Bridging AI and Health on Time Series Analysis and Explainability Using the Case Study of EEG Channel Selection Problem

Vandana Srivastava[1,2], **Biplav Srivastava**[1]

[1]Artificial Intelligence Institute, University of South Carolina
[2]University Libraries, University of South Carolina
{vandana, biplav.s}@sc.edu

Abstract

Time series (TS) analysis is an active application area for Artificial Intelligence (AI) methods where the objective is to analyze numeric quantities indexed by time for tasks like classification, forecasting, and abnormality detection. In health, TS manifests as biosignals like the electroencephalogram (EEG), where electrical signals from the brain are analyzed. AI and health communities can tremendously benefit each other in TS, with the former offering advanced analytical methods while the latter provides complex data sets and trust-sensitive use cases. But the communities also need to overcome confusing terminologies, hidden assumptions, and a lack of necessary domain contexts for result evaluation and interpretation. In this paper, we attempt to bridge the gap using the problem of channel selection in EEG. We outline challenges in working with EEG data, demonstrate via two experiments how simple explainable AI (XAI) methods can be quite effective for channel selection irrespective of EEG tasks/paradigms, and argue that recent TS trends in AI, like LLMs and XAI methods, can benefit health as well. We hope that this work will bring researchers working on TS problems at the intersection of AI and health closer to work in AI trustworthiness so that they can better leverage results from their respective areas to overcome common challenges. All code and resources are released on GitHub to help others replicate.

Code — https://github.com/vsrivas/XAI-EEG-Analysis

Introduction

In time series (TS) analysis, the objective is to analyze numeric quantities indexed by time for tasks like classification, forecasting, and abnormality detection (Hamilton 1994). TS is an active application area for Artificial Intelligence (AI) methods and has wide real-world potential in domains as diverse as finance, power, water, weather, and health. In health, biosignals, such as the electroencephalogram (EEG), are an example of TS, where electrical signals from the brain are analyzed. Biosignals are distinguished by having multiple channels (variables) capturing the same phenomenon simultaneously but from different biological regions. Although the data volume may be large, it is often characterized by missing values, redundant data, non-linear dependencies, and requires high computing resources to process.

Researchers in TS today are exploring new methods like deep learning (Elsayed et al. 2021) and Large Language Models (LLMs) (Zhang et al. 2024) for improving performance and explainability (XAI) (Rojat et al. 2021) to help users understand AI output and gain their trust for wider-scale adoption. AI and health communities can benefit each other tremendously in TS, with the former offering advanced analytical methods while the latter providing complex data sets and use cases. Furthermore, to build user trust with their models, they need to employ AI trustworthiness techniques like explainability. But the communities also need to overcome confusing terminologies, hidden assumptions, and a lack of necessary domain contexts for result evaluation and interpretation.

In this paper, we attempt to bridge the gaps using the problem of channel selection in EEG. Channel selection helps answer whether all the channels, interpreted as features, should be analyzed or some can be dropped and thus, data reduced with minimal loss of accuracy for increased analysis efficiency. We outline challenges in working with this data modality and tackle two binary classification problems: resting state in LEMON (Babayan, Erbey, and Kumral 2019) and mental arithmetic in EEGMAT(Zyma et al. 2019). Seen together, we demonstrate a drastic reduction in channels needed (only 3 needed v/s all; $\geq 85\%$) while retaining performance ($\geq 91\%$) measured on two metrics, AUC and accuracy, while saving over 74% processing time. We release all code and resources on GitHub to help others replicate. Building on the case study, we outline how general TS trends can benefit health if the latter were to adopt recent LLMs and XAI trends. Furthermore, EEG data is more complex than normal TS data used by AI and XAI researchers in TS. Hence, adopting EEG for AI research and evaluation will generalize their methods for even wider applications. We, thus, hope that this paper will bring researchers working on TS problems at the intersection of AI, XAI, and health to better leverage results from their respective areas to overcome common challenges.

In the rest of the paper, we start with preliminaries and related work, followed by describing two EEG datasets and how we performed a classification task using four methods. We demonstrate how to use both whitebox (also called glassbox) and blackbox XAI methods to select very few channels without compromising performance. Finally, we dis-

cuss how new emerging methods may be used and conclude.

Background and Related Work

In this section, we provide necessary background and discuss related prior work in the areas of time series, bio-signals and explainability.

Analyzing Time Series

Let a time-series be represented by $\{x_{t-n+1}, x_{t-n+2}, ..., x_t, x_{t+1}, ..., x_{t+d}\}$, where each x_{t-n+i} represents a numeric quantity, like voltage or stock price, t represents time instant of interest and n and d are parameters. n is called the sliding window size. If instead of a single variable (quantity) x we have a set of two or more *inter-related* variables X, this becomes a multivariate TS data (Mendis, Wickramasinghe, and Marasinghe 2024). We now describe two prominent TS tasks - forecasting and classification; see a more comprehensive discussion on TS in (Hamilton 1994) and on tasks in (Shyalika, Wickramarachchi, and Sheth 2024; Zhang et al. 2024).

