Федеральное государственное автономное образовательное учреждение

высшего образования «Национальный исследовательский университет   
«Высшая школа экономики»

###### **Факультет компьютерных наук**

###### **ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

Классификация состояний мозга с использованием жидких нейронных сетей на EEG данных

Brain State Classification Using Liquid Neural Networks on EEG Data

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Науки о данных»

Студент группы МНОД231

Селиверстов Алексей Владимирович

(Ф.И.О.)

Руководитель КР

Сохраби Маджид

(Ф.И.О.)

Москва, 2025

**Table of contents**

[Brain State Classification Using Liquid Neural Networks on EEG Data 3](#_Toc196263919)

[Data 3](#_Toc196263920)

[Methodology 4](#_Toc196263921)

[Training setup 4](#_Toc196263922)

[Quality metrics 4](#_Toc196263923)

[Models 5](#_Toc196263924)

[Experimental Results 5](#_Toc196263925)

[Conclusions 9](#_Toc196263926)

[References 10](#_Toc196263927)

# Brain State Classification Using Liquid Neural Networks on EEG Data

## Data

This thesis is based on a publicly available dataset originally designed to explore cognitive responses—specifically confusion—in students viewing online educational video content. The data were collected under controlled experimental conditions, with the aim of studying how EEG signals reflect varying levels of cognitive load during learning.

The dataset includes recordings from ten college students, each of whom viewed a set of ten distinct educational videos. These videos were selected and categorized prior to the experiment into two groups:

* Non-confusing: Topics presumed to be familiar and readily understandable to the average student, such as introductory algebra or geometry.
* Confusing: Advanced topics such as quantum mechanics or stem cell research, selected for their potential to induce confusion in students unfamiliar with the material.

All students watched the same ten videos, evenly split between the two categories. Each video was approximately two minutes long, but only the central one-minute segment was used for EEG data analysis. The beginning and end of each clip were trimmed to minimize transitional and non-content-related cognitive responses.

EEG signals were recorded using a single-channel wireless EEG headset positioned to monitor activity over the frontal lobe. The headset recorded data using one electrode on the forehead and two reference electrodes placed near the ears. The device sampled neural activity every 0.5 seconds, resulting in approximately 120 EEG samples per video per participant.

Each EEG sample consists of power values across multiple standard brainwave frequency bands:

* Delta
* Theta
* Alpha1, Alpha2
* Beta1, Beta2
* Gamma1, Gamma2

These eight channels were the sole features used in this study. All other available information—such as participant demographics or specific details about the videos—was excluded from the analysis. The objective was to evaluate confusion purely from the perspective of neural signal data.

The primary label used in this thesis is the predefined confusion label, assigned based on the categorization of each video as either “confusing” or “non-confusing.” This classification was made independently of the participants' self-assessments and remains consistent across all viewers. The use of this label enables an investigation into whether confusion—operationalized at the video level—can be predicted from EEG features alone.

The dataset comprises over 12,000 EEG samples, corresponding to 100 individual recording sessions (10 participants × 10 videos). Each sample includes:

* A timestamp (0.5-second intervals)
* Eight EEG frequency band values (delta through gamma2)
* The predefined confusion label (binary)

These data were used to develop and evaluate predictive models aimed at distinguishing between confusing and non-confusing learning experiences based solely on patterns in brainwave activity.

## Methodology

### Training setup

     The models in this study were trained using only EEG-derived features—specifically, the power values across eight standard frequency bands: delta, theta, alpha1, alpha2, beta1, beta2, gamma1, and gamma2. No participant demographic data or video metadata were included. This design choice allowed for an isolated exploration of how neural signals alone relate to cognitive confusion.

Each EEG time series corresponds to a one-minute segment, sampled at 0.5-second intervals, resulting in from 120 up to 144-time steps per recording. To ensure consistency across samples, each sequence was padded to a fixed length of 144-time steps, using zero-padding at the end of each signal. This allowed all models to accept uniform input dimensions while preserving the temporal structure of the original signals.

The dataset was split into training and testing subsets using a 70–30 ratio. To ensure robust and reliable evaluation, multiple distinct and fixed data splits were created. All models were trained and evaluated on the same series of partitions, ensuring comparability across experiments while mitigating the effect of variance due to random sampling. Each model was trained for the same number of epochs, standardizing training duration across all experiments.

In some experiments, a lagged embedding of the EEG signal was applied to enrich the input with temporal dependencies beyond the original sampling rate. This involved concatenating delayed versions of the input signal to provide the models with access to past temporal context within a given time window.

### Quality metrics

Model performance was assessed using a combination of three evaluation metrics:

* Accuracy: Measures the proportion of correctly classified samples out of all predictions. It provides a general measure of model performance across classes.

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

* AUC (Area Under the Receiver Operating Characteristic Curve): Evaluates the trade-off between true positive rate (sensitivity) and false positive rate across different classification thresholds. A higher AUC indicates better model discrimination between the classes.

TPR (True Positive Rate) =

FPR (False Positive Rate) =

* Entropy: Used as a measure of prediction uncertainty. Lower entropy indicates higher confidence in the model’s prediction. For binary classification, entropy of a prediction with probability is calculated as:

where is the Shannon entropy, and logarithms are typically base 2 or natural logarithm depending on implementation.

