Neural Computing and Applications (2023) 35:10051–10068
https://doi.org/10.1007/s00521-022-07809-x

S.I.: INTERPRETATION OF DEEP LEARNING

XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series

Dominik Raab[1] · Andreas Theissler[1] · Myra Spiliopoulou[2]

Received: 31 March 2022 / Accepted: 6 September 2022 / Published online: 29 September 2022

Abstract
In clinical practice, algorithmic predictions may seriously jeopardise patients' health and thus are required to be validated by medical experts before a final clinical decision is met. Towards that aim, there is need to incorporate explainable artificial intelligence techniques into medical research. In the specific field of epileptic seizure detection there are several machine learning algorithms but less methods on explaining them in an interpretable way. Therefore, we introduce XAI4EEG: an application-aware approach for an explainable and hybrid deep learning-based detection of seizures in multivariate EEG time series. In XAI4EEG, we combine deep learning models and domain knowledge on seizure detection, namely (a) frequency bands, (b) location of EEG leads and (c) temporal characteristics. XAI4EEG encompasses EEG data preparation, two deep learning models and our proposed explanation module visualizing feature contributions that are obtained by two SHAP explainers, each explaining the predictions of one of the two models. The resulting visual explanations provide an intuitive identification of decision-relevant regions in the spectral, spatial and temporal EEG dimensions. To evaluate XAI4EEG, we conducted a user study, where users were asked to assess the outputs of XAI4EEG, while working under time constraints, in order to emulate the fact that clinical diagnosis is done - more often than not - under time pressure. We found that the visualizations of our explanation module (1) lead to a substantially lower time for validating the predictions and (2) leverage an increase in interpretability, trust and confidence compared to selected SHAP feature contribution plots.

Keywords Explainable AI · SHAP · Deep learning · Machine learning · Epileptic seizures · EEG time series

1 Introduction

In recent years, machine learning (ML) has revealed its merit in many tasks that typically require human intelligence [1] and has even demonstrated better performance than that of human experts for certain task [2–4]. Driven by the increased popularity of ML systems in various domains, products and services [5–7] and the resulting significant impact on society [8], explaining and interpreting these systems has become a crucial skill. Hence, the research field of Explainable Artificial Intelligence (XAI) has emerged with vast amount of research interest [9–18].

In the medical domain, where a variety of ML applications are increasingly introduced for medical diagnostics, treatment and prevention [19], is need for deployment of XAI methods as algorithmic predictions here impact human lives. At this, explainability and interpretability are essential for ensuring that a prediction is made based on traceable reasons [13] and for addressing the lack of transparency that already has led to incidents [20]. To enhance clinical decision-making there is increasing interest in ML for the analysis of multivariate electroencephalography (EEG) time series, e.g. for detection of Parkinson's disease [21], schizophrenia [22, 23] and the detection of epileptic seizures [24].

✉ Dominik Raab
dominik.raab@hs-aalen.de

Andreas Theissler
andreas.theissler@hs-aalen.de

Myra Spiliopoulou
myra@ovgu.de

[1] Aalen University of Applied Sciences, 73430 Aalen, Germany

[2] Otto-von-Guericke University Magdeburg, 39106 Magdeburg, Germany

More than 65 individuals worldwide suffer from epilepsy [25], making this condition one of the most common and serious neurological diseases. Epilepsy leads to a continual susceptibility towards epileptic seizures [26] whereby the seizures occur more frequently in the neonatal phase, especially within the first postnatal week [27]. Thus, epileptic seizures are the most frequent neurological emergency in neonates [28] having the greatest susceptibility to seizures of any age group [29] with an incidence of 1.8 to 3.5 per 1,000 live births [30, 31].

Motivation As most seizures among neonates are acute symptomatic events [32] constituting a neurological emergency and often implying serious dysfunction or impairments of the immature brain [29], an instant detection and appropriate treatment with antiepileptic drugs is required. The authors of [33] highlight the overall mortality rate of 4% for neonates suffering from seizures. Though, medical expertise in intensive care units is not continuously available [34], making an automatic detection valuable. However, interpreting neonatal EEG is not an easy task [35] and generally requires a neurophysiologist or paediatric neurologist with specific expertise [36].