For **forecasting** TS task, d is the number of future quantities a (learning) model predicts. In this sequence, let $X_t = \{x_{t-n+1}, x_{t-n+2}, ..., x_t\}$, and let $\hat{Y_t} = \{x_{t+1}, x_{t+2}, ..., x_{t+d}\}$, where $\hat{Y_t} = f(X_t)$ i.e., based on the quantity's value at previous n timesteps, let the function f represent the predicted value for the next d timesteps. Let Y_t denote the true quantity values for the next d timesteps. Examples of representative works in TS forecasting are (Muppasani et al. 2022; Lakkaraju et al. 2024; Mendis, Wickramasinghe, and Marasinghe 2024).

As a **classification** TS problem, we are given a schema F consisting of a list of features $F = \{f_1, .. ,f_m\}$ that a sensor is able to capture. They are also called columns in the data. Adopting the notations of (Goodfellow, Bengio, and Courville 2016) for a classification problem, $x \in \mathbb{R}^n$ can be seen as a collection of observations. Each observation, or row in the data, has a structure consistent with F. The classification or state identification problem is to produce a function $f : \mathbb{R}^n \rightarrow \{1, ..., k\}$. When $s = f(x)$, the state corresponding to x is a number capturing the category assigned to the observation. Examples of representative works in TS classification are (Muppasani et al. 2024).

There are many methods proposed for addressing TS tasks, including rule-based, classical machine learning based and ensemble-based (Elsayed et al. 2021). Although deep learning based (Wen et al. 2022; Mendis, Wickramasinghe, and Marasinghe 2024) and LLM-based (Zhang et al. 2024) methods are the latest rage, classical machine learning methods, especially ensemble-based, have still been found quite competitive in many settings (Elsayed et al. 2021).

Analyzing Biosignals

Biosignals are physiological signals generated by living beings that can be measured continuously in time(Fuentes-Aguilar, Pérez-Espinosa, and de-la Cruz 2022). They can be electrical (like Electroencephalogram: EEG and Electrocardiogram: ECG) and non-electrical (like Magnetoencephalogram: MEG). Biosignals provide valuable information about the different physiological processes taking place inside the body. AI has played an important role in identifying hidden patterns and markers from these signals, which was not possible for the clinicians to do with the naked eye (Dukyong et al. 2020). In this paper, we will discuss the EEG biosignal as an exemplary use case for the application of XAI on time-series data.

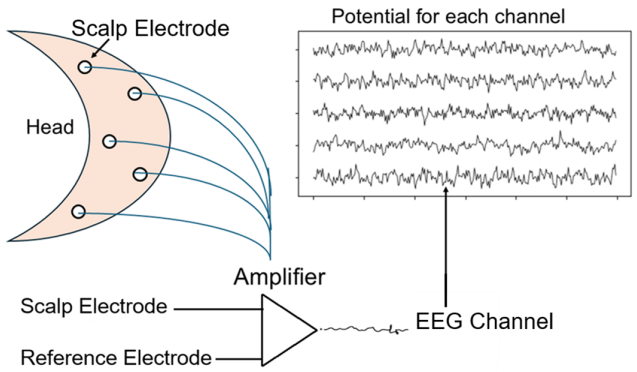

Figure 1: EEG recording setup: Capturing neural activity via scalp electrodes and recording through channels using a differential amplifier.

EEG as a time series. EEG is a biosignal to capture electrical activity inside the brain. Tiny electrodes, attached to a wire, are placed on the scalp at special positions. The electrodes detect small charges generated by neuronal activity in the brain and record them when presented with a specific cognitive activity *(tasks or events)*. EEG systems use a differential amplifier to produce each *channel* or trace of activity at a spatial region (Nagel 2019; Smith 2024)

The names of EEG channels are based on their placement on the head (like F for frontal, C for central, T for temporal, etc.) and channel number (even numbers for right hemisphere location and odd numbers for left brain hemisphere). The commonly used international standard for channel placement is called *10-20 system* (W. 2009).

EEG is either carried out to study neurological conditions/diseases or to study the brain response for specific tasks. If a set of multiple tasks is created and a participant has to switch between these tasks in a systematic way then it is called a *"paradigm"* (I et al. 2010). For example, mental math and digit-span are tasks, and *oddball experiment* is a paradigm (I et al. 2010). EEG signals are generated using multiple channels ranging from 16 (minimum number according to American Clinical Neurophysiology Society guidelines (SR et al. 2016)) to 128 (high resolution for complex analysis), each channel being one time-series. Hence, the complexity of EEG data increases with an increase in channels, making channel selection an important aspect of EEG analysis. Also, there are some other properties of EEG that make its analysis very challenging, such as non-stationary and noisy signals, nonlinearity, and high inter-subject variability.

Explanation Methods

AI methods, especially those based on learning, deep learning, and LLMs, face a major challenge in proving their trustworthiness regardless of the mode in which they take input - numeric, timeseries, textual, audio, visual, or multimodal. They are notorious for their fragility (lack of robustness) and other characteristics (e.g., opaqueness, alignment to human values) that go beyond performance to contribute to users' trust of technology(Varshney 2022). For example, small variations in the inputs to a Machine Learning (ML) model may result in drastic swings in its output. This uncertainty about robustness is amplified by the lack of interpretability of many ML models due to their black-box nature (Longo et al. 2020). As a result, such systems face challenges in gaining acceptance and trust from end-users, hampering their widespread adoption.