All metrics were averaged across the multiple fixed splits to obtain reliable, variance-reduced estimates of model performance.

### Models

The predictive models employed in this work consisted of three major types of neural network architectures, each with multiple internal variations:

* GRU-based networks: These models leveraged Gated Recurrent Units, which are well-suited for capturing temporal dependencies in sequential data while maintaining a relatively compact architecture. Several variations were tested with different numbers of layers and hidden dimensions.
* LSTM-based networks: Networks built with Long Short-Term Memory units were also explored due to their effectiveness in retaining long-range dependencies and mitigating vanishing gradient issues. As with the GRU models, multiple architectures were tested.
* Liquid State Neural Networks (LSNNs): Inspired by biologically plausible models, these networks incorporate liquid time-constant dynamics, allowing them to adapt their memory and response patterns over time. LSNNs were included to investigate whether models with non-static internal dynamics could outperform more traditional RNN variants on EEG time series data.

In each category, a range of architectures was evaluated to explore the design space and identify the most effective configurations for confusion detection.

All hyperparameters—including learning rate, batch size, number of layers, and hidden unit size – were manually selected based on preliminary testing and domain knowledge. These settings were held constant across similar models within each experimental group to ensure fair comparisons.

### Experimental Results

**Baseline**

As a baseline, a fully connected (FC) network architecture was implemented based on a previous study. This model flattened the EEG signal across time and frequency dimensions before feeding it into a deep multilayer perceptron. The architecture consisted of five sequential dense layers with decreasing units (200 → 100 → 50 → 16 → 1), each followed by a ReLU activation function except for the output layer, which used a sigmoid activation for binary classification.

This FC model achieved an average validation accuracy of 0.63, consistent with the source paper. However, the model contained 54,983 parameters, and its inability to account for temporal dynamics limited its effectiveness when compared to time-series-oriented approaches.

**Implemented models**

To better capture the temporal structure of the EEG signals, three categories of neural network models were evaluated: Gated Recurrent Units (GRU), Long Short-Term Memory networks (LSTM), and Liquid State Neural Networks (LSNN). All models were trained exclusively on the raw EEG frequency band features (delta, theta, alpha1, alpha2, beta1, beta2, gamma1, gamma2) and used the same training conditions:

* Learning rate: 0.002
* Weight decay: 1e-7
* EMA (Exponential Moving Average): enabled
* Same number of training epochs and fixed data splits across models

Each model processed input sequences of 144-time steps per experiment, padded to a uniform length. Their architectures are described schematically below:

**GRU-based Model**

This model used a deep recurrent structure with four GRU layers and dropout applied before each:

* Input: (144 time steps, 8 EEG features)
* GRU(20 units, return sequences)
* GRU(20 units, return sequences)
* GRU(10 units, return sequences)
* GRU(10 units, return final)
* Dense(1 unit, sigmoid activation)

Performance:

* Accuracy: 0.873
* Loss: 0.278
* AUC: 0.945
* Parameters: 5,951

**LSTM-based Model**

Similar in depth and layout to the GRU model, this version replaced GRU cells with LSTM units:

* Input: (144 time steps, 8 EEG features)
* LSTM(20 units, return sequences)
* LSTM(20 units, return sequences)
* LSTM(10 units, return sequences)
* LSTM(10 units, return final)
* Dense(1 unit, sigmoid activation)

Performance:

* Accuracy: 0.853
* Loss: 0.337
* AUC: 0.940
* Parameters: 7,691

**LSNN-based Model**

This model used two LTC layers with adaptive time constants based on Neural Circuit Policies (NCPs):

* Input: (144 time steps, 8 EEG features)
* LTC layer configured with an AutoNCP topology of 64 neurons and 32 interconnections (return sequences)
* LTC layer configured with AutoNCP topology of 32 neurons and 16 interconnections (return final)
* Dense(1 unit, sigmoid activation)

Performance:

* Accuracy: 0.820
* Loss: 0.310
* AUC: 0.924
* Parameters: 27,105

This architecture leveraged the structure of NCPs to enhance temporal expressiveness while maintaining interpretability and compactness. Although the model had a relatively high parameter count, its performance remained competitive, suggesting that NCP-based LSNNs can effectively model cognitive states from EEG signals.

Overall, all three temporal models clearly outperformed the fully connected baseline across all evaluation metrics. Among them, the GRU-based model achieved the best performance while maintaining a relatively low number of parameters, making it an optimal choice in terms of both efficiency and accuracy. The LSNN, while having the highest parameter count, still demonstrated strong performance, indicating its potential for modeling complex temporal dynamics in EEG data.

**Temporal Embeddings for Enhanced Representation**

In addition to evaluating the models on raw EEG frequency-band signals, further experiments explored the integration of temporal embeddings to enrich the input representations. These embeddings were designed to expose temporal dynamics more explicitly by incorporating historical context into each time step.

The following types of lag-based embeddings were considered:

* Lag-1 embedding: Each timestamp was augmented with the feature values from the immediately preceding time step.
* Lag-2 embedding: The EEG signal at each timestamp was additionally augmented with the values from two steps before.
* Lag-1 + Lag-2 embedding: Both first and second-order lagged values were concatenated to each current timestamp, resulting in a broader temporal window for each input vector.
* First-order difference (Δlag-1): Rather than using raw lagged values, this embedding encoded the difference between the current timestamp and the one immediately before it, highlighting the change in brain activity over time rather than the absolute value.