So far, the international gold standard for detecting neonatal seizures is the visual detection of electrographic discharges in the multi-channel EEG by medical experts [37–39]. However, the visual inspection of continuous EEG recordings is a time consuming, monotonous and error-prone process and misdiagnosis can be very harmful [40] leading to injury and even death [26]. Although several ML algorithms for detecting seizures have been proposed in literature [36, 41–45], there are less methods on explaining them in an interpretable way, impeding an embedding into clinical practice. Here, the major challenge is to bridge the technical ML components and the established medical environments [46, 47] by designing interactive explanation interfaces that enable clinical decision-makers to validate the algorithmic predictions.

EEG time series are characterized by (a) spectral, (b) spatial and (c) temporal dimensions and since all are crucial for seizure detection, *we argue that an explanation of an algorithmic prediction must encompass these three dimensions.* For a seizure, the spectral dimension refers to the frequency bands where a seizure became obvious [48], the spatial dimension reflects the location on the scalp [49], and the temporal dimension refers to the point in time during the recording [50].

Proposed approach Medical experts cannot be expected to be familiar with the internal decision making of ML algorithms nor with their mathematical/statistical assumptions. Therefore, the main objective of this work lies in designing visual explanations that are adjusted to the given problem setting – detecting epileptic seizures in multivariate EEG time series – and are interpretable for medical experts. To this end, we propose XAI4EEG: an application-aware approach for an explainable and hybrid deep learning-based detection of seizures in multivariate EEG time series. In XAI4EEG, we model the task of seizure detection as a binary classification problem. For this problem, we design two seizure detection methods – 1D-CNN and 3D-CNN – that we fuse in XAI4EEG, where each method incorporates into a Convolutional Neural Network three types of domain knowledge – (a) frequency bands, (b) location of EEG leads and (c) temporal characteristics. Explainability arises from integrating an ensemble of SHAP explainers[1] generating local explanations. To deal with the complexity of the returned explanations and to enhance interpretability for medical experts, we couple our methods with a mechanism that maps computed SHAP values to an explanation module that highlights the location of decision-relevant regions in the spectral, spatial and temporal EEG dimensions.

To the best of our knowledge, we are the first to introduce a visual explanation schema for deep learning-based seizure detection that displays feature contributions in all three EEG dimensions. The contributions of this work are as follows:

1. We propose an explanation module visualizing SHAP values obtained by two SHAP explainers, each explaining the predictions of one of two deep learning models. The resulting visual explanations enable the identification of decision-relevant regions in the spectral, spatial and temporal EEG dimensions.
2. We incorporate the explanation module into a hybrid seizure detection approach that encompasses EEG data preparation, two deep learning models (1D-CNN and 3D-CNN) and a flow of patterns to the explanation module.
3. We introduce an evaluation scenario that emulates the fact that clinical diagnosis is done under time pressure, and follows the human-grounded evaluation principle proposed in [14]. In an initial study with the aforementioned scenario, we report on the effectiveness of the explanation module and show that it leads to a substantially lower time for validating the predictions compared to selected feature contribution plots implemented in the SHAP package.
4. We provide reproducible research by offering the prototype, source code and a tutorial video[2].

The rest of the paper is organized as follows: In Sect. 2, we survey related work. Section 3 presents XAI4EEG. Section 4 comprises performance indicators of the deep learning models. In Sect. 5, we conduct a user study to

[1] Both comprised in the SHAP package introduced in [16].

[2] Prototype and video: www.ml-and-vis.org/xai4eeg

validate XAI4EEG. In Sect. 6, we discuss the outcomes and Sect. 7 concludes this paper.

2 Related work

Since algorithmic predictions directly influence and may potentially harm patients' health, obvious concerns for the adoption of ML models raised with regard to the liability for medical malpractice law [51]. Moreover, the ethical, legal and moral issues of ML systems within the medical domain have already gained attention in recent years [52]. Since black box models are difficult to implement into the medical work flow and routines [53], it is essential to provide interpretable explanations for their predictions. Studies actually highlighted the positive effect of explanations provided for clinical end-users [54]. Adjusting the visual representation of feature contributions to the given problem setting [55], aims to ensure the explanations are interpretable [56] and useful [57] for an application expert and thus support clinical decision making [58]. Particularly, this is important for automatic neonatal seizure detection systems in clinical practice [59]. Two recent examples for XAI in medical applications are [60] presenting explainability for the task of skin cancer prediction and [61] addressing the interpretation of cortical EEG source signals during the preparation of hands' sub-movements.

In the following, we discuss first studies on deep-learning-based seizure detection algorithms that process EEG time series. Thereafter, we present studies pursuing an explainable seizure detection approach.