When such systems are deployed in critical areas like healthcare (Asan, Bayrak, and Choudhury 2020) and finance (Boukherouaa et al. 2021), the consequences of their uncertain behavior could cause critical failures. Some promising ideas to manage user trust in time series are to (1) *explain* model behavior (also called XAI methods) (Rojat et al. 2021; Nori et al. 2023) and (2) *communicate* the behavior of AI systems through ratings that are assigned after assessing AI systems from a third-party perspective (without access to the system's training data). The rating methods (Lakkaraju et al. 2024; Srivastava and Rossi 2018, 2020; Srivastava et al. 2024; Lakkaraju, Srivastava, and Valtorta 2024) are particularly valuable to users to make informed decisions when *a choice of AI models* is available for a task to decide from.

Channel Selection

An illustrative and important problem in EEGs is channel subselection, where one needs to decide if all the channels, interpreted as features, should be analyzed or some can be dropped, and thus, the data reduced with minimal loss of accuracy and increased analysis efficiency. The advantages of selecting a lower number of channels are that it can lead to lower power requirements and a lower chance of overfitting (Yulan and Wenshan 2024).

For biosignals like EEGs, researchers have started to look at XAI methods (Apicella et al. 2022). For channel selection, (Choel-Hui et al. 2025) discusses variants of LIME and SHapley additive exPlanation (SHAP) explainability methods. In this paper, we look at whitebox and blackbox XAI methods as available in the InterpretML tool(Nori et al. 2023). In (Yulan and Wenshan 2024), the authors discuss channel selection in the context of EEG and epilepsy classification. They use the CHB-MIT dataset, LSTM for channel selection, and SHapley additive exPlanation (SHAP). The paper also gives good motivations and background details. In (Ahmad et al. 2024), authors consider EEG for epilepsy classification, but do not focus on channel selection. They consider Bonn and UCI-EEG datasets datasets Bagged Tree-based classifier (BTBC), and SHapley Additive exPlanation (SHAP) for explanations.

A Case Study in Choosing and Explaining Selection of EEG Channels

We wanted to explore the following questions:

1. Can XAI methods help in deciding and selecting a minimum number of channels needed to get a (reasonably) high performance for an EEG task?
2. Are the selected channels physiologically relevant in the context of the EEG task?

We will use two publicly available EEG datasets, LEMON (Babayan, Erbey, and Kumral 2019) and EEGMAT (Zyma et al. 2019) from physionet (Goldberger et al. 2000).

Dataset 1: LEMON

In the LEMON dataset, we chose the preprocessed EEG data of 39 participants. Each preprocessed *.tar.gz* EEG file was approximately 55 MB. The 62-channel EEG (61 scalp electrodes) of duration 16 min was recorded at rest at a sampling rate of 2500Hz. The EEG session had 16 events, 8 eyes closed (EC) and 8 eyes-open(EO), each 60 seconds long. The EO and EC events took place alternatively, with recording starting at the eyes-closed condition. During the EO session, the participants were seated in front of a computer screen and asked to stay awake with their eyes focused on a black cross displayed on a white background. During preprocessing, the EEG data were downsampled from 2500Hz to 250Hz, bandpass filtered within 1-45Hz (8th order, Butterworth filter), and split into EO and EC conditions for further analyses. The final dataframe for modeling had dimension (9,151,114 x 39) with 4,539,017 instances for 'EO' and 4,612,097 for 'EC.'

We call the machine learning task here as of *classifying for resting states with labels of EO and EC*. We read the EEG data for both EO and EC events and found that there were only 38 channels (CP2, C5, FC6, Oz, F6, CPz, PO4, Pz, FT7, C2, O2, PO8, P5, F4, AF4, C3, P1, F1, C4, P8, PO3, FC3, AFz, FC5, P4, CP4, F2, TP8, CP1, POz, P6, P2, F3, P3, AF3, FC4, CP5, CP3) that were common in all the subjects. We combined the data for all 39 participants to create a time-series binary classification problem with channels as 'features' and condition 'EO' and 'EC' as two target labels.

Dataset 2: EEGMAT

The EEGMAT dataset (Zyma et al. 2019) contains EEG recordings of 36 participants before and during mental arithmetic tasks. The 23-channel EEG recordings, sampled at 500Hz, were preprocessed using a high-pass filter with a 30Hz cut-off frequency and Independent Component Analysis (ICA) for removing artifacts. All recordings were 60 seconds long in separate EEG files with code '1' or '2'; 1 for recording before the mental math task and 2 for recording during the mental math task. Each *.edf* recording for state '1' was 3.7 MB, and '2' was 1.3 MB.