These augmentations provided the models with richer temporal features, potentially aiding in the identification of subtle signal shifts indicative of cognitive states such as confusion.

**Convolutional Embedding**

In a parallel approach, temporal feature engineering was explored via causal convolution. A 1D convolutional layer with 50 filters, a kernel size of 3, ReLU activation, and causal padding was applied to the raw EEG signals. This method created new time-dependent embeddings that preserved the sequence’s chronological order while capturing short-term local patterns. The convolutional layer effectively summarized nearby temporal trends within the signal and passed the enriched representation to subsequent model layers.

To assess the impact of different temporal embeddings, performance metrics—accuracy, loss, and AUC—were monitored across training epochs for each embedding type and model architecture. These evaluations were conducted across GRU, LSTM, and LSNN-based models, allowing for a comprehensive comparison of embedding effectiveness. In the case of the best-performing LSNN architecture, not all embeddings were included in the final evaluation phase. This decision was based on earlier results, where certain embeddings consistently underperformed relative to others, and were therefore omitted to streamline experimentation.

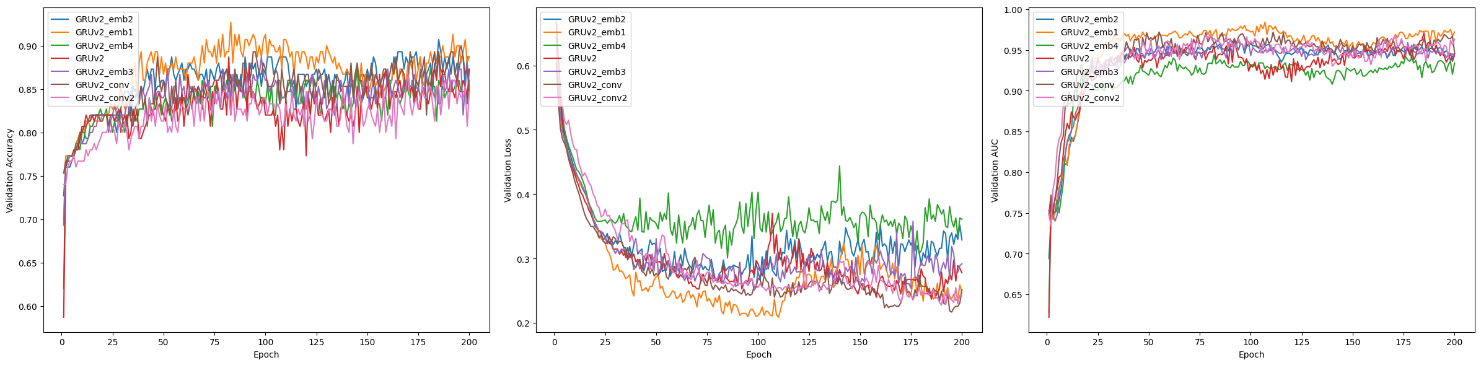


Figure 1. Best GRU performance

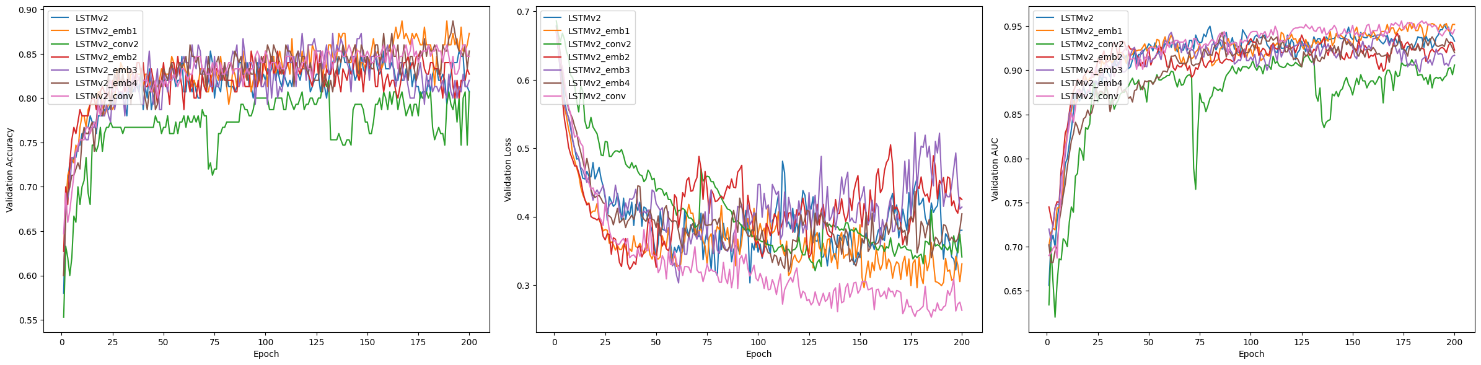


Figure 2. Best LSTM performance

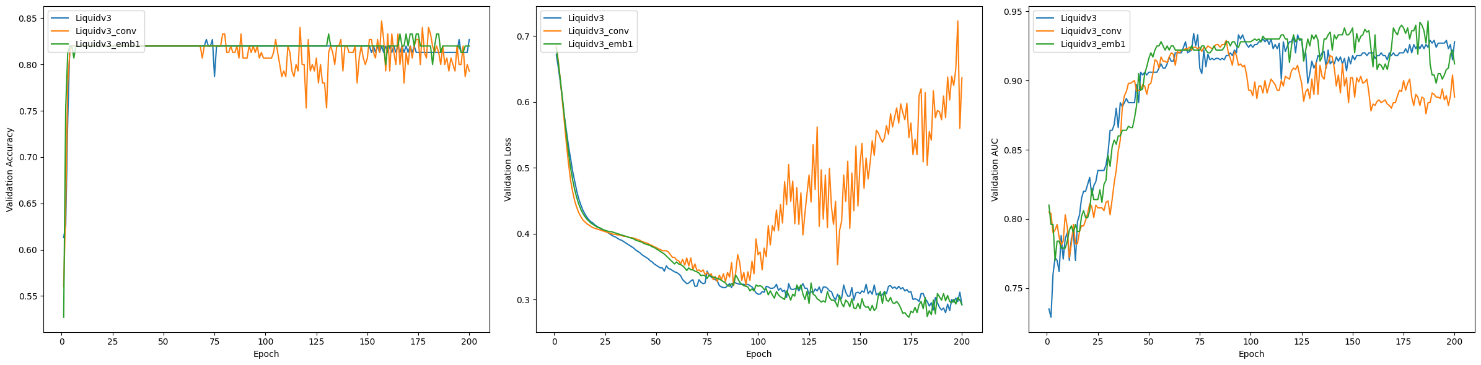


Figure 3. Best LNNS performance

Among all embedding strategies, the lag-1 embedding and the convolutional embedding consistently led to the most significant improvements across models. These two approaches not only increased predictive accuracy but also tended to stabilize the training process, reducing variance and preventing overfitting in several configurations.

The lag-1 embedding provided a minimal but meaningful temporal context, enriching the model’s ability to recognize transitions in brain signal patterns associated with cognitive states. Similarly, the convolutional embedding, by leveraging local temporal filters, effectively captured short-range dependencies that were often missed by raw features alone.