2.1 Deep learning-based seizure detection in EEG time series

In [62], Hu et al. design a deep bidirectional Long short-term memory (LSTM) network comprising two independent LSTM networks for the detection of onset seizures in adults. The proposed architecture enables to process information prior and after the analyzed sequence resulting in a sensitivity of 93.61% and a specificity of 91.85% on average based on an unseen test set. The authors of [63] propose a nested LSTM network that explores the inherent temporal dependencies in EEG. Based on unseen data, sensitivity and specificity values were 97.47 and 96.17%. In [64], Abdelhameed et al. introduce an architecture using 1D-CNN as preprocessing front-end followed by a bidirectional LSTM. Given a multi-class classification problem distinguishing normal, interictal and ical events, the authors achieve an average overall accuracy of 98.89% based on k-fold cross-validation. The authors of [65] propose 1D-CNN for detecting neonatal seizures and achieve an average AUC of 0.971 based on leave one-patient-out cross-validation.

Time series imaging is widely used for encoding time series as images. For time-frequency decomposition of EEG data, time-frequency maps can be created per electrode. These 2D images containing spectral and temporal information can be used in 2D-CNN classification tasks. In [66], Yuan et al. compare four different 2D-CNNs working on time-frequency maps. The proposed architectures are based on a transfer learning VGG-net that was introduced in [67]. Two of the networks reach a sensitivity and specificity of more than 90%. Note that since EEG signals are characterized by spectral, temporal, and spatial information, the fact of multi-channel signal processing [68] is ignored when classifying single EEG time-frequency maps. Here, the location of electrodes over the scalp is discarded. To address this, the authors of [69] designed a 3D-CNN to detect inter-ictal, pre-ictal, and ictal EEG stages. The performance indicators are based on an unseen test set and are as follows: sensitivity of 88.90% and specificity of 93.78%. A further approach to maintain at least parts of spatial information is proposed in [70]. Here, three time-frequency maps from different electrodes are fused into a single output image that is used as input for a 2D-CNN. The average accuracy using VGG16, VGG19, and ResNet50 [71] is 97.75, 98.26, and 96.17%, respectively. In our hybrid approach we incorporate both, a 3D-CNN as described by [69] in combination with a 1D-CNN as described e.g. in [65].

2.2 Explainable seizure detection

The authors of [72] propose an explainable seizure detection approach based on connectivity features using EEG time series. Seven connectivity features are computed from each of the common EEG frequency bands, arranged as a tensor and finally fed as features to the model. The model combines a 2D-CNN and a bidirectional LSTM achieving a sensitivity of 97.65% and a specificity of 96.58%. By implementing a self attention layer, the authors computed the relevance of each input feature for a certain decision using the weights of the model stored in activation values and provided an explanation in the spectral and temporal EEG dimensions. The proposed feature relevance extraction is computationally expensive. The authors of [73] introduce a CNN with an attention mechanism that automatically extracts the importance of each electrode and thus pays more attention to the important ones. Furthermore, this mechanism enables to visualize the important brain regions on topographies leveraging a spatial explanation. Sensitivity and specificity values of 97.4 and of 88.1% are achieved. Combining 1D-CNN with Grad-CAM producing heatmaps that are overlaid on the original input

is proposed in [74]. This enables the identification of reccurring patterns and a temporal explanation. A sensitivity of 66.9%, and a specificity of 83.0% is reached. One recent example for use of SHAP in seizure detection is [75] presenting a visual channel-level SHAP map that highlights the highest contributing EEG channel. The two 2D-CNNs reach a accuracy of 88.81 and 91.54%, respectively. The explanation appears in the spatial and temporal dimensions. More recently, the authors of [76] propose superimposition of computed SHAP values on a 2D grayscale image that is composed of the raw EEG signal representation of four EEG channels. This approach constitutes an explanation in the spatial and temporal dimensions. A F1-score of 0.873 using 2D-CNN is reached.

Differently from the aforementioned works, we introduce a novel representation of SHAP values providing a full-fledged visual explanation in all three EEG dimensions: (a) spectral, (b) spatial and (c) temporal.

3 Methodology

In this section, we first begin the description of our approach with two core definitions. Thereafter, we give an overview of XAI4EEG and then we describe each component in turn.

While the underlying idea of our approach might be useful for other problem domains, we focus on the special characteristics of EEG data: The used data is a multivariate time series stemming from electrodes attached to the scalp. The underlying EEG data corresponds to a multivariate time series X_t containing M univariate time series $X_t{}^{(j)}$, one for each signal of $j : 1 \ldots M$ electrodes. X_t is subdivided into non-overlapping intervals $I = X_{[t_i,t_i+w)}$, where w defines the length of the interval in seconds. The used data encompasses annotations in seconds and each second is ascribed a seizure or normal that we modelled as a binary classification problem. Thus, we dissect the intervals I into subsequences $S = X_{[t_i,t_i+1)}$ of length 1 seconds using a non-overlapping window. This notation appears in Table 1.

During EEG recording seizure and normal patterns can alternate. Thus, our intervals can enclose subsequences that are annotated as both seizure and normal. To define an overall label and classification result for each interval, we define intervals as "seizure" if they contain ≥ 1 subsequences that are annotated as seizure (adopted from [77]). An example is shown in Fig. 1.

Normal subsequences are denoted by S_{normal}, seizure subsequences by S_{seizure}. The intervals were either labeled as normal, termed as I_{normal}, or as seizure I_{seizure}. The classification results are then determined as follows, where $I_{\text{classified=normal}}$ is an interval classified as normal and $I_{\text{classified=seizure}}$ an interval classified as seizure:

$$TP : \exists S_{\text{seizure}} \in I_{\text{classified=seizure}}$$
$$TN : \nexists S_{\text{seizure}} \in I_{\text{classified=normal}}$$
$$FP : \nexists S_{\text{seizure}} \in I_{\text{classified=seizure}}$$
$$FN : \exists S_{\text{seizure}} \in I_{\text{classified=normal}}$$

Our goal is to build two classifiers both conducting the seizure detection on the basis of these intervals' feature set (see Definition (1)). We refer to this classifiers as *interval-based classifiers*. Detected seizure or normal pattern is the result of both classifiers.

Definition 1 (*Interval-based seizure detection*) We define *interval-based seizure detection* as the classification of an interval I of the multivariate EEG time series X_t as either "seizure" or "normal". We assess an interval I as a true positive, if it was classified as "seizure" and it contains at least one subsequence S that was annotated as "seizure".

The main idea is to incorporate multiple (in our case two) classifiers and – transferring the idea of ensembles to explanations – multiple explanations. We refer to this as hybrid explainable seizure detection (see Definition (2)). Since seizure detection is on the basis of intervals, so is the explanation.

Definition 2 (*Hybrid explainable seizure detection*) We define *hybrid explainable seizure detection* as the detection of seizures by more than one classifier (in our case interval-based classifier), each explained by at least one explanation.

Table 1 Notation: methodology

Notation	Description
X_t	multivariate EEG time series containing M univariate time, one for each signal of $j : 1 \ldots M$ electrodes.
$X_t{}^{(j)}$	univariate time series of one EEG electrode
interval I	non-overlapping interval $X_{[t_i,t_i+w)}$ of the continuous EEG data, with length w in seconds
subsequence S	subsequence $X_{[t_i,t_i+1)}$ of an interval I with a 1-second-duration

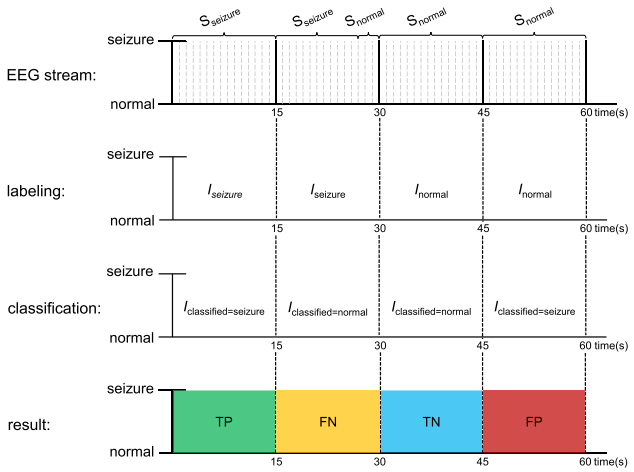

Fig. 1 Labeling and classification of intervals I based on the annotations obtained for subsequences, showing correctly and falsely detected seizures (TP, FP), correctly detected normal intervals (TN) and undetected seizures (FN)

3.1 XAI4EEG

In the following, we describe the fundamental flow of patterns in XAI4EEG. An overview of the components of our approach is given in Fig. 2.

In XAI4EEG, we propose to use two seizure detection methods, each composed of a preprocessing step that is common to both methods, followed by steps for feature extraction, an interval-based classification model, and a post-hoc explanation component. The first is denoted as *Detector1D* – inspired by the underlying 1D-CNN classifier – and classifies the EEG data that is transformed into the frequency domain, referred to as *Inp1D*. The second method, *Detector3D* using a 3D-CNN classification model, classifies the data transformed into time-frequency domain, referred to as *Inp3D*. This notation appears in Table 2. The post-hoc explanation component of both methods encompass a SHAP explainer explaining the classifier prediction, and our proposed explanation module used to visualize the computed feature contributions. The resulting visual explanations are denoted by *electrode-wise explanation*, *Explanation1D* and *Explanation3D* (see Fig. 2, right).

From the technical perspective, *Detector1D* and *Detector3D* are two standalone, independently built seizure detection methods that we fuse in XAI4EEG. We modelled XAI4EEG so that both methods concurrently process and classify an interval, and provide an explanation to it. Thus, for each interval, XAI4EEG outputs two classification results and two local explanations.

3.2 Preprocessing: filtering

As a first step, low and high frequencies are filtered out. As neonatal seizures have shown to emerge in frequencies between 0.5 and 12.5 Hz and the dominant frequencies range between 0.5 and 6 Hz [78], we implemented a bandpass filter (see Fig. 3, left) using the finite-impulse-response filtering while keeping a low frequency of 0.5 Hz (high-pass) and a high frequency of 12.5 Hz (low-pass). In order to enhance our classifiers' ability to distinguish between neonatal seizures and minor signal noise, we deliberately do not denoise the EEG signals.

3.3 Detector1D

Detector1D (see Fig. 2, top) processes and classifies the EEG signals transformed into the frequency domain. The proposed feature extraction steps capture all of the EEG dimensions.

3.3.1 Feature extraction: interval dissection and recomposing

In order not to discard the variability of the power spectrum within an interval that may indicate a seizure onset, the intervals I are dissected into subsequences $S = X_{[t_i,t_i+1)}$ of length 1 seconds using a non-overlapping window. Afterwards, for each subsequence a spectral analysis is conducted, i.e. the power spectrum is computed and divided into three frequency bands (see Sect. 3.3.2). Since EEG

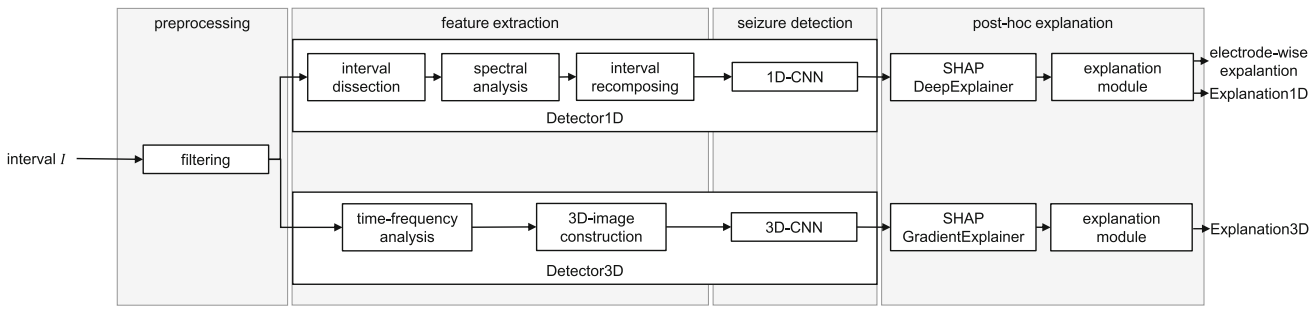

Fig. 2 Overview of the components of XAI4EEG encompassing two seizure detection methods (referred to as *Detector1D* and *Detector3D*), each composed of steps for preprocessing and feature extraction, followed by the seizure detection algorithm, and a post-hoc explanation component

Table 2 Notation: XAI4EEG

Notation	Description
Detector1D	Interval-based seizure detection method using a 1D-CNN and EEG data transformed into the frequency domain.
Inp_{1D}	Input tensor containing the transformed EEG data for the 1D-CNN
Detector3D	Interval-based seizure detection method using a 3D-CNN and EEG data transformed into time-frequency domain.
Inp_{3D}	Input tensor containing the transformed EEG data for the 3D-CNN

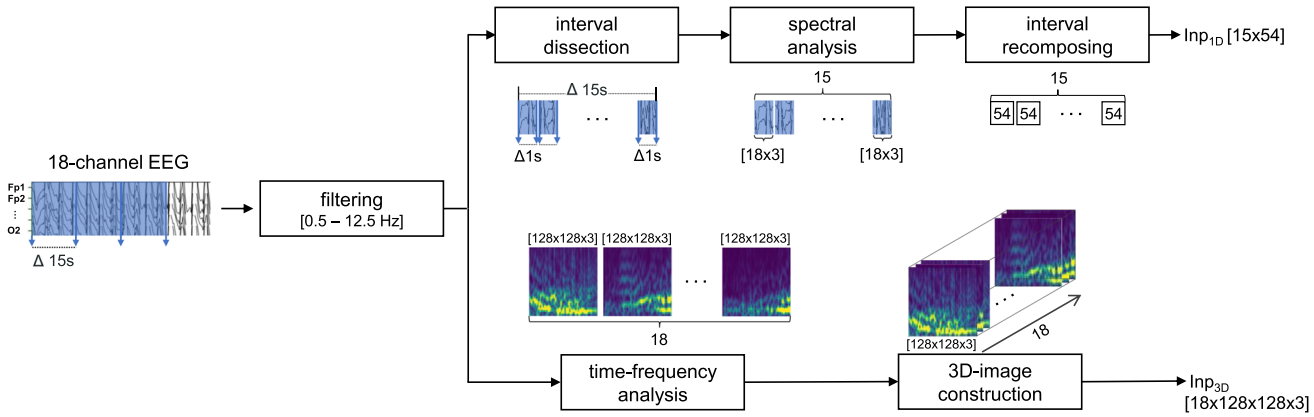

Fig. 3 EEG preprocessing and feature extraction steps, here with interval length w = 15 seconds: *Interval dissection*: dissection of the interval into 15x1-s non-overlapping subsequences, *spectral analysis*: computation of power spectrum per subsequence with Welch's method subdivided into delta, theta, alpha-frequency band resulting in 54 features per subsequence (18 electrodes à three frequency bands), *interval recomposing*: recomposing the interval from the subsequences. *Time-frequency analysis*: generating time-frequency maps (128 x 128 pixels) per electrode with Morlet wavelets, *3D-image construction*: 3D multi-channel image construction by concatenating the single images

data from 18 electrodes are considered in this paper, a total of 54 features are obtained from each subsequence. Finally, the subsequences are unified resulting in a tensor Inp_{1D} with dimensions $[w \times 54]$ (see Fig. 3, top).

3.3.2 Feature extraction: spectral analysis

We transform the EEG time series into the frequency domain [79] using spectral analysis with Welch's method [80] (see Fig. 3, top). After subdividing the EEG signals into sinusoidal oscillations with a known wavelength, the verification of each wavelength for accordance with the signal is realizable through convolution analysis. The power spectrum, as a result of the Welch's method, allows an estimation of the distribution of the frequencies of the EEG signal [81, 82]. The EEG power spectrum is traditionally divided into the five frequency bands: alpha (7.5 ...12.5 Hz), beta (12.5 ...30 Hz), theta (3.5 ...7.5 Hz), delta (0.5 ...3.5 Hz) and gamma (> 30 Hz). REMOVE: Since neonatal seizures have predominantly shown to emerge in the delta, theta and alpha frequency band, we do not consider the beta and gamma band in this work.

3.3.3 Seizure detection: 1D-CNN

Besides recurrent neural networks (RNNs) and its subtypes like LSTM network [83], a 1D-CNN constitutes an effective deep learning technique for processing both univariate and multivariate time series data of variable length. While in 2D-CNNs the kernel is convolved both horizontally and vertically across an image, in 1D-CNNs the kernel is convolved across the data along one dimension, for time series data along the time axis.

We use an architecture consisting of three consecutive hidden layers, each followed by a batch normalization, a dropout layer with a rate of 0.2, and a pooling layer. We set the number of filters in the convolution layer to 64, 128, and 256. The kernel sizes are 3, 3, and 2 respectively. In addition, we apply L2 weight regularization with a regularization parameter of 1e-2 to prevent the model from overfitting. The set of hidden layers is followed by a flatten layer, and a fully connected layer containing 16 nodes. In order to conduct final seizure detection, a fully connected layer with one node and a logistic activation function is used. For the remaining layers the rectifier activation function is used.

3.4 Detector3D

Detector3D (see Fig. 2, bottom) processes and classifies the EEG signals transformed into the time-frequency domain. To not discard spatial information, the electrodes' time-frequency maps are concatenated in order to form a 3D multi-channel image. Classification and explanