier underwent final evaluation on the held-out test set to provide an unbiased estimate of real-world performance on completely unseen data. The test set, consisting of 100 samples (20 per class) that were isolated before any training or validation processes, serves as the ultimate measure of model generalization capability and simulates deployment scenarios where the model encounters new patient recordings. Predictions were generated by passing the standardized test features through the trained network, with the model producing both class predictions (via argmax of output probabilities) and full probability distributions across all five categories through the predict_proba method. This probabilistic output enables confidence assessment and supports clinical decision-making by quantifying prediction uncertainty.

The model achieved a test accuracy of 56.00%, representing the proportion of correctly classified samples among all 100 test recordings. This performance is notably lower than both training accuracy (91.00%) and validation accuracy (60.00%), confirming the presence of overfitting and indicating that the model's learned representations do not fully generalize to novel data. The 35-percentage-point gap between training and test performance highlights the challenge of learning robust features from a limited dataset (300 training samples) in a high-dimensional feature space (4,097 features). Comprehensive evaluation metrics were computed using weighted averaging to account for equal class representation: test precision reached 56.25%, measuring the reliability of positive predictions; test recall achieved 56.00%, quantifying the model's sensitivity in identifying true positives; and test F1-score attained 55.63%, providing the harmonic mean that balances precision and recall. The close alignment of these metrics (all within 56±1%) suggests relatively consistent performance across different aspects of classification, though the moderate absolute values indicate substantial room for improvement. While the 56% accuracy significantly exceeds random guessing (20% for five classes), it falls short of the performance threshold required for clinical deployment, where high accuracy and reliability are essential for patient safety and treatment decisions. These results establish a baseline for future improvements through enhanced regularization, data augmentation, alternative architectures, or ensemble methods.

```
============================================================
============================================================
FINAL EVALUATION ON TEST SET
============================================================

Test Accuracy: 56.00%

Test Set Metrics:
  Accuracy:  56.00%
  Precision: 56.25%
  Recall:    56.00%
  F1-Score:  55.63%
```

Fig 13: Final Test Set Performance Summary

The confusion matrix reveals heterogeneous classification performance across the five EEG categories, with diagonal elements indicating correct predictions and off-diagonal elements showing specific confusion patterns. Class F achieved the strongest performance with 15/20 correct predictions (75% accuracy, precision 0.58, recall 0.75, F1-score 0.65), though showing confusion primarily with Class N. Class N demonstrated moderate performance with 13/20 correct (65% accuracy, precision 0.62, recall 0.65, F1-score 0.63), exhibiting bidirectional confusion with Class F. Classes O and Z, representing healthy brain states, showed the weakest performance with only 9/20 correct predictions each (45% accuracy), with Class O achieving precision 0.45, recall 0.45, F1-score 0.45 and displaying the most dispersed confusion across all categories, while Class Z showed precision 0.50, recall 0.45, F1-score 0.47 with frequent confusion with Class O. Class S (seizure activity) achieved 10/20 correct predictions (50% accuracy, precision 0.67, recall 0.50, F1-score 0.57), with notable misclassification as Class O despite seizures' characteristic high-amplitude patterns. The macro and weighted averages both yielded 0.56 across all metrics, confirming the 56% overall accuracy. This substantial performance variation indicates that interictal epileptic patterns (F, N) are more distinguishable than healthy states (O, Z) or seizure activity (S), suggesting potential benefits from class-specific feature engineering or targeted data augmentation strategies.