Overall, both of these embedding methods generally enhanced model performance, suggesting that even lightweight temporal transformations of EEG data can provide substantial benefits when predicting cognitive states like confusion.

**Summary of Final Model Performance**

To provide a comprehensive comparison, a summary graph is included that illustrates the final validation metrics—accuracy, AUC, and loss—for the best-performing configuration of each model type: GRU, LSTM, and LSNN. These models were selected based on their performance across all previous experiments, including variations with and without temporal embeddings.

The plotted results clearly show that GRU-based models significantly outperform the others across all metrics. Notably, the GRU model enhanced with lag-1 embedding achieved the best overall results, with accuracy exceeding 0.90 and AUC surpassing 0.95. This performance reflects a substantial improvement over the baseline fully connected model from the referenced paper, which reported an average validation accuracy of only 0.63.

While LSNN models (based on sparsely connected neural circuit policies) demonstrated competitive behavior, their overall performance was consistently lower than that of GRU and LSTM counterparts. Nevertheless, their results still validate the potential of biologically inspired recurrent architectures in EEG-based classification tasks.

These findings emphasize the importance of sequence-aware architectures and temporal embeddings when working with EEG signals. The GRU’s ability to model temporal dependencies effectively, especially when coupled with lightweight lagged context, appears to be particularly well-suited for confusion detection tasks using frontal EEG data.

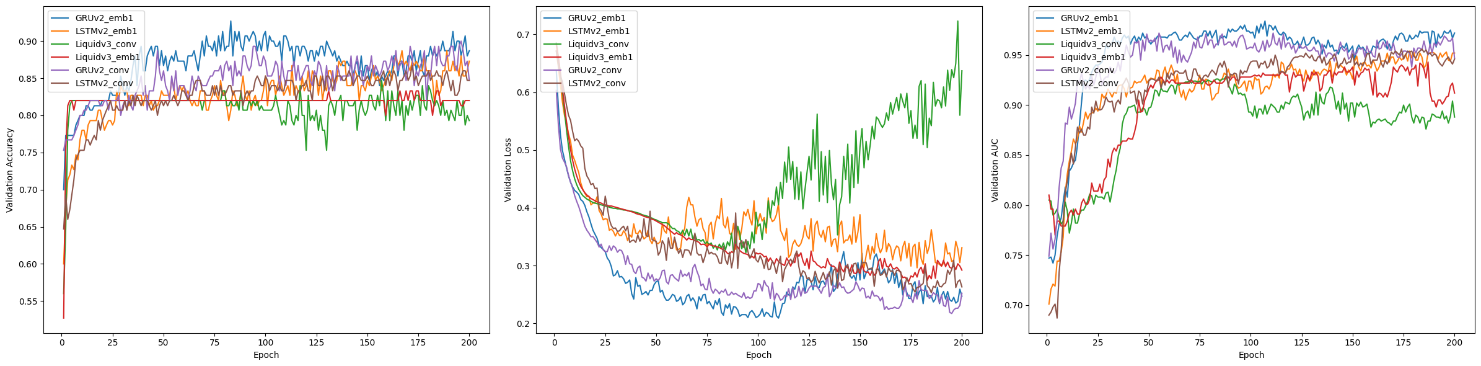


Figure 4. Final model performance

**Model Interpretability with SHAP Analysis**

To explore the interpretability of the trained models, I employed SHAP (SHapley Additive exPlanations) values, a widely adopted method for explaining predictions made by machine learning models. SHAP assigns an importance value to each feature for individual predictions, enabling insight into how specific EEG signals contribute to the final output.