We call the machine learning task here as of *classifying for mental arithmetic activity with labels of 1 and 2*. We merged the two recordings for all subjects based on the 20 common channels (T3, T5, O1, Cz, C4, Pz, T4, Fp2, F8, F7, O2, A2-A1, P3, P4, F4, Fz, C3, T6, F3, Fp1) with labels

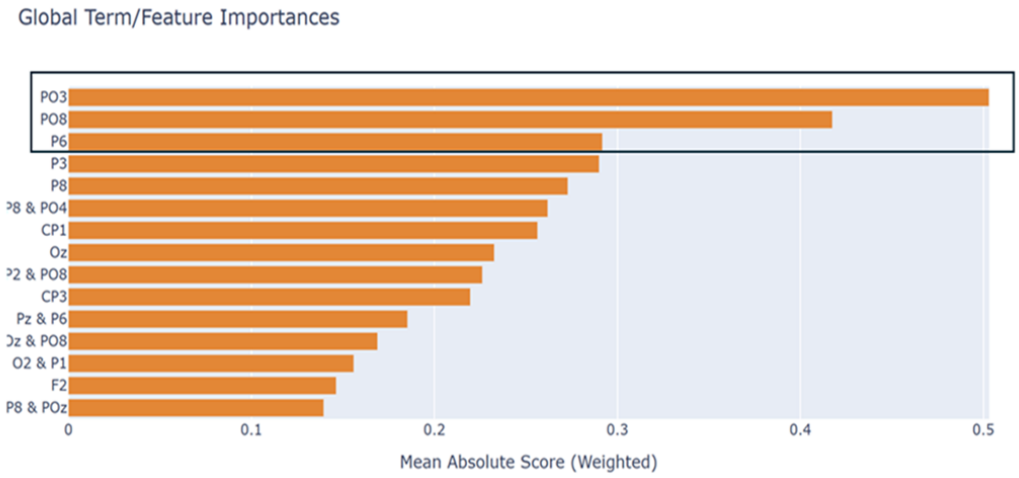

Figure 2: Feature importance summary for Glassbox Explainable Boosting Method for LEMON data

"1" or "2". The final dataframe for modeling had dimension (4,338,000 x 21) with 3,222,000 instances for '1' and 1,116,000 for '2.'

Methods

For both binary classification problems (resting state in LEMON and mental arithmetic in EEMGMAT), the channel data were used as input features and state 'EO'/'EC' and '1'/'2' as target labels. We used the *mne* python package to read the data and convert it to a pandas dataframe. We did not consider "time" or subject id "sub" as an input feature to the model. We used high-performance computing (HPC) clusters during model creation.

We created explainable classification models (Glassbox-Explainable Boosting Method (EBM), Logistic Regression, and Decision Tree) for both data using the package *interpretML* to predict the EEG labels. InterpretML is an open-source python toolkit for developing explainable models. For comparison, we also used the Blackbox Logistic Regression method.

First, we developed prediction models using ALL the channels as input and computed the ROC AUC score/accuracy for each model (see Table 1). We observed that EBM had the highest AUC score among all the methods. The feature importance summary of EBM showed that the top three channels that contributed most to the model prediction in LEMON data were PO3, PO8, and P6 (Figure 2), and in EEGMAT were FP1, FP7&FP1, FP3&FP1 (Figure 3). We repeated the model creation process using the top 3 channels (PO3, PO8, and P6 for LEMON and FP1, FP7, FP3 for EEGMAT) and the top channel (PO3 and FP1 for LEMON and EEGMAT, respectively). The model performances are given in Table 1.

For modeling, we used 70% of the data for training and 30% for testing. The raw channel values were used as input, and we did not create any new features. The models were created using their default configuration and without any hyperparameter tuning in any of the classification methods, as our goal was not to find the most efficient model but to see the impact of reducing channels identified by explainable methods.

Results

We first created a model using ALL channels. The top-3 channels are identified by the feature importance summary (FIS) chart and feature importance value (FIV). The FIS chart displays "which" features have a significant impact on the target label, and FIV determines "how" much the impact is on the target label. We will refer to the blackbox Logistic Regression method as BLR, explainable Logistic Regression as ELR, and explainable Decision Tree as EDT in the following section. We compared both AUC score and accuracy for models in case 2 ("Top-3") and case 3 ("Top-1") with case 1 ("All") by dividing the case 2 and case 3 values from case 1 values and reporting the result as %.

Resting State Classification - LEMON Dataset

1. Case 1 ("All"): In the first round, we created models using *all* the channels. The AUC scores were between .497 - .695, the lowest .497 for both BLR and ELR methods, and highest for EBM (.695). Hence, we chose EBM to identify the top-3 channels. (Figure 2) using FIS and FIV.
2. Case 2 ("Top-3"): The top-3 channels (or features) by EBM using FIS were found to be PO3 (.503), PO8 (.417), and P6 (.292). The FIV values are in brackets. We created four models again, now using just the top-3 channels. We observed that the new AUC scores were .493 for both BLR and ELR, which was 99.1% of the original AUC score with all channels. Similarly, the new AUC score for EBM was 94.5% and for EDT was 100% of case 1. Also, the accuracy values for models with top-3 channels were 100% for ELR and EDT, 96.8% for EBM, and 128% for BLR.(Figure 2)
3. Case 3 ("Top-1"): With PO3 as the only channel, the AUC score and accuracy of EBM reduced to 85.6% and 89%, respectively in comparison to case 1. The AUC