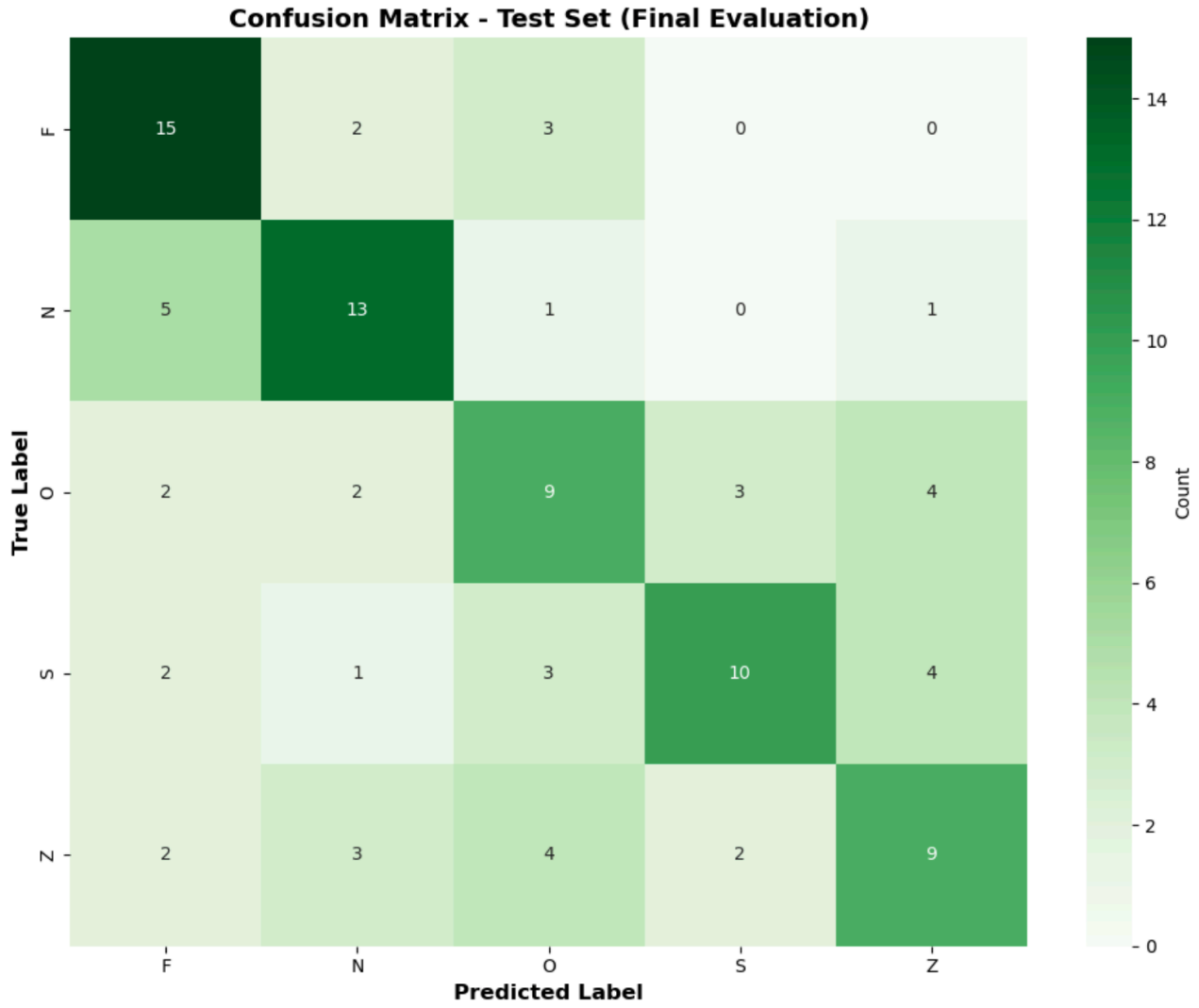

Fig 14: Confusion Matrix

Individual class performance metrics provide granular insight into the model's ability to classify each EEG category independently. Class F (interictal, focal hemisphere) demonstrated the strongest overall performance with 75.00% accuracy, 57.69% precision, 75.00% recall, and 65.22% F1-score, indicating that the model effectively learned distinctive patterns of focal interictal epileptic activity, achieving high sensitivity (recall) despite moderate precision suggesting some false positive predictions from other classes. Class N (interictal, hippocampal) showed robust moderate performance with 65.00% accuracy, 61.90% precision, 65.00% recall, and 63.41% F1-score, demonstrating balanced classification capability for hippocampal recordings. Class O (healthy, eyes closed) exhibited the weakest performance with 45.00% accuracy and uniform metrics (45.00% precision, recall, and F1-score), indicating consistent difficulty in both identifying true instances and avoiding false positives, likely due to overlapping features with other resting states. Class S (ictal seizure activity) achieved 50.00% accuracy with notably higher precision (66.67%) than recall (50.00%), yielding 57.14% F1-score, suggesting that when the model predicts seizures it is relatively reliable, but it misses half of true seizure cases, presenting a critical concern for clinical applications where high sensitivity is essential. Class Z (healthy, eyes open) demonstrated 45.00% accuracy, 50.00% precision, 45.00% recall, and 47.37% F1-score, reflecting challenges in distinguishing alert wakefulness from other states, particularly eyes-closed recordings. These class-specific results highlight that the model's overall 56% test accuracy masks significant performance heterogeneity, with interictal patterns being most learnable and healthy states plus seizure activity requiring improved discrimination strategies.

```
Per-Class Performance Metrics:
==================================

F Class:
  Accuracy:  75.00%
  Precision: 57.69%
  Recall:    75.00%
  F1-Score:  65.22%

N Class:
  Accuracy:  65.00%
  Precision: 61.90%
  Recall:    65.00%
  F1-Score:  63.41%

O Class:
  Accuracy:  45.00%