In the SHAP summary plots, each point represents a feature contribution for a single sample. The size of the point corresponds to the absolute magnitude of the SHAP value, indicating the strength of that feature's influence. Red-colored points indicate positive SHAP values, meaning that the feature increased the model’s predicted probability of confusion, while blue-colored points reflect negative contributions, lowering the predicted probability.

To maintain clarity across signal types, the same EEG channel was consistently represented with the same shade of red or blue, allowing visual tracking of individual signal behavior across samples. This consistent color-coding helped highlight the relative importance of each frequency band (e.g., delta, theta, alpha1, etc.) throughout the dataset.

In experiments involving temporal embeddings, such as lagged values, I aggregated the SHAP values of the lagged features with their corresponding original features. This allowed for a more interpretable comparison across models and provided a clearer understanding of the role each signal type played in both its original and embedded forms.

A particularly striking observation from the SHAP summary plots is the presence of specific timestamps where all EEG signal features exhibit notably high importance. At these moments, the absolute SHAP values across all frequency bands increase sharply, suggesting that the model places disproportionate weight on certain temporal segments within the one-minute recording. This implies that cognitive confusion may manifest more strongly at specific points in the video, potentially aligning with moments of conceptual difficulty or topic transitions. While this observation is consistent across multiple samples and models, it warrants further investigation to verify whether these signal spikes consistently correspond to semantically confusing segments in the educational content itself.

Another consistent finding across experiments is the dominant importance of the theta band in predicting confusion. SHAP values frequently show that the theta signal contributes more significantly—both positively and negatively—to the model's decision-making process compared to other frequency bands. This trend is in agreement with prior research, which has linked increased theta activity in the frontal lobe with elevated cognitive load and attentional engagement, particularly in tasks that involve comprehension or problem-solving. The model's reliance on this band therefore provides additional validation for both the data and the model’s alignment with neuroscientific understanding of EEG signal interpretation.

While the GRU and LSTM models yielded interpretable patterns through SHAP analysis, LSNN models unfortunately did not provide useful insights into feature importance. The SHAP value distributions for LSNNs were generally less consistent, lacked clear structure, and often failed to highlight meaningful temporal or spectral patterns. This may be attributed to the architectural sparsity and internal dynamics of neural circuit policies, which, although biologically inspired, can obscure direct attribution of output predictions to individual input features. As a result, while LSNNs demonstrated reasonable predictive performance, their interpretability remains limited, reducing their utility for detailed signal-level or time-resolved analysis in this context.