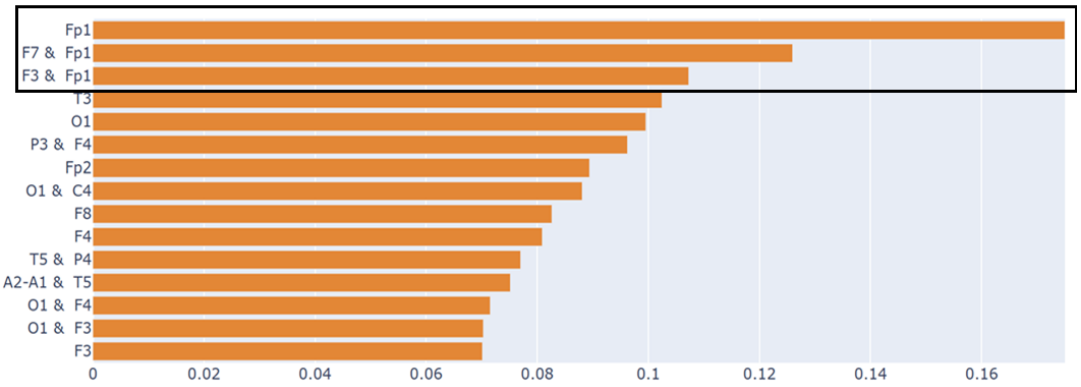

Figure 3: Feature importance summary for Glassbox Explainable Boosting Method for EEGMAT data

score for ELR and BLR was 99.6% and 100%, respectively, of case 1, and the accuracy was equal to case 1. The AUC and accuracy of EDT were found to be 98.4% and 98.3%, respectively, in comparison to case 1.

Mental Arithmetic Classification - EEGMAT Dataset

1. Case 1 ('All'): In first round, we created models using *all* the channels. The AUC scores were between .500 - .664, the lowest .500 for both BLR and ELR methods and highest for EBM (.664). Hence we chose EBM to identify the top-3 channels. (Figure 2) using FIS and FIV.
2. Case 2 ('Top-3'): The top-3 channels (or features) by EBM using FIS were found to be FP1 (.175), F7&FP1 (.126), and F3&FP1 (.107). The FIV values are in brackets. The new models were created using the top-3 channels as FP1, F7, and F3. We observed that the new AUC scores were .499 for both BLR and ELR, which was 99.8% of the original AUC score with all channels. Similarly, the new AUC score for EBM was 91.4% and for EDT was 98.7% for case 1. The accuracy values for models with top-3 channels were 100% in case 1 for all the models.(Figure 2)
3. Case 3 ('Top-1'): With FP1 as the only channel, the AUC score and accuracy of EBM reduced to 83.1% and 98.6%, respectively, in comparison to case 1. The AUC score for both ELR and BLR was 99.6% for case 1, and the accuracy was equal to case 1. The AUC and accuracy of EDT were found to be 97.9% and 100%, respectively, in comparison to case 1.

Summary

Reviewing the results in Table 1, we find that **three channels** are usually sufficient to achieve classification results comparable to those with all the channels when measured with accuracy and AUC. In both datasets, three channels led to the best performance ($\geq$100% accuracy) except for EBM, and on AUC, the performance was $\geq$ 91% of that with all channels.

When we compared the running time of the programs for the above three cases (Table 2), we found that the running time for all models on LEMON dataset with all channels was 230 mins, with three channels was 18.95 mins and 10.79 mins with one channel representing at most 8.2% ($\approx$ 92% reduction) and 4.7% ($\approx$ 95% reduction) of time off all-channel results, respectively. For the EEGMAT dataset, it took 56.41 minutes to run the program with all channels in comparison to 14.81 min (74% reduction) with three channels, and 8 minutes (92% reduction) with one channel. The programs were run on High Performance computing clusters (HPC) with a single node of 24 cores and 24gb memory per CPU.

We thus see that XAI methods can be used to make EEG processing more efficient. Furthermore, emerging trends in AI and XAI, discussed next, have the potential to transform EEG processing drastically if they could be easily applied.

Discussion: Leveraging Synergies

We now discuss areas of synergy between AI trends in TS and biosignal analysis in health.

Complex Data Despite being a time-series data, EEG signals differ from regular time-series like weather or stock prices, in many ways. These differences arise from the following properties of EEG (Roy et al. 2019):

- Noisy: EEG signals contain a lot of noise due to heartbeat, sweat, head movement, jaw clinching, eye movement, etc.
- Non-stationary: EEG signals are non-stationary, so their statistical properties change over time. Classifiers trained on a certain duration of data may perform poorly on data at a different time interval.
- High inter-subject variability: The two EEG signals from two different individuals are distinct due to physiological differences between individuals. This leads to poor performance when generalizing the models.
- Volume conduction: Due to the electrical conduction spread, more than one channel could pick up the signal from the same area of the brain.
- Multivariate time-series: Each channel generates a time-series for a different brain region, generating a multivariate dataset.
- Multi-frequency signal: Brain activity consists of frequencies of different ranges for different brain functions.