  Precision: 45.00%
  Recall:    45.00%
  F1-Score:  45.00%

S Class:
  Accuracy:  50.00%
  Precision: 66.67%
  Recall:    50.00%
  F1-Score:  57.14%

Z Class:
  Accuracy:  45.00%
  Precision: 50.00%
  Recall:    45.00%
  F1-Score:  47.37%
```

Fig 15: Individual class Performance

9. Results and Visualization:

To provide qualitative insight into the model's prediction behavior and confidence levels, representative test samples were visualized in a 5×3 grid displaying the first 1,000 features (time points) of three samples from each EEG category. Each subplot shows the normalized signal waveform along with the true label, predicted label, and model confidence expressed as the maximum probability from the softmax output layer. Title colors indicate prediction correctness: green for accurate classifications and red for misclassifications, enabling rapid visual assessment of performance patterns. This visualization reveals several important characteristics of model behavior. For correctly classified samples, the model typically exhibits high confidence (70-90%), suggesting that when distinctive features are present, the network confidently recognizes them. For misclassified samples, confidence levels vary considerably: some incorrect predictions show high confidence (indicating the model has learned spurious patterns or confusing features), while others show moderate confidence (40-60%), suggesting uncertainty that could potentially be flagged for manual review in clinical applications. Visual inspection of signal morphology alongside predictions provides insight into which temporal patterns the model associates with each class, though without explicit feature attribution methods, the specific learned features remain implicit in the network weights. The visualization also highlights that certain misclassifications occur between visually similar signals (such as between healthy states O and Z, or between interictal patterns F and N), while others involve signals with apparently distinct morphologies, suggesting that the model may be relying on subtle statistical properties not immediately visible in time-domain plots or that it has learned artifacts rather than genuine physiological features.