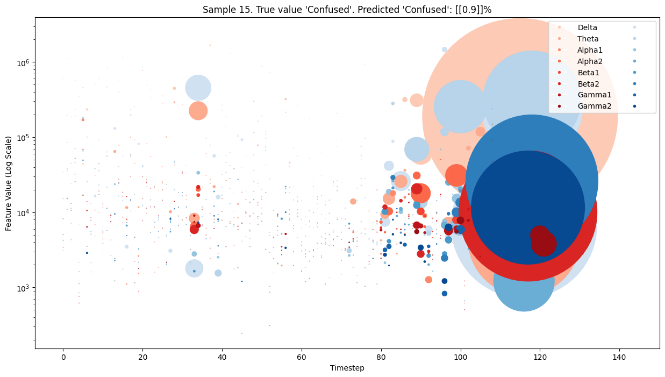
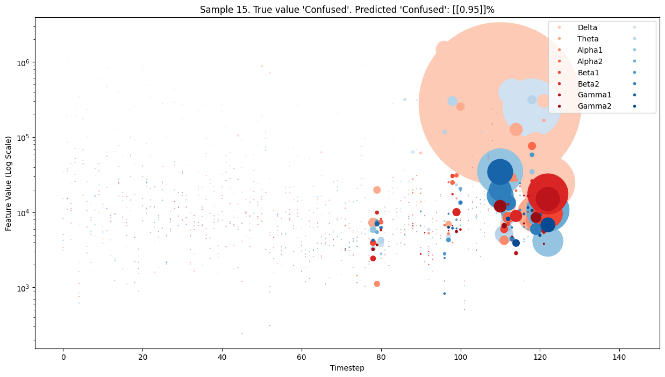
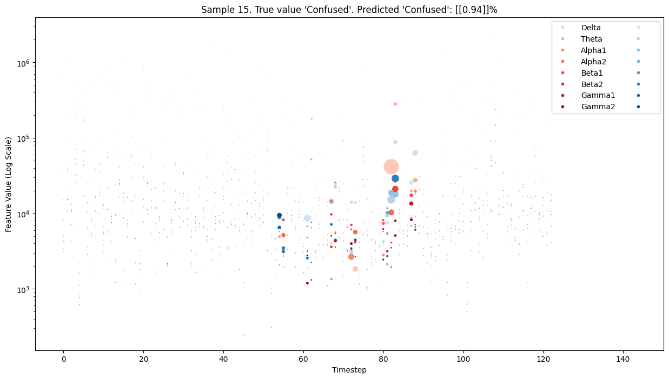
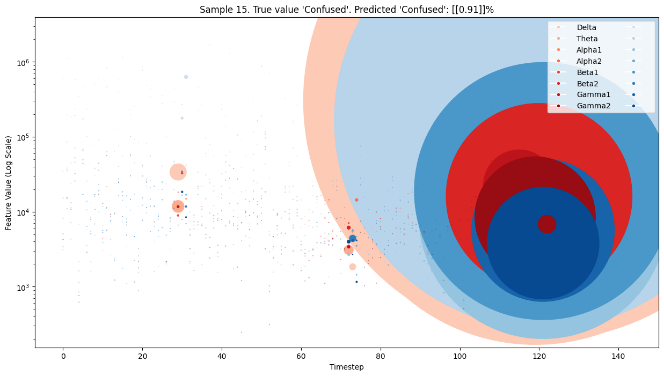
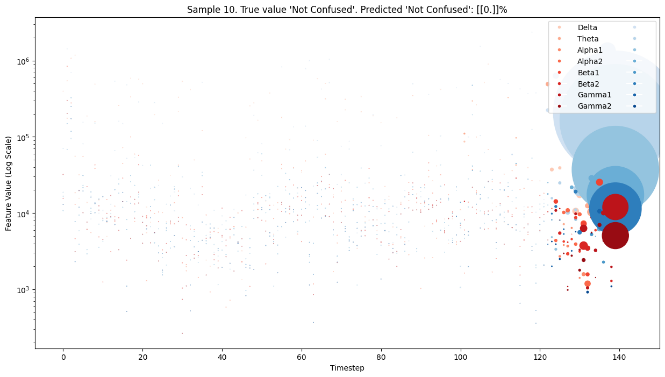
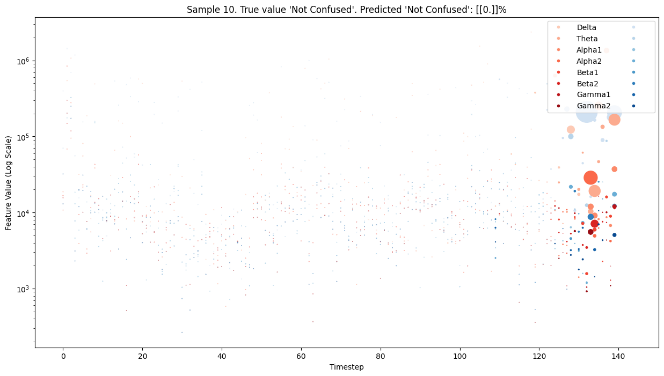
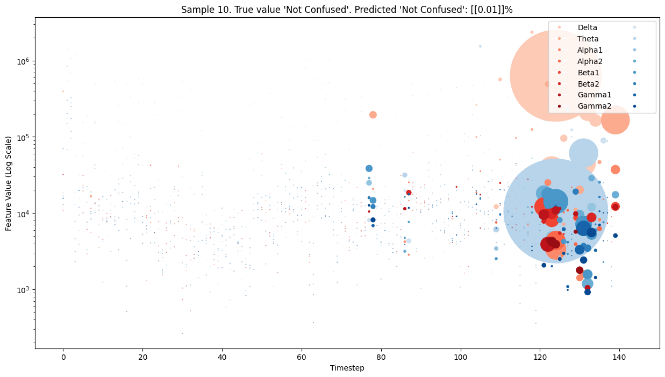