	All		Top-3		Top-1	
LEMON dataset	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**
Blackbox Logistic Regression	.497	.50	.493 (99.1)	**.64 (128)**	**.497 (100)**	**.50 (100)**
Glassbox Explainable Boosting	.695	.64	.657 (94.5)	.62 (96.8)	.595 (85.6)	.57 (89)
Glassbox Logistic Regression	.497	.50	.493 (99.1)	**.50 (100)**	.495 (99.6)	**.50 (100)**
Glassbox Decision Tree	.583	.58	**.586 (100)**	**.58 (100)**	.574 (98.4)	.57 (98.3)
EEGMAT dataset	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**
Blackbox Logistic Regression	.500	.74	.499 (99.8)	**.74 (100)**	.498 (99.6)	**.74 (100)**
Glassbox Explainable Boosting	.664	.75	.607 (91.4)	**.75 (100)**	.552 (83.1)	.74 (98.6)
Glassbox Logistic Regression	.500	.74	.499 (99.8)	**.74 (100)**	.498 (99.6)	**.74 (100)**
Glassbox Decision Tree	.543	.74	.536 (98.7)	**.74 (100)**	.532 (97.9)	**.74 (100)**

Table 1: Summary of results on two datasets. Numbers in brackets represent the value as a percentage compared to the corresponding value in the All-channels results. Results of 100% or more are highlighted. Note that the accuracy of the 3-channel result for logistic regression on LEMON is higher than for All-channels.

	All		Top-3		Top-1	
LEMON dataset	230.24	(100)	18.95	**(8.2)**	10.79	(4.7)
EEGMAT dataset	56.41	(100)	14.81	**(26.3)**	4.54	(8.0)

Table 2: Summary of the time taken (in minutes) to run the program with All, Top-5, and Top-1 channels in two datasets. In the LEMON dataset, All takes more than 3 hours (180 mins). Numbers in brackets represent the value as a percentage compared to the corresponding value in the All-channels results. Results for 3 channels are highlighted.

They range from delta (0 - 4Hz during sleep) to gamma (>32Hz for high cognitive activity) and alpha, beta, theta in between (M et al. 2017).

Our results show that a reduction in the number of channels for both LEMON (from 38 to 3; 92% fewer channels) and EEGMAT (from 21 to 3; 86% fewer channels) did not impact the performance of the models significantly. The models with only 3 channels that were identified by the XAI method could capture more than 90% of the AUC and accuracy scored by models with all the channels.

Also, the top-3 channels recognized by the XAI method for the LEMON dataset were PO3, PO8, and P6, which were tracing activity from the parietal-occipital (PO) and parietal (P) region of the brain. The PO/P region is responsible for visual and spatial processing (Baker et al. 1996), which is very relevant in the context of eyes-open (EO) and eyes-closed (EC) tasks.

In the EEGMAT dataset, the task involved mental math, and the top-3 channels identified were FP1, F3, and F7, all representing the frontal lobe region. This region is responsible for executive functions like attention, problem-solving, and judgement (Scott and Schoenberg 2011); all required for a mental math task. The channels identified by XAI are also very appropriate for the dataset.

We saw that XAI methods were able to correctly recognize the channels based on the datasets and task. This establishes that XAI can play an important role in analyzing EEG. With this case study, we could see that:

- XAI methods can be used successfully for channel selection.
- XAI methods can correctly identify the channels that have physiological relevance in the context of the EEG task.

Deep Learning, LLMs and Compact Models

A growing body of research in TS is on using deep-learning models and LLMs for TS tasks (Miller et al. 2024). Here, one line of work is exploring different deep learning models like Long-short Term Model (LSTM) using the numeric data or transforming the data into other formats (e.g., images) and exploring Convolutional Neural Networks (CNNs) (Zeng et al. 2023a). A second line of work is exploring the application of pre-trained LLMs to the tasks (Jia et al. 2024) where the training data may be general-purpose or time-based, as well as fine-tuning them. A third and more recent line of work is designing language models from scratch, attempting to reduce model size without sacrificing performance(Ekambaram et al. 2024).

These trends are slowly being adopted in the analysis of biosignals as well. However, the standard benchmarks used in TS papers still do not use EEG or other biosignal datasets (Ekambaram et al. 2024; Zeng et al. 2023b; Jia et al. 2024). We hope that this paper will help AI researchers working in TS and XAI in adopting biosignal/EEG data for developing and evaluating their methods.

Explainable AI

XAI is undergoing a period of rethink on how it should be positioned with respect to end users. Although XAI methods have been around, as discussed earlier (Ali et al. 2023), their adoption is mostly high among model developers and not end users. As a result, it has been recently argued that the traditional role of XAI, where it recommends a human to accept or reject AI's output, is outdated since this ignores humans' cognitive processes in making a decision after con-

sidering all aspects of a decision(Miller 2023). Instead, in a new proposed framework called *Evaluative AI*, XAI methods are positioned in a machine-in-the-loop setup to provide evidence for and against decisions to be made by a user.

We envisage such a role shift to happen especially in the context of EEG, since it deals with personal end-user health and well-being, which both patients and providers have high stakes in, not just AI developers. Furthermore, EEG data is more complex than normal TS data, making the need for explainability even more important.

Conclusion

AI and health communities can tremendously benefit each other in TS, with the former offering advanced analytical methods while the latter provides complex data sets and trust-sensitive use cases. But the communities also need to overcome confusing terminologies, hidden assumptions, and a lack of necessary domain contexts for result evaluation and interpretation. In this paper, we attempted to bridge their gaps using the problem of channel selection in EEG. We outlined challenges in working with EEG data, showed how general TS trends can benefit health, like using LLMs and XAI methods, and demonstrated via two experiments how simple XAI methods can be quite effective for channel selection. We hope that this work will bring researchers working on TS problems at the intersection of AI and health to better leverage results from their respective areas to overcome common challenges.

References

Ahmad, I.; Yao, C.; Li, L.; Chen, Y.; Liu, Z.; Ullah, I.; Shabaz, M.; Wang, X.; Huang, K.; Li, G.; Zhao, G.; Samuel, O. W.; and Chen, S. 2024. An efficient feature selection and explainable classification method for EEG-based epileptic seizure detection. *Journal of Information Security and Applications*, 80: 103654.

Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99: 101805.

Apicella, A.; Isgrò, F.; Pollastro, A.; and Prevete, R. 2022. Toward the Application of XAI Methods in EEG-based Systems. In *XAI.it@AI*IA*.

Asan, O.; Bayrak, E.; and Choudhury, A. 2020. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. In *Journal of Medical Internet Research 22 (6), e15154, At SSRN: https://ssrn.com/abstract=3676111*.

Babayan, A.; Erbey, M.; and Kumral, D. e. a. 2019. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. In *Sci Data 6, 180308. https://doi.org/10.1038/sdata.2018.308*.

Baker, S.; Rogers, R.; Owen, A.; Frith, C.; Dolan, R.; Frackowiak, R.; and Robbins, T. 1996. Neural systems engaged by planning: a PET study of the Tower of London task. *Neuropsychologia*, 34(6): 515–526.

Boukherouaa, E. B.; Shabsigh, G.; AlAjmi, K.; Deodoro, J.; Farias, A.; Iskender, E. S.; Mirestean, A. T.; and Ravikumar, R. 2021. Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance. In *International Monetary Fund, ISBN: 9781589063952, https://doi.org/10.5089/9781589063952.087*.

Choel-Hui, L.; Daesun, A.; Hakseung, K.; Jin, H. E.; Jung-Bin, K.; and Dong-Joo, K. 2025. NeuroXAI: Adaptive, robust, explainable surrogate framework for determination of channel importance in EEG application. *Expert Systems with Applications*, 261: 125364.

Dukyong, Y.; Jong-Hwan, J.; Jin, C. B.; Young, K. T.; and Ho, H. C. 2020. Discovering hidden information in biosignals from patients using artificial intelligence. *Korean J Anesthesiol*, 73(4): 275–284.

Ekambaram, V.; Jati, A.; Dayama, P.; Mukherjee, S.; Nguyen, N. H.; Gifford, W. M.; Reddy, C.; and Kalagnanam, J. 2024. Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series. In *Neurips. On Arxiv at: https://arxiv.org/abs/2401.03955*.

Elsayed, S.; Thyssens, D.; Rashed, A.; Jomaa, H. S.; and Schmidt-Thieme, L. 2021. Do we really need deep learning models for time series forecasting? *arXiv preprint arXiv:2101.02118*.

Fuentes-Aguilar, R. Q.; Pérez-Espinosa, H.; and de-la Cruz, M. A. F. 2022. Chapter 2 - Biosignals analysis (heart, phonatory system, and muscles). In Torres-García, A. A.; Reyes-García, C. A.; Villaseñor-Pineda, L.; and Mendoza-Montoya, O., eds., *Biosignal Processing and Classification Using Computational Learning and Intelligence*, 7–26. Academic Press. ISBN 978-0-12-820125-1.

Goldberger, A.; Amaral, L.; Glass, L.; Hausdorff, J.; Ivanov, P. C.; Mark, R.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. In *Circulation [Online]. 101 (23), pp. e215–e220.*

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton University Press.

I, D.; SJ, G.; PW, B.; and LJ., O. 2010. Neural correlates of task and source switching: similar or different? In *Biol Psychol. 2010 Mar;83(3):239-49. doi: 10.1016/j.biopsycho.2010.01.008.*

Jia, F.; Wang, K.; Zheng, Y.; Cao, D.; and Liu, Y. 2024. GPT4MTS: Prompt-based Large Language Model for Multimodal Time-series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21): 23343–23351.

Lakkaraju, K.; Kaur, R.; Zeng, Z.; Zehtabi, P.; Patra, S.; Srivastava, B.; and Valtorta, M. 2024. Rating Multi-Modal Time-Series Forecasting Models (MM-TSFM) for Robustness Through a Causal Lens. arXiv:2406.12908.

Lakkaraju, K.; Srivastava, B.; and Valtorta, M. 2024. Rating Sentiment Analysis Systems for Bias Through a Causal Lens. *IEEE Transactions on Technology and Society*, 1–1.

Longo, L.; Goebel, R.; Lecue, F.; Kieseberg, P.; and Holzinger, A. 2020. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In Holzinger, A.; Kieseberg, P.; Tjoa, A. M.; and Weippl, E., eds., *Machine Learning and Knowledge Extraction*, 1–16. Cham: Springer International Publishing. ISBN 978-3-030-57321-8.

M, R.-A.; L, A.; S, H.; and M., A. 2017. Changes of the brain's bioelectrical activity in cognition, consciousness, and some mental disorders. In *Med J Islam Repub Iran. 2017 Sep 3;31:53. doi: 10.14196/mjiri.31.53.*

Mendis, K.; Wickramasinghe, M.; and Marasinghe, P. 2024. Multivariate Time Series Forecasting: A Review. In *Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition*, CVIPPR '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400716607.

Miller, J. A.; Aldosari, M.; Saeed, F.; Barna, N. H.; Rana, S.; Arpinar, I. B.; and Liu, N. 2024. A Survey of Deep Learning and Foundation Models for Time Series Forecasting. arXiv:2401.13912.

Miller, T. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 333–342. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Muppasani, B.; Anand, C. J.; Appajigowda, C.; Srivastava, B.; and Johri, L. 2024. A Dataset and Baseline Approach for Identifying Usage States from Non-intrusive Power Sensing with MiDAS IoT-Based Sensors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15545–15550.

Muppasani, B. C.; Anand, C. J.; Appajigowda, C.; Srivastava, B.; and Johri, L. 2022. Power Forecasting and Anomaly Detection with MIDAS IoT-based Sensor. In *DOI: 10.13140/RG.2.2.17358.33600*.

Nagel, S. 2019. *Towards a home-use BCI: fast asynchronous control and robust non-control state detection*. Ph.D. thesis, Universitat Tubingen.

Nori, H.; Jenkins, S.; Koch, P.; and Caruana, R. 2023. InterpretML: A Unified Framework for Machine Learning Interpretability. In *https://arxiv.org/abs/1909.09223*.

Rojat, T.; Puget, R.; Filliat, D.; Ser, J. D.; Gelin, R.; and D'iaz-Rodr'iguez, N. 2021. Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey. *ArXiv*, abs/2104.00950.

Roy, Y.; Banville, H.; Albuquerque, I.; Gramfort, A.; Falk, T. H.; and Faubert, J. 2019. Deep learning-based electroencephalography analysis: a systematic review. In *J Neural Eng . 2019 Aug 14;16(5):051001. doi: 10.1088/1741-2552/ab260c.*

Scott, J. G.; and Schoenberg, M. R. 2011. Frontal Lobe/Executive Functioning. In *The Little Black Book of Neuropsychology: A Syndrome-Based Approach, DOI 10.1007/978-0-387-76978-3*$_1$0.

Shyalika, C.; Wickramarachchi, R.; and Sheth, A. P. 2024. A Comprehensive Survey on Rare Event Prediction. *ACM Comput. Surv.*, 57(3).

Smith, E. 2024. https://www.ebme.co.uk/articles/clinical-engineering/introduction-to-eeg. In *EBME Biomedical and Clinical Engineering*.

SR, S.; L, S.; D, S.; D, S.-J.; KE, D.; JJ, H.; AJ, H.; FW, D.; and MM, S. 2016. American Clinical Neurophysiology Society Guideline 1: Minimum Technical Requirements for Performing Clinical Electroencephalography. In *J Clin Neurophysiol. 2016 Aug;33(4):303-7. doi: 10.1097/WNP.0000000000000308. Erratum in: J Clin Neurophysiol. 2021 May 1;38(3):e16. doi: 10.1097/WNP.0000000000000817.*

Srivastava, B.; Lakkaraju, K.; Bernagozzi, M.; and Valtorta, M. 2024. Advances in Automatically Rating the Trustworthiness of Text Processing Services. In *AI Ethics 4, 5–13. https://doi.org/10.1007/s43681-023-00391-5. Preprint on Arxiv at: https://arxiv.org/abs/2302.09079*.

Srivastava, B.; and Rossi, F. 2018. Towards Composable Bias Rating of AI Systems. In *2018 AI Ethics and Society Conf. (AIES 2018), New Orleans, Louisiana, USA, Feb 2-3*.

Srivastava, B.; and Rossi, F. 2020. Rating AI Systems for Bias to Promote Trustable Applications. In *IBM Journal of Research and Development*.

Varshney, K. R. 2022. Trustworthy Machine Learning. *ISBNL 979-8411903959*.

W., K. 2009. Everything you wanted to ask about EEG but were afraid to get the right answer. In *Nonlinear Biomed Phys 3, 2 (2009). https://doi.org/10.1186/1753-4631-3-2*.

Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; and Sun, L. 2022. Transformers in Time Series: A Survey. In *arXiv:2202.07125*.

Yulan, D.; and Wenshan, Z. 2024. Channel Selection for Seizure Detection Based on Explainable AI With Shapley Values. *IEEE Sensors Journal*, 24(16): 26126–26135.

Zeng, Z.; Kaur, R.; Siddagangappa, S.; Balch, T.; and Veloso, M. 2023a. From Pixels to Predictions: Spectrogram and Vision Transformer for Better Time Series Forecasting. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, 82–90. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702402.

Zeng, Z.; Kaur, R.; Siddagangappa, S.; Balch, T.; and Veloso, M. 2023b. From Pixels to Predictions: Spectrogram and Vision Transformer for Better Time Series Forecasting. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, 82–90.

Zhang, X.; Chowdhury, R. R.; Gupta, R. K.; and Shang, J. 2024. Large Language Models for Time Series: A Survey. In *Proc. 33 Int. Joint Conf. on AI, IJCAI-24*.

Zyma, I.; Tukaev, S.; Seleznov, I.; Kiyono, K.; Popov, A.; Chernykh, M.; and Shpenkov, O. 2019. Electroencephalograms during Mental Arithmetic Task Performance. *Data*, 4(1).