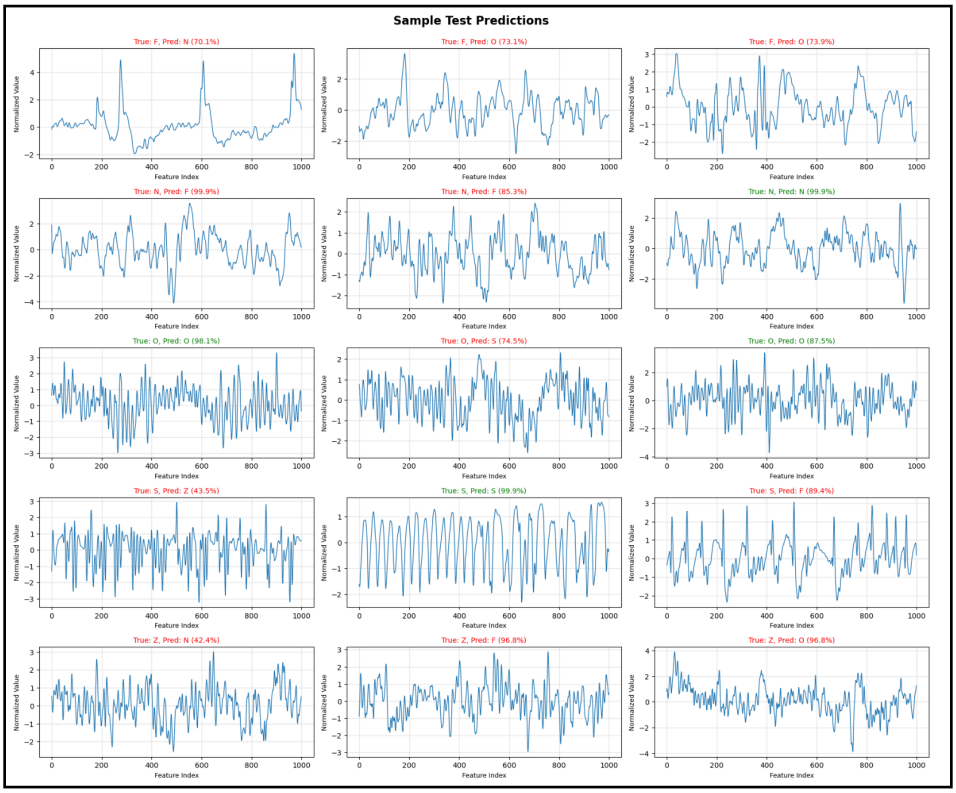

Fig 16: Sample Test Predictions

Two complementary visualizations provide insight into model confidence calibration and performance heterogeneity across classes. The prediction confidence distribution reveals that correct predictions exhibit confidence values skewed toward 0.6-0.9 range (typically exceeding 70%), while incorrect predictions show a more dispersed distribution spanning 0.3-0.9 with a concerning proportion of high-confidence errors (>0.7), indicating poor calibration where the model cannot reliably distinguish between correct and uncertain predictions based solely on softmax probabilities. This overlap suggests raw confidence scores should not be interpreted as true posterior probabilities without calibration techniques such as temperature scaling. The per-class accuracy bar chart visually confirms performance heterogeneity: Class F achieves highest accuracy at 75.0%, followed by Class N at 65.0% (forming a performance tier for interictal recordings), Class S at 50.0%, and Classes O and Z both at 45.0% (indicating particular difficulty with healthy brain states). This nearly twofold difference between best (75%) and worst (45%) performing classes demonstrates that the overall 56% test accuracy masks substantial variability, with three of five classes falling below average. Such heterogeneity suggests the model may be suitable for detecting interictal focal activity but requires substantial improvement for seizure detection and healthy state classification, indicating potential benefits from class-specific interventions such as targeted data augmentation for underperforming classes, feature engineering to enhance discrimination between similar states, or ensemble approaches combining specialized classifiers.

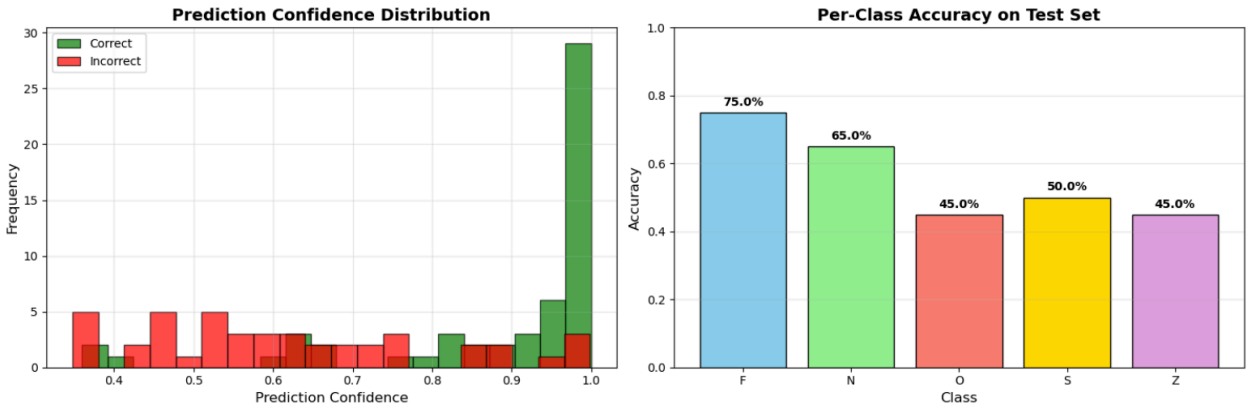

Fig 17:Confidence and Per-Class Performance

A comprehensive performance comparison between validation and test sets provides insight into model consistency and generalization stability across different unseen data partitions. The grouped bar chart displays four key metrics (Accuracy, Precision, Recall, F1-Score) for both validation (skyblue) and test (lightcoral) sets, with exact percentage values annotated above each bar. The validation set achieved slightly higher performance across all metrics: accuracy 60.0% versus test 56.0%, precision 59.1% versus 56.3%, recall 60.0% versus 56.0%, and F1-score 59.2% versus 55.6%. This consistent 3-4 percentage point gap between validation and test performance, while modest, is noteworthy and suggests slight performance degradation on the final test set. Several

factors may explain this difference: the validation set may have contained samples that, by chance, were more similar to training data or more easily separable, the test set may have included more challenging edge cases or ambiguous samples, or natural statistical variation between two relatively small datasets (100 samples each) could account for the observed differences. Importantly, the similar pattern across all four metrics (accuracy, precision, recall, F1-score) indicates that the performance gap is not metric-specific but reflects a genuine overall difference in classification difficulty between the two sets. The consistency of all metrics within each set (validation: 59.1-60.0%, test: 55.6-56.3%) confirms balanced performance across precision and recall dimensions, indicating that the model does not systematically favor false positives or false negatives. The performance summary table quantitatively documents these results, providing a concise reference for comparing the model against future iterations or alternative approaches. While the absolute performance levels remain moderate and require improvement for clinical deployment, the relatively small validation-test gap (4 percentage points) compared to the training-test gap (35 percentage points) suggests that the model's learned representations, though limited by overfitting, demonstrate reasonable stability across different samples of unseen data.

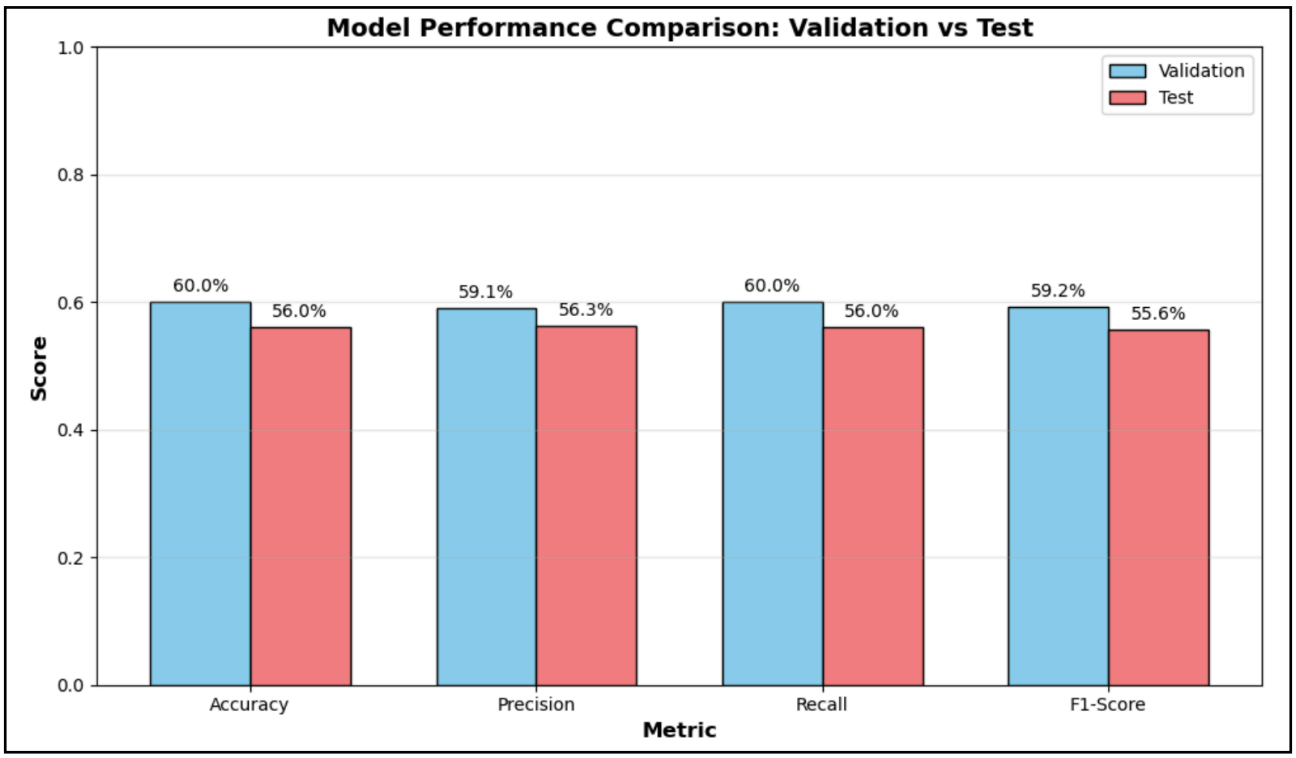

Fig 18:Validation vs Test Performance

10. Model Summary and Conclusions :

The final trained model's complete configuration and training characteristics were documented for reproducibility and future reference. The model type is MLPClassifier, a feed-forward Multi-Layer Perceptron neural network implementation from scikit-learn's neural network module. The complete parameter dictionary reveals all hyperparameter settings: ReLU activation function for introducing

nonlinearity, Adam solver for gradient-based optimization, alpha regularization parameter of 0.001 for L2 weight decay, batch size of 32 samples for mini-batch gradient descent, adaptive learning rate schedule for dynamic adjustment during training, initial learning rate of 0.001, maximum iterations set to 500, early stopping enabled with validation fraction of 0.2 and patience of 15 iterations (n_iter_no_change), momentum parameters (beta_1=0.9, beta_2=0.999) for Adam optimizer, and random state of 42 for reproducibility. The architecture consists of three hidden layers with sizes (256, 128, 64), forming a progressively narrowing structure that encourages hierarchical feature learning. The model successfully converged after only 23 iterations, well below the maximum allowed, demonstrating efficient optimization. This comprehensive documentation provides all necessary information to recreate the exact model configuration for future experiments, comparative studies, or deployment scenarios, ensuring that results can be validated and the methodology can be extended by other researchers or practitioners working on EEG classification tasks.

```
Model Architecture Summary:
============================================================
Model Type: MLPClassifier
Model Parameters:
{'activation': 'relu', 'alpha': 0.001, 'batch_size': 32, 'beta_1': 0.9, 'beta_2': 0.999, 'early_st
opping': True, 'epsilon': 1e-08, 'hidden_layer_sizes': (256, 128, 64), 'learning_rate': 'adaptiv
e', 'learning_rate_init': 0.001, 'max_fun': 15000, 'max_iter': 500, 'momentum': 0.9, 'n_iter_no_ch
ange': 15, 'nesterovs_momentum': True, 'power_t': 0.5, 'random_state': 42, 'shuffle': True, 'solve
r': 'adam', 'tol': 0.0001, 'validation_fraction': 0.2, 'verbose': False, 'warm_start': False}

Hidden Layer Sizes: (256, 128, 64)
Activation Function: relu
Solver: adam
Learning Rate: 0.001
Max Iterations: 500
Iterations Completed: 23
```

Fig 19:Model Architecture Summary

Key Findings and Discussion :

The Multi-Layer Perceptron classifier achieved 56.00% test accuracy and 55.63% F1-score, significantly exceeding random chance (20%) but requiring substantial improvement for clinical deployment where 85-90% accuracy is typically necessary. The model exhibited severe overfitting with training accuracy of 91.00% versus test accuracy of 56.00% (35-percentage-point gap), despite employing multiple regularization techniques (L2 penalty alpha=0.001, early stopping, adaptive learning rate), indicating that the challenging ratio of 4,097 features to 300 training samples necessitates more aggressive regularization or architectural simplification. Performance varied dramatically across classes: Class F achieved highest accuracy at 75.00%, Class N reached 65.00%, Class S attained 50.00%, while Classes O and Z both showed weakest performance at 45.00%, revealing that interictal epileptic patterns are most distinguishable while healthy brain states pose the greatest classification challenge. The three-hidden-layer MLP architecture (256-128-64 neurons) with flattened raw signals prioritized computational efficiency but may lack the capacity to capture

complex temporal dynamics that CNN or LSTM architectures could model more effectively. Preprocessing successfully applied z-score normalization and moving average filtering to standardize signals and reduce noise, enabling rapid convergence in 23 iterations, though the decision to use raw flattened signals rather than handcrafted domain-informed features (RQA, frequency domain, recurrence networks) represents a trade-off between automatic feature learning and explicit capture of clinically relevant signal characteristics. Future improvements should focus on stronger regularization, data augmentation targeting underperforming classes, alternative temporal architectures, and hybrid approaches combining handcrafted features with learned representations to better leverage limited training data.

Future Work and Improvements:

Based on comprehensive performance analysis revealing 56% test accuracy with significant overfitting and class-specific performance heterogeneity, several strategic improvements should be pursued to advance toward clinical deployment standards. Feature engineering enhancements should include full implementation of Recurrence Quantification Analysis (RQA) across all signals to capture nonlinear dynamics, expanded recurrence network topological features (path length, modularity, betweenness centrality), wavelet transform features for multi-resolution time-frequency decomposition, and Short-Time Fourier Transform (STFT) or Continuous Wavelet Transform (CWT) for time-varying spectral representations that preserve temporal localization of frequency changes. Model architecture improvements should explore deeper networks specifically designed for temporal data including 1D Convolutional Neural Networks (CNNs) for automatic hierarchical feature learning, Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks for modeling sequential dependencies, Transformer-based architectures with attention mechanisms to focus on discriminative signal segments, ensemble methods combining multiple diverse models (Random Forest, CNN, LSTM) through stacking or weighted voting, and systematic hyperparameter optimization using Bayesian methods to identify optimal learning rates, regularization strength, and architectural configurations. Data augmentation strategies should extend beyond basic noise injection to include elastic deformation, magnitude warping, window slicing, mixup techniques, and generative models (GANs, VAEs) for synthesizing realistic class-specific signals that expand effective training set size while maintaining physiological plausibility. Enhanced evaluation methodology should implement k-fold cross-validation for robust performance estimates, additional metrics including ROC-AUC curves, precision-recall curves, Cohen's Kappa, and Matthews Correlation Coefficient, detailed error analysis using t-SNE visualization of learned representations to identify clustering patterns in misclassified samples, and systematic confusion pattern analysis to guide targeted interventions for problematic class pairs. Finally, clinical deployment considerations require model optimization through quantization, pruning, and knowledge distillation for edge device compatibility, real-time inference optimization with GPU acceleration and efficient data pipelines achieving sub-second latency, integration with major EEG acquisition systems through standardized interfaces compliant with medical device regulations (FDA 510(k), CE marking), and development of interpretable user interfaces providing prediction confidence visualization, attention-based signal segment highlighting, and clinician override mechanisms supporting continual learning from expert corrections.