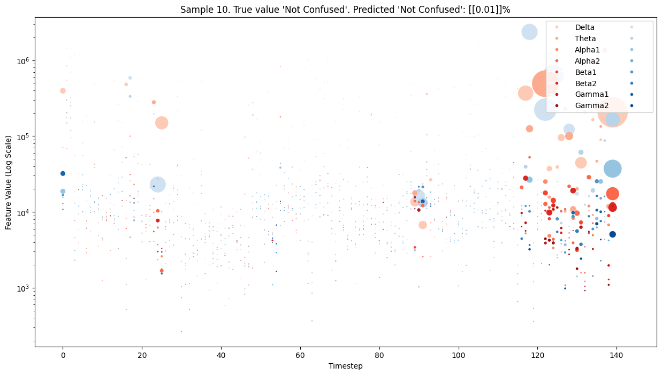
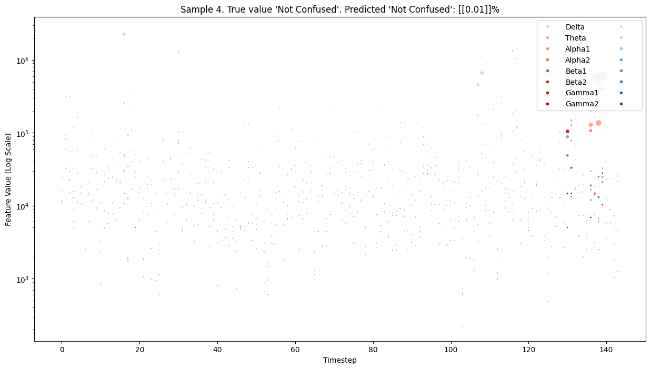
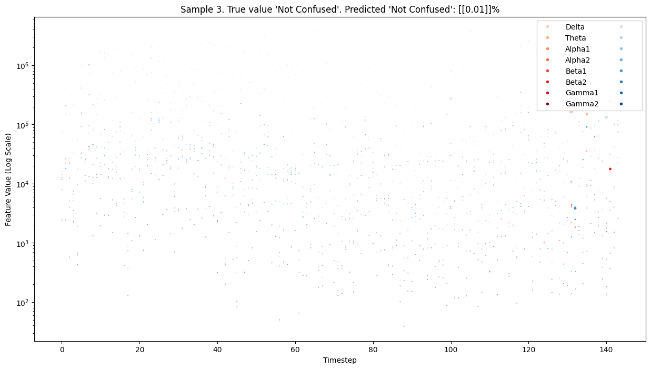
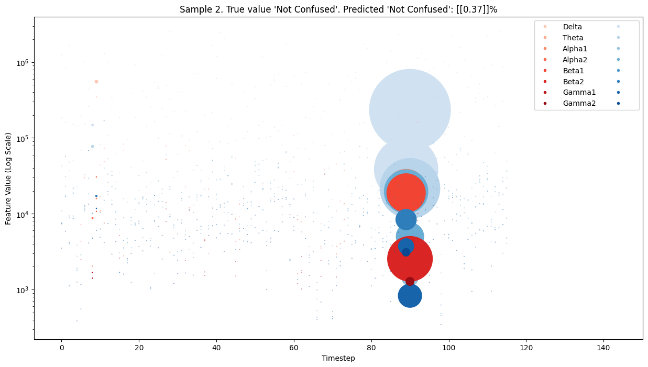
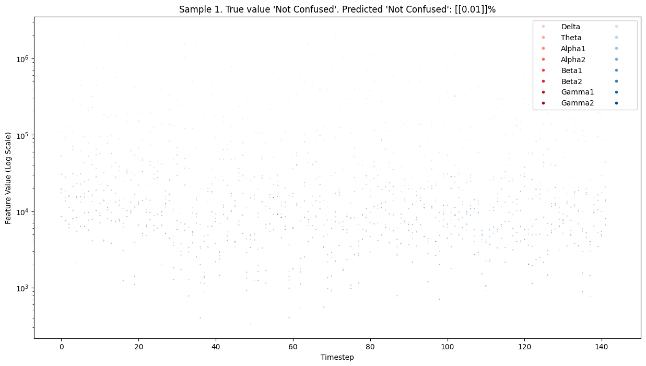
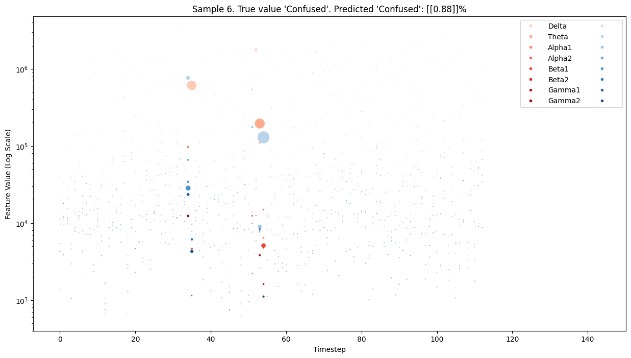
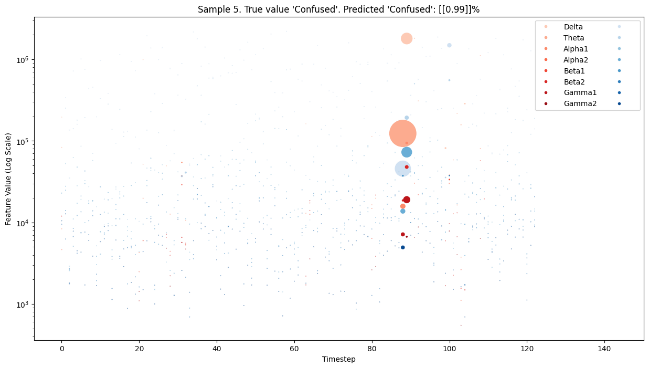


Figure 5. SHAP for Samples 10 and 15 for GRU and LSTM with lag-1 embeddings and convolution





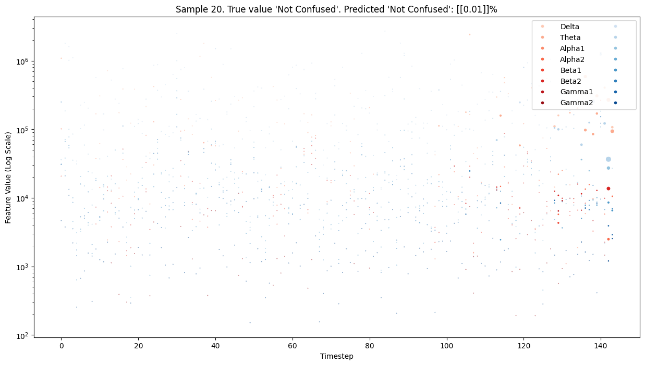
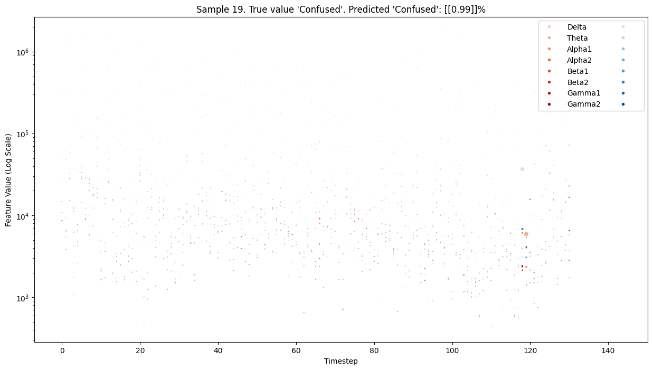
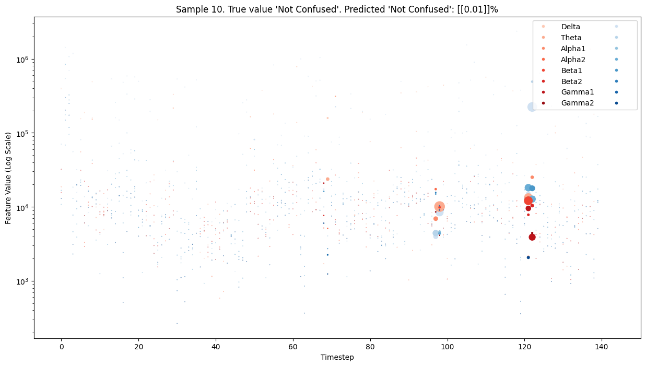
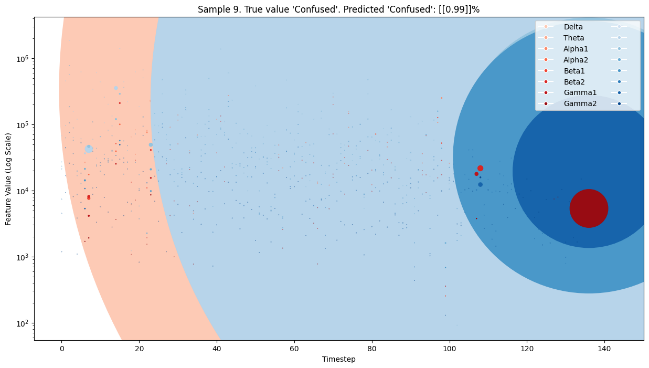


Figure 6. SHAP for GRU with embeddings (best model)

## Conclusions

## References

1. **Ayoub, O., Andreoletti, D., Knapińska, A., Goścień, R., Lechowicz, P., Leidi, T., Giordano, S., Rottondi, C., & Walkowiak, K.** (2024). Liquid Neural Network-based Adaptive Learning vs. Incremental Learning  for Link Load Prediction amid Concept Drift due to Network Failures. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2404.05304>
2. **Bidollahkhani, M., Atasoy, F., & Abdellatef, H.** (2023). LTC-SE: Expanding the potential of Liquid Time-Constant Neural Networks for scalable AI and embedded systems. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2304.08691>
3. **Hasani, R., Lechner, M., Amini, A., Liebenwein, L., Ray, A., Tschaikowski, M., Teschl, G., & Rus, D.** (2022). Closed-form continuous-time neural networks. Nature Machine Intelligence, 4(11), 992–1003. https://doi.org/10.1038/s42256-022-00556-7
4. **Hasani, R., Lechner, M., Amini, A., Rus, D., & Grosu, R.** (2021). Liquid Time-constant Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 35(9), 7657-7666. <https://doi.org/10.1609/aaai.v35i9.16936>
5. **Huang, Z., Contreras, L. F. H., Leung, W. H., Yu, L., Truong, N. D., Nikpour, A., & Kavehei, O.** (2024). Efficient Edge-AI models for robust ECG abnormality detection on Resource-Constrained hardware. PubMed. <https://doi.org/10.1007/s12265-024-10504-y>
6. **Lechner, M., Hasani, R., Amini, A., Henzinger, T. A., Rus, D., & Grosu, R.** (2020). Neural circuit policies enabling auditable autonomy. Nature Machine Intelligence, 2(10), 642–652. <https://doi.org/10.1038/s42256-020-00237-3>
7. **Lechner, M., Hasani, R. M., & Grosu, R.** (2018). Neuronal circuit policies. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.1803.08554>
8. **Nerrise, F., Sosanya, A. S., & Neary, P.** (2024). Physics-Informed Calibration of Aeromagnetic Compensation in Magnetic Navigation Systems using Liquid Time-Constant Networks. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2401.09631>
9. **Nye, L.** (2023). Digital twins for patient care via knowledge graphs and Closed-Form Continuous-Time liquid neural networks. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2307.04772>
10. **Yang, J., Shi, R., & Ni, B.** (2021). MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 191-195). Nice, France. doi: <https://doi.org/10.1109/ISBI48211.2021.9434062>
11. **Zheng, Z., & Jia, X.** (2023). Complex Mixer for MEDMNIST Classification Decathlon. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2304.10054>
12. **Zhu, F., Wang, X., Huang, C., Jin, R., Yang, Q., Alhammadi, A., Zhang, Z., Yuen, C., & Debbah, M.** (2024). Robust Continuous-Time Beam Tracking with Liquid Neural Network. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2405.00365>