Sleep Stage Classification with Classical Machine Learning and Deep Learning

Final Project Report, EEE 598 Fall 2025

Aman Gupta, Preston Garrett, Humberto Delgado, Bhavya Minesh Shah

December 2025

1 Abstract

Polysomnography (PSG) serves as the gold standard for diagnosing sleep disorders, relying on the manual classification of physiological signals into distinct sleep stages. However, this process is labor-intensive, time-consuming, and subject to inter-rater variability. This project investigates the efficacy of Deep Learning (DL) architectures for automated sleep stage classification using the Sleep-EDF dataset. We implemented and compared four distinct architectures across two signal modalities: Time-Domain models (RNN and Transformer) and Frequency-Domain models (CNN and Vision Transformer). The pipeline involved preprocessing single-channel EEG signals into 30-second epochs and generating spectrograms for spectral analysis. Our evaluation reveals that the Time-Domain Transformer model achieved the highest overall performance with an accuracy of 42.15% and an F1 score of 0.396. While all models demonstrated high sensitivity in detecting deep sleep (N3), with the CNN achieving 99.28% recall for this class, performance dropped significantly for transitional stages like N1 and REM. This disparity highlights the challenge of class imbalance and the difficulty of distinguishing morphologically similar sleep stages without multimodal sensor data. This study demonstrates the potential of automated systems while underscoring the necessity for advanced data augmentation and class-balancing strategies in future clinical applications.

2 Introduction

Sleep is a complex physiological process essential for human health, affecting memory consolidation, cognitive function, and cardiovascular health [13]. Far from a state of inactivity, the sleeping brain exhibits predictable, dynamic patterns of electrical activity that are categorized into distinct stages [6]. These stages are broadly divided into Rapid Eye Movement (REM) and Non-REM (NREM) sleep [6]. NREM is further

subdivided into three stages: N1 (light sleep), N2 (deeper sleep characterized by sleep spindles), and N3 (deep or slow-wave sleep) [14]. A typical sleep cycle progresses through these stages in approximately 90 to 120-minute intervals [12].

The architecture of these sleep stages, specifically their duration, arrangement, and transitions, provides critical diagnostic information for sleep-related disorders such as sleep apnea, narcolepsy, and parasomnias [11]. Disruptions in sleep architecture are also clinically relevant indicators for depression, aging, and traumatic brain injuries [11]. Consequently, accurate identification of sleep stages is paramount for effective diagnosis and treatment planning.

2.1 Polysomnography and Data Acquisition

The standard diagnostic tool for monitoring sleep architecture is the Polysomnography (PSG) study. A PSG is a comprehensive test that records multiple physiological parameters simultaneously [6]. While considered non-invasive, the procedure is often intrusive and uncomfortable for patients due to the extensive array of sensors required, usually involving more than 20 electrodes attached to the body [2]. A standard PSG setup typically involves:

- **EEG (Electroencephalogram):** To record brain wave activity and classify sleep stages [9].
- **EOG (Electrooculogram):** To track eye movements, essential for identifying REM sleep [9].
- **EMG (Electromyogram):** To monitor muscle tone and atonia [9].
- **ECG and Respiratory Sensors:** To monitor heart rate, blood oxygen levels (pulse oximetry), and breathing effort (nasal cannula, chest movement belts) [9].

The complexity of the wiring and the unfamiliar environment often lead to patient discomfort, which can alter sleep patterns, a phenomenon known as the ”first-night effect”, potentially necessitating repeat testing [3].

2.2 Signal Characteristics and Manual Scoring

Following data acquisition, sleep technicians manually analyze the continuous physiological data by dividing it into 30-second windows known as ”epochs” [11]. Each epoch is assigned a sleep stage based on specific frequency bands and waveform morphologies observed in the EEG signals [5]:

- **Awake:** Characterized by high-frequency Beta waves (13–30 Hz) and Alpha waves (8–12 Hz) when drowsy [11].

- **N1 (Stage 1):** Defined by a transition to Theta waves (4–7 Hz) and the presence of muscle tone [11].
- **N2 (Stage 2):** Distinctive due to the presence of K-complexes (high amplitude biphasic waves) and sleep spindles (bursts of 11–16 Hz oscillatory activity) [5].
- **N3 (Stage 3):** Characterized by high-amplitude, low-frequency Delta waves (< 4 Hz) [11].
- **REM:** Exhibits mixed-frequency activity similar to wakefulness (Beta waves) combined with rapid eye movements and muscle atonia [11].

2.3 Motivation for Automation

While manual scoring remains the clinical standard, it is a bottleneck in sleep medicine. It requires highly specialized training, is time-intensive, and is prone to subjective interpretation, leading to discrepancies between scorers. The subtle distinctions between stages—such as differentiating the Theta waves of N1 from the background activity of N2 without distinct spindles—present significant challenges. To address these limitations, this project explores the application of Machine Learning (ML) and Deep Learning techniques to automate sleep stage detection. By leveraging both Time-Domain approaches (Recurrent Neural Networks and Transformers) and Frequency-Domain approaches (Convolutional Neural Networks and Vision Transformers), we aim to develop a robust model capable of classifying sleep stages from single-channel EEG data, potentially reducing the reliance on cumbersome multi-sensor setups and manual scoring efforts.

3 Method/Strategy

3.1 Exploratory Data Analysis and Feature Extraction

Although the goal is to have a proven Machine Learning (ML) model for stage classification, it was important to run Exploratory Data Analysis (EDA) of the database to identify how the data was skewed and what feature engineering was necessary for any choice of algorithm to work correctly.

Due to the patient-ID labeling from the database, we selected 101 patients for the classical ML models, since we faced limitations in data loading, but from the EDA these were enough samples, especially when breaking the data into 30-second epochs. As expected, the data set was highly imbalanced, and a majority of the samples were classified as Stage-2 NREM, as expected from [1]. To mitigate this we used SMOTE to improve imbalance which was specifically for Stage-4 NREM.

Previous literature shows significant accuracy in classical ML models with good feature engineering, specifically by extracting features from the EEG(front and posterior regions), EOG, and EMG signals. All of which were present from the PSG signal of the database. In total, 42 features were extracted so there

was high dimensionality in the features, and as expected from all the features, there was significant overlap in our features even when PCA and T-SNE(for visualizing) was applied. One important thing to note from dimension reduction is that features have linear relationships, since T-SNE showed no significant clusters of the data. The PCA plot shown below demonstrates the significant overlap among the extracted features.

Although our data was highly overlapped due to feature engineering, the ML pipeline was still able to achieve valid results compared to other studies not using advanced or deep learning algorithms. The overall algorithm consisted of initially cleaning the signals, i.e., the EEG, EMG, and EOG signal. Although these signals contributed to the overall accuracy of the models, one of the most important features is the EEG signal band ratio power and spindle features. The background is in the physiological interpretation since at each sleep stage each frequency band holds different amounts of power.

Following feature extraction, we applied PCA to reduce the dimensionality of our features since several of our tested models seemed to be suffering from too many dimensions, specifically SVM with RBF Kernel, K-Nearest Neighbors, and our top performer XGBoost. Following dimensionality reduction, the data were split into an 80/20 split and we ensured that the data were split patient wise, instead of epoch wise. To ensure that the model would learn features and not patients. One crucial design choice in the ML pipeline was the use of Synthetic Minority Oversampling Technique, although the pipeline incorporated this in the early stages it was observed that this technique actually did not help all models. At this point our pipeline deviated depending on the algorithm of choice. Following SMOTE, we trained and evaluated K-Nearest Neighbor, Support Vector Machine, Random Forest, and XGBoost. Following the evaluation feature importance was evaluated with Shapely values.

3.2 Deep Learning Approach

To complement the classical machine learning pipeline, we designed a deep learning framework that operates more directly on the EEG time series and their time–frequency representations. The goal of this part of the project is to evaluate whether sequence models (RNNs and Transformers) and image-based architectures (CNN and Vision Transformer) can better capture the rich non-linear structure of sleep dynamics than classical models built on hand-crafted features.

3.2.1 Dataset and Problem Formulation

For the deep learning experiments, we used a subset of the Sleep-EDF database (sleep cassette recordings). Each nocturnal polysomnography (PSG) recording contains EEG, EOG and EMG channels as well as an accompanying hypnogram with 30-second sleep stage annotations. From each subject we primarily used the

EEG channels Fpz–Cz and Pz–Oz when available, following common practice in the sleep staging literature.

Signals were resampled or treated at a sampling rate of $100\,\mathrm{Hz}$ and segmented into non-overlapping 30-second epochs, so that each epoch corresponded to exactly one hypnogram label. Epochs annotated as "movement", "unknown" or unlabeled intervals were discarded. Stages 3 and 4 were merged into a single deep sleep class (N3), yielding a five-class classification problem:

- Wake.
- N1 (light sleep).
- N2 (intermediate NREM).
- N3 (deep NREM, merged stages 3+4).
- REM.

Labels were encoded as integers $\{0, \ldots, 4\}$ and all subsequent models were trained to predict this stage index for each 30-second epoch.

3.2.2 Signal Preprocessing

Deep models are sensitive to noise and artifacts, so we implemented a common preprocessing pipeline applied channel-wise to all EEG signals before segmentation:

1. **Band-pass filtering:** A 5th-order Butterworth band-pass filter between 0.5 and $30\,\mathrm{Hz}$ removes slow drifts and high-frequency muscle artifacts while preserving the EEG alpha, beta, theta and delta bands. The filter order controls the steepness of the transition band; here an order of 5 offers a compromise between sharpness and numerical stability.
2. **Notch filtering:** An IIR notch filter centered at $50\,\mathrm{Hz}$ with quality factor $Q = 30$ suppresses narrow-band power-line interference. The Q-factor controls how narrow the removed band is: high Q removes a very narrow range around $50\,\mathrm{Hz}$, preserving neighboring frequencies while efficiently attenuating the mains component.
3. **Artifact removal:** Amplitude-based artifact detection is applied with a threshold of 3 standard deviations from the mean. Samples exceeding this range are treated as artifacts and replaced by linear interpolation between neighboring clean points. The threshold determines the trade-off between rejecting artifacts and retaining high-amplitude but physiologically plausible events (e.g., K-complexes).
4. **Normalization:** Each continuous EEG trace is z-score normalized (zero mean, unit variance). This step prevents channels with large absolute amplitudes from dominating the gradients during training and improves numerical stability across different subjects.

After preprocessing, signals are segmented into 30-second epochs aligned with the hypnogram annotations and stored as contiguous arrays for dataset construction.

3.2.3 Time and Frequency Domain Representations

We constructed two complementary deep learning datasets from the preprocessed EEG:

Time-domain Dataset For the recurrent and Transformer models, each 30-second epoch is represented as a sequence of raw samples from one EEG channel. With a sampling rate of $100\,\mathrm{Hz}$ this corresponds to $30 \times 100 = 3000$ samples per epoch. Epochs shorter than the target length are padded with zeros; longer epochs are truncated. Each example thus has shape $[3000, 1]$ (sequence length by channel), and the full dataset consists of

$$\{\mathbf{x}_i \in \mathbb{R}^{3000\times 1},\, y_i \in \{0, \ldots, 4\}\}_{i=1}^{N}.$$

Frequency-domain Dataset For the CNN and Vision Transformer, we use 2-D time–frequency representations derived from the same 30-second segments:

- *Power spectral density (PSD):* Welch's method with 2-second windows (200 samples) and 50% overlap estimates the average power in canonical bands (delta, theta, alpha, beta and low gamma). These band powers are primarily used for exploratory analysis.
- *Spectrograms:* A short-time Fourier transform (STFT) with the same windowing yields a spectrogram $S(f, t)$, which is log-scaled and limited to frequencies below $30\,\mathrm{Hz}$. Each epoch is thus represented as an image-like matrix with frequency along one dimension and time along the other.

Spectrograms are then normalized across the dataset and zero-padded so that all examples share a common height and width. A singleton channel dimension is added to obtain tensors of shape $[1, H, W]$ suitable for 2-D convolutions and patch-based Vision Transformers.

3.2.4 Deep Learning Architectures and Hyperparameters

We implemented four deep architectures: a bidirectional LSTM, a Transformer encoder, a convolutional neural network (CNN) and a Vision Transformer (ViT). All four models use a common set of optimization hyperparameters unless stated otherwise:

- Batch size: 8.
- Optimizer: AdamW with weight decay 10^{-4}.

- Initial learning rate: 10^{-4} for the RNN and CNN, 5×10^{-5} (half the base rate) for the Transformer and ViT to improve stability on deeper attention-based models.
- Dropout rate: 0.3 in fully connected layers and dropout2d in convolutional layers.
- Maximum epochs per fold: 100 with early stopping.
- learning-rate scheduler: ReduceLROnPlateau with factor 0.3 and patience 4 epochs.

The learning rate determines the step size of gradient updates; smaller values, especially in Transformers and ViTs, prevent divergence and allow more stable convergence. Dropout and weight decay act as regularizers that discourage overfitting on the relatively small dataset. The scheduler automatically reduces the learning rate when the validation loss plateaus, allowing finer convergence near minima.

Bidirectional LSTM (RNN). The recurrent model treats each epoch as a sequence of 3000 scalar samples:

- Three stacked LSTM layers with hidden size 128 per direction.
- Bidirectional processing so that forward and backward hidden states are concatenated. effectively doubling the temporal context.
- A classification head consisting of a linear projection to 64 units, ReLU, dropout and a final linear layer to the 5 output classes.

The hidden size controls the capacity of the recurrent state: larger values allow more complex dynamics but also increase parameter count and risk of overfitting. The number of layers controls modelling depth and the ability to capture long-range dependencies.

Transformer Encoder. For the time-domain Transformer, each sample is first linearly projected from dimension 1 to a model dimension $d_{\text{model}} = 128$ and augmented with sinusoidal positional encodings. The projected sequence is then passed through:

- Four Transformer encoder layers, each with 8 self-attention heads and feed-forward sublayers with dropout 0.3.
- A final layer normalization, followed by a two-layer MLP classifier (128→64→5) with ReLU and dropout.

The model dimension sets the size of the shared representation space; the number of heads controls how many distinct attention patterns the model can learn in parallel, and the number of layers controls depth and receptive field over the sequence.

CNN on Spectrograms. The convolutional model operates on spectrograms $[1, H, W]$ using a hierarchical feature extractor:

- Four convolutional blocks with filter counts 32, 64, 128 and 256 (kernel size 3×3), each followed by batch normalization, ReLU, 2×2 max pooling and dropout2d.
- An adaptive average pooling layer to aggregate spatial information into a fixed-size 256-dimensional vector regardless of input resolution.
- A fully connected classifier (256→128→64→5) with ReLU activations and dropout.

Increasing the number of filters per layer allows the network to learn more diverse local patterns at each scale, while pooling layers expand the effective receptive field and reduce spatial resolution.

Vision Transformer on Spectrograms. The ViT treats the spectrogram as a set of non-overlapping image patches:

- The spectrogram is cropped so that its height and width are divisible by a patch size of 5×5.
- A convolutional patch embedding with kernel size and stride 5 converts each patch into a 128-dimensional token.
- A learnable classification token (CLS) is prepended to the token sequence; all tokens receive trainable positional encodings.
- Four Transformer encoder layers with 8 heads process the token sequence.
- The final CLS token is passed through a linear classifier to produce the 5 logits.

The patch size trades off spatial resolution and sequence length: smaller patches preserve more local detail but create longer token sequences and higher computational cost; we found 5×5 to be a practical compromise.

3.2.5 Training, Validation and Evaluation

We used 5-fold cross-validation with patient-wise splits, ensuring that all epochs from a given subject appear either in the training or validation set but never in both. For each fold, we computed empirical class frequencies and derived inverse-frequency weights to construct a *WeightedRandomSampler* for the training set. This sampler preferentially draws rare classes (such as N1 and REM), partially compensating for class imbalance during optimization.

All models were trained using a cross-entropy loss, which naturally handles the multi-class setting. Early stopping monitored the validation F1-score with a patience of 10 epochs to prevent overfitting. At the end

of each fold, we recorded:

- Overall accuracy, precision, recall and F1-score (weighted by class frequency).
- One-vs-rest AUROC averaged in a class-weighted manner.
- Per-class precision, recall and F1.
- Normalized confusion matrices for qualitative error analysis.

The final performance for each architecture is reported as the average across the five folds.

3.2.6 Expected Outcome

Given the physiological characteristics of sleep stages, we expected deep models to perform best on N3 (deep sleep), which has distinctive high-amplitude delta activity, and to struggle most with N1 and REM, which can exhibit overlapping spectral features and are under-represented in the dataset. We further expected attention-based models (time-domain Transformer and spectrogram-based ViT) to outperform purely recurrent or convolutional approaches due to their ability to capture both local and long-range dependencies and to model complex temporal interactions across the full 30-second epoch.

4 Results

4.1 Results for Classical Machine Learning Approach

Classical ML Models showed promising results, if proper feature extraction was given, as with any model. From a pipeline perspective we expected the data to be imbalanced since humans typically spend the majority of sleep in the N2 stage, however this could easily be mitigated with an increased in samples, since we would have large representations of every class. However, since this stage appears the most in our data we also expected models to perform better in this stage, but this only occurred when not using an oversampling technique.

One important feature that drastically improved the accuracy of all models was the log ratio frequency band power and the spindle density for each 30 second epoch. The *Moelle 2011* algorithm was used for spindle detection and we counted how many times a spindle was detected and averaged it over the epoch time length. Due to the time, the *WONAMBI* library was used for python and this library is not yet proven, so results still have room for improvement. Given that spindle density can be tracked another way.

Follow the initial processing and EDA, we extracted the features and decided to visualize the features using PCA and T-SNE. Regardless, of the dimensionality reduction technique our visual results were suffering, but regardless of dimensionality reduction only some models showed improvement, such as the SVM with RBF Kernel.

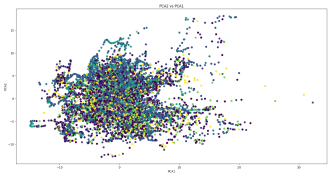


Figure 1: PCA plot (PCA2 vs PCA1) with Class Assignments

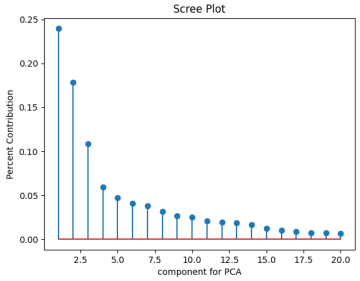


Figure 2: Scree Plot for PCA

As expected from the classical ML model, results were quite underwhelming with our best accuracy plateauing at about 35-40 percent accuracy and our best recall score of 0.62. Below is a table describing the results:

	N1	N2	N3	N4	WAKE	REM
KNN	0.34	0.28	0.24	0.10	0.08	0.18
SVM	0.32	0.13	0.28	0.28	0.25	.38
Random F.	0.22	0.30	0.10	0.0	0.29	0.20
XGBoost	0.37	0.42	0.27	0.05	0.18	0.38

Table 1: F1 Score Per Class

Although from first glance the XGBoost algorithm seems the most accurate based off F1 score, the confusion matrices show that SVM does best at predicting True Positive(TP) and True Negatives(TN). While XGBoost reveals that it suffers a lot from the imbalanced data set since the Recall for class N3 is very close to zero.

From observing the confusion matrix, all models are very weak in classifying N4, WAKE, and REM stage. This is without a doubt, a result of weak feature engineering and being unable to find good data features that represents all stages. Even when observing the SHAP value beeswarm plot there are few feature that have significant impacts on the model in the last three classes mentioned above. Overall, the ML pipeline is giving us true performing results, but currently not the best results. Future results could easily be improved given more time is dedicated to feature engineering to improve upon EEG activity given the specific sleep stage.

4.2 Deep Learning Results

4.2.1 Overall Model Comparison

Table 2 summarizes the 5-fold cross-validation metrics for the four deep architectures. Values are averaged over folds and reported as weighted scores to account for class imbalance.

Model	Accuracy	Precision	Recall	F1-score	AUROC
RNN (BiLSTM)	0.3465	0.3057	0.3465	0.2745	0.6794
Transformer	**0.4215**	**0.4505**	**0.4215**	**0.3960**	**0.7654**
CNN (Spectrogram)	0.3193	0.2239	0.3193	0.2119	0.7588
Vision Transformer	0.3710	0.4077	0.3710	0.3150	0.7538

Table 2: Global performance of deep learning models (5-fold cross-validation, weighted metrics).

The time-domain Transformer achieves the best performance across all global metrics, with an accuracy of approximately 42% and a weighted F1-score close to 0.40. The Vision Transformer ranks second, outperforming both the RNN and the CNN. The RNN provides moderate performance but is clearly inferior to the attention-based models, while the spectrogram CNN yields the lowest F1-score despite a competitive AUROC, indicating that its probability estimates separate classes reasonably well but its hard predictions are biased towards a subset of stages.

Training dynamics also differ across architectures. The best epochs (highest validation F1) typically occur around epochs 7–19 for the RNN (mean $\approx$ 15), epochs 2–24 for the Transformer (mean $\approx$ 12), epochs 1–2 for the CNN (mean ≈ 1.2), and epochs 3–13 for the ViT (mean $\approx$ 7). The CNN in particular tends to reach its peak performance in the very first epoch and then overfit, suggesting that a smaller learning rate or stronger regularization might be required for more stable training.

4.2.2 Per-stage Performance

To better understand how each architecture behaves across sleep stages, Table 3 reports the per-class F1-scores averaged over folds.

Stage	RNN	Transformer	CNN	ViT
Wake	0.1798	**0.3968**	0.1456	0.2978
N1	0.1688	0.2118	**0.2726**	0.2692
N2	0.1865	**0.3364**	0.0385	0.2048
N3	0.5874	**0.7247**	0.4712	0.5720
REM	0.1853	**0.2593**	0.0352	0.1544

Table 3: Per-class F1-scores for deep learning models (5-fold average).

Deep sleep (N3) : All models classify N3 relatively well, with F1-scores between 0.47 and 0.72. The Transformer attains both the highest N3 precision and recall, whereas the CNN and ViT reach extremely high N3 recall (≈ 0.99 and 0.98, respectively) but lower precision, indicating that they tend to over-predict N3 at the expense of other stages. This behavior is consistent with the physiology of N3, which exhibits prominent low-frequency delta power that is easily recognized in both time and time–frequency domains.

Wake : Wake epochs are recognized with moderate success. The Transformer again leads with an F1-score close to 0.40. The ViT obtains higher precision but lower recall, meaning it is conservative when labeling Wake and misses a substantial fraction of actual Wake epochs. RNN and CNN perform poorly on this stage, likely because quiet wakefulness can resemble light NREM in both amplitude and spectral content.

N1 and REM : As expected, N1 and REM are the most difficult stages. They are underrepresented in the dataset and share features with neighboring stages (e.g., N1 with Wake and N2; REM with Wake-like EEG but NREM-like muscle tone). For N1, the CNN and ViT obtain slightly higher F1-scores than the Transformer and RNN, but all values remain below 0.28. For REM, the Transformer clearly outperforms the other models (F1 ≈ 0.26), thanks to its much higher recall (≈ 0.58) compared to the CNN (recall ≈ 0.04) and ViT (recall ≈ 0.19). The CNN rarely predicts REM, leading to both low precision and recall.

N2 : N2 performance lies between the extremes, with the Transformer again delivering the best F1-score (0.34). The CNN essentially fails to learn this class (F1 ≈ 0.04), misclassifying many N2 epochs as N3, which further supports the observation that the CNN is overly biased towards deep sleep.

4.2.3 Confusion Matrix Analysis

Figure 4 shows the normalized confusion matrices for the four deep learning models. These visualizations align with the quantitative metrics described above.

Across all four matrices, the N3 row exhibits the strongest diagonal, indicating that deep sleep is consistently classified correctly. The Transformer and ViT additionally show relatively solid diagonals for Wake and N2, whereas the CNN matrix reveals a strong bias towards predicting N3 regardless of the true stage. Significant off-diagonal mass between N1 and REM in the Transformer matrix indicates specific confusion between these two light/REM stages, which is physiologically plausible and consistent with the low per-class F1-scores.

4.2.4 Evaluation Metrics

Given a confusion matrix $\mathbf{C}$, we calculate overall accuracy as $(\sum C_{kk})/(\sum C_{ij})$. For each class k, we compute $\text{Precision}_k = C_{kk}/\sum_i C_{ik}$ and $\text{Recall}_k = C_{kk}/\sum_j C_{kj}$. The F1-score is the harmonic mean of precision and recall. We report the weighted average for all metrics based on class support w_k.

The Area Under the Receiver Operating Characteristic curve (AUROC) is computed in a one-vs-rest fashion from the predicted class probabilities by sweeping a decision threshold over $[0, 1]$, computing the true positive rate and false positive rate at each threshold, and numerically integrating the resulting ROC curve. The per-class AUROC values are then averaged with the same class-frequency weights w_k.

Applying these definitions to the confusion matrices and prediction scores of each model over all cross-validation folds yields the summary in Table 4

Model	**Accuracy**	**Precision**$_w$	**Recall**$_w$	**F1**$_w$	**AUROC**
RNN (BiLSTM)	0.3465	0.3057	0.3465	0.2745	0.6794
Transformer	**0.4215**	**0.4505**	**0.4215**	**0.3960**	**0.7654**
CNN (Spectrogram)	0.3193	0.2239	0.3193	0.2119	0.7588
Vision Transformer	0.3710	0.4077	0.3710	0.3150	0.7538

Table 4: Accuracy, weighted precision, weighted recall, weighted F1-score and AUROC for each deep learning model, computed from their confusion matrices and prediction scores over the 5-fold cross-validation.

4.2.5 Effect of Hyperparameters

The results also reflect the impact of key hyperparameters discussed in Section 3.2:

- **Learning rate and regularization.** The lower learning rate used for the Transformer and ViT improves their stability and final performance compared with the RNN and CNN, which use a higher rate and tend to overfit more quickly. The CNN's rapid convergence within one epoch suggests that either the learning rate is too aggressive for the current architecture or that additional regularization (e.g., stronger weight decay, data augmentation, or label smoothing) is needed.
- **Model capacity.** The deeper and wider attention-based models (Transformer and ViT) are able to leverage their higher capacity without catastrophic overfitting, thanks in part to dropout, weight decay and early stopping. In contrast, increasing hidden size and layer count in the LSTM does not close the performance gap, highlighting the advantage of self-attention for modeling long-range temporal dependencies in EEG.
- **Class imbalance handling.** The use of a WeightedRandomSampler based on inverse class frequencies provides some improvement on minority classes such as REM and N3; all models achieve reasonable N3 recall despite fewer examples. However, N1 and REM performance remains modest, indicating that sampling alone cannot fully resolve severe imbalance when classes are also intrinsically hard to separate.

4.2.6 Summary of Deep Learning Performance

Overall, the deep learning experiments demonstrate that:

- Attention-based models (time-domain Transformer and spectrogram-based ViT) consistently outperform both the recurrent and convolutional baselines in terms of global metrics and per-stage F1-scores.
- Deep sleep (N3) is the easiest stage to classify, with all models achieving high recall, whereas N1 and REM remain the bottleneck due to their transitional and heterogeneous nature.
- Properly chosen hyperparameters—notably a smaller learning rate, strong regularization and patient-wise cross-validation—are crucial for training stable and interpretable deep models on relatively small, imbalanced clinical datasets.

These results indicate that the deep learning pipeline, especially the Transformer-based architectures, provides a clear improvement over the classical machine learning models and forms a promising foundation for future refinements such as multi-channel inputs, additional modalities and more sophisticated imbalance-aware loss functions.

5 Discussion

5.1 Classical Machine Learning Models

Sleep stage classification is currently a difficult problem especially for lightweight or classical approaches as it requires extensive feature engineering. Not only was it challenging, but we also encountered problems with loading the dataset so we had a limited amount of patients(101) to train and test our models, creating the major class imbalance and under-sampling. However, we are highly confident in the structure of our pipeline because we improved it to a point where we saw a 15% increase in f1 score and a bump up in recall score. This was achieved by adding time context to our training features where each current feature "knew" the result of the past and previous features. We decided to added this because physiologically speaking an actual sleep cycle, specifically NREM sleep, is dependent on previous stages of sleep. From interpreting the SHAP values and the PCA Scree plot it is quite obvious that an improvement in valuable features could put our best performing model(SVM with RBF Kernel) into the 70-80% accuracy range, but to do this, a deeper understanding of sleep spindles, k-complexes, and wave signatures for REM is needed. Identifying characteristic features for REM is one of the most important features, since we spent more time researching and implementing characteristics of NREM sleep stages we believe that is why most of our models perform so strong in N1 and N2. Future works will include a much larger dataset and more slow wave features to better represent the REM stage of sleep.

6 Conclusion

This project examined both classical machine learning methods and modern deep learning architectures for automated sleep stage classification using single channel EEG data from the Sleep EDF dataset. The task remained challenging throughout, shaped by class imbalance, limited data availability, and the subtle physiological similarities that exist between neighboring sleep stages. Classical machine learning models relied heavily on carefully crafted features, yet their performance ultimately remained modest, reaching only 35 to 40 percent accuracy. Deep learning models demonstrated clearer strengths, especially the time domain Transformer, which achieved the highest accuracy and F1 score among all models. Even so, all approaches, classical or deep, struggled most with transitional and underrepresented stages such as N1 and REM. These patterns reveal both the promise and the current limitations of automated sleep staging from a single EEG channel.

The superior performance of attention based architectures suggests that models capable of learning long range temporal patterns are better suited to the complexities of sleep dynamics. Nevertheless, the results also point toward several important directions for future work. Larger and more balanced datasets, the inclusion

of additional PSG modalities such as EOG and EMG, and improved imbalance aware training strategies will be essential if automated systems are to approach the reliability required for clinical use. Sleep is not a static event but a continuous physiological journey, and models that recognize this temporal structure will almost certainly provide more faithful and more clinically meaningful predictions.

Beyond the technical achievements, this project serves as a reminder of why sleep matters in the first place. It is not merely a stage to be scored or a sequence of signals to be analyses. Sleep is a profound act of restoration that sustains our physical health and shapes our cognitive well being. A good night of sleep strengthens the immune system, protects the heart, restores muscle function, and preserves hormonal balance. Equally important, it clears the mind, steadies the emotions, and prepares the brain to learn, remember, and create. When sleep falters, life itself grows heavier. When sleep improves, the whole human experience becomes lighter, clearer, and more resilient. The work of understanding sleep is therefore not only a scientific pursuit but also a deeply human one. It aims, in the end, to help people move through their days with greater clarity, steadiness, and vitality.

A Figures

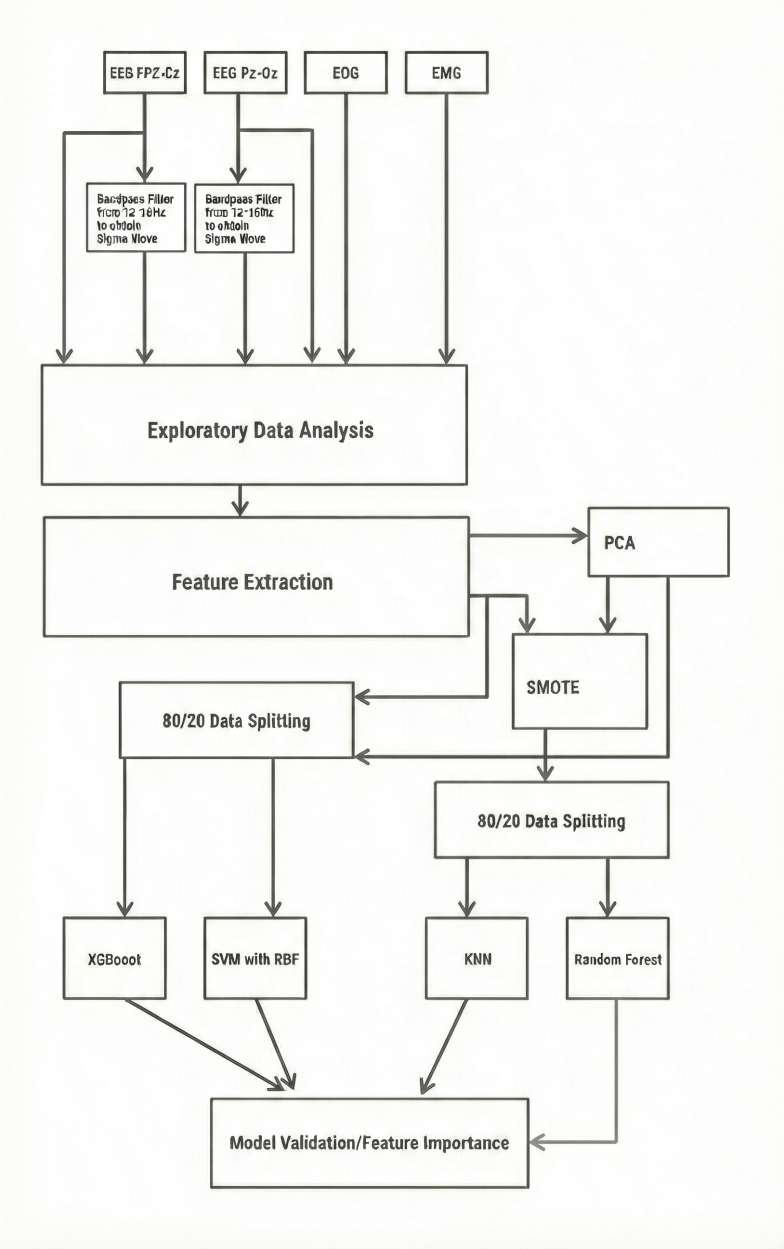
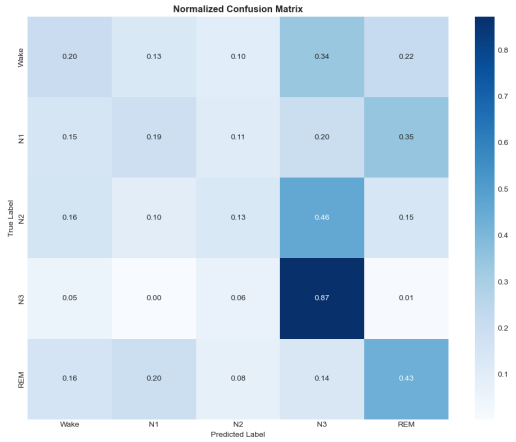
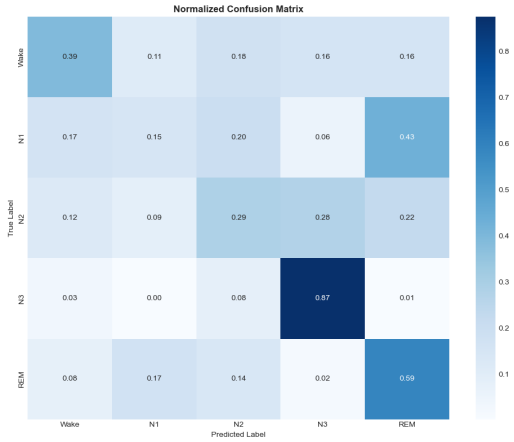


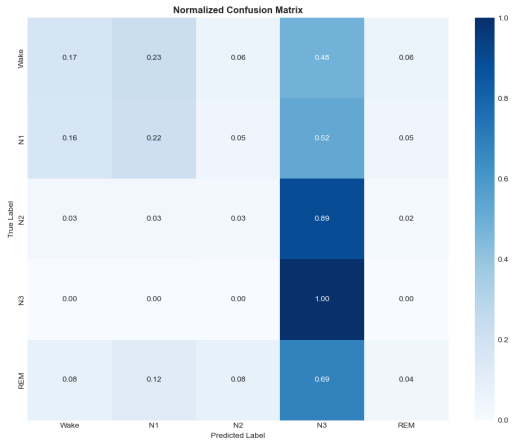
Figure 3: Block Diagram of the ML Pipeline for Sleep Stage Classification



(a) RNN



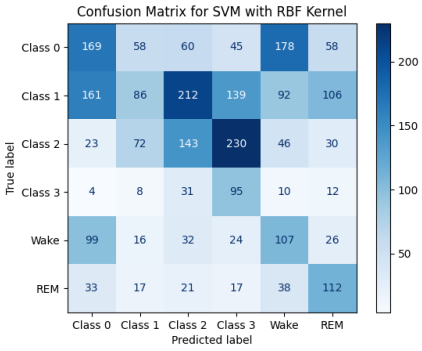
(b) Transformer



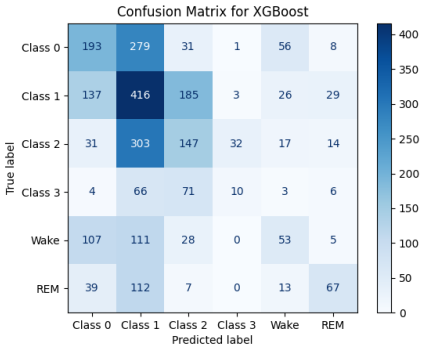
(c) CNN

(d) Vision Transformer

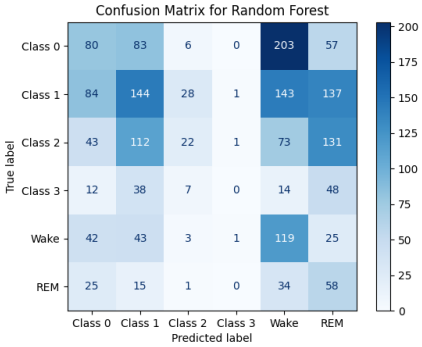
Figure 4: Normalized confusion matrices for deep learning models on the held-out folds. Rows correspond to true labels and columns to predicted labels.



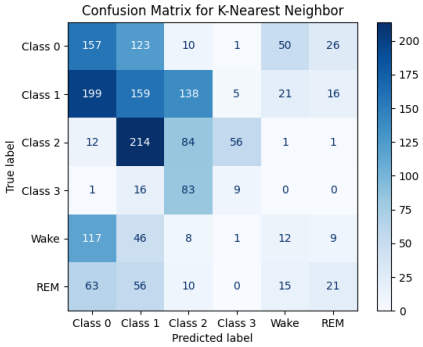
(a) SVM Confusion Matrix



(b) XGBoost Confusion Matrix

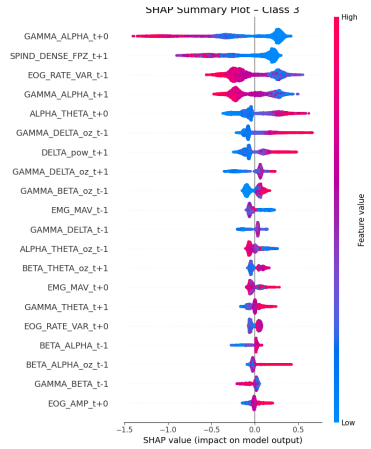


(c) Random Forest Confusion Matrix

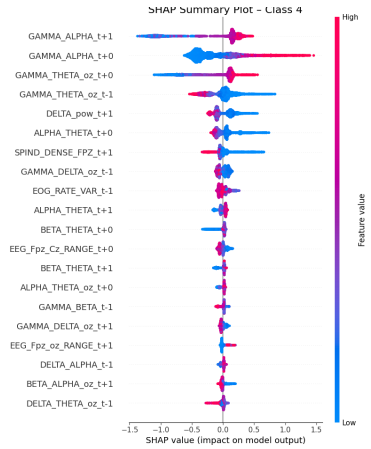


(d) KNN Confusion Matrix

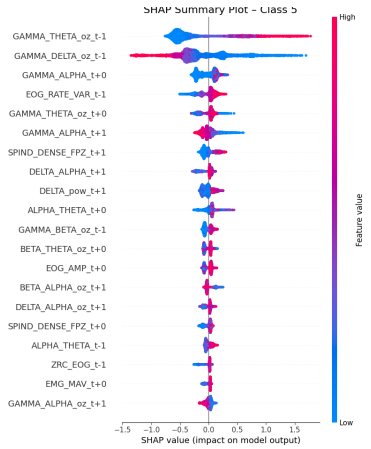
Figure 5: Confusion Matrices for Classical ML Models.



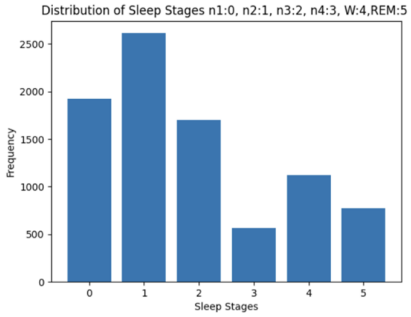
(a) SVM Confusion Matrix



(b) XGBoost Confusion Matrix



(c) Random Forest Confusion Matrix



Bar Plot Showing Distribution of Classes for 101 Patients

EEG Signal(Cenro-Parietal and Zero Midline)	EOG Signal	EMG Signal
EEG-Fpz-Cz Range,EEG-Fpz-oz Range, EEG-Fpz-oz Range,Sigma Spindle Density, Zero-Cross, $\frac{\alpha}{\theta}$ power-cz, $\frac{\gamma}{\delta}$ power-cz, $\frac{\gamma}{\beta}$ power-cz, $\frac{\gamma}{\alpha}$ power-cz, $\frac{\gamma}{\theta}$ power-cz, $\frac{\delta}{\beta}$ power-cz, $\frac{\delta}{\alpha}$ power-cz, $\frac{\delta}{\theta}$ power-cz, $\frac{\beta}{\alpha}$ power-cz, $\frac{\beta}{\theta}$ power-cz, $\frac{\alpha}{\theta}$ power-oz, $\frac{\gamma}{\delta}$ power-oz, $\frac{\gamma}{\beta}$ power-oz, $\frac{\gamma}{\alpha}$ power-oz, $\frac{\gamma}{\theta}$ power-oz, $\frac{\delta}{\beta}$ power-oz, $\frac{\delta}{\alpha}$ power-oz, $\frac{\delta}{\theta}$ power-oz, $\frac{\beta}{\alpha}$ power-oz, $\frac{\beta}{\theta}$ power-oz, γ power-cz, δ power-cz, β power-cz, α power-cz, θ power-cz, γ power-oz, δ power-z, β power-oz, α power-oz, θ power-oz,	EOG Rate Zero-Cross EOG EOG Amplitude EOG Rate Variance	Zero-Cross EMG EMG Variance EMG Max Amplitude Value EMG Amplitude

Features Extracted

Figure 6: Bar Plot and Feature Extracted

References

[1] ”Brain Basics: Understanding Sleep”

[2] Armon, C., Johnson, K., Roy, A., Talavera, F., Meyers, A., Alvarez, N., Murro, A., & Nowack, W. (2023). Polysomnography: Overview of polysomnography, parameters monitored, staging of sleep. *Medscape*.

[3] Back2Sleep. (2024, May 5). Polysomnography: Definition, indication and interpretation of results.

[4] Boe, A. J., McGee Koch, L. L., O'Brien, M. K., Shawen, N., Rogers, J. A., Lieber, R. L., Reid, K. J., Zee, P. C., & Jayaraman, A. (2019). Automating sleep stage classification using wireless, wearable sensors. *Npj Digital Medicine*, 2(1), 131.

[5] Charles River Laboratories. (n.d.). Quantitative EEG and EMG Services.

[6] Cleveland Clinic. (2023, June 19). Sleep Basics.

[7] Cleveland Clinic. (2023, February 10). Sleep study: What it is, what to expect, types & results.

[8] Cummings, J., & Sanders, L. (2014). *Introduction to Psychology*. Saskatchewan Open Educational Resources.

[9] Mayo Clinic. (n.d.). Polysomnography (Sleep study).

[10] National Heart, Lung, and Blood Institute. (2022, March 24). Sleep deprivation and deficiency - what are sleep deprivation and deficiency?

[11] Patel, A. K., Reddy, V., Shumway, K. R., & Araujo, J. F. (2024). Physiology, sleep stages. In *StatPearls*. StatPearls Publishing.

[12] Schlafgut. (2013). Hypnogram—One sleep cycle—PSG [Graphic]. Wikimedia Commons.

[13] Stanford Medical. (n.d.). Idiopathic Hypersomnia. Retrieved December 1, 2025.

[14] Suni, E., & Singh, A. (2021, December 2). Stages of sleep: What happens in a normal sleep cycle? *Sleep Foundation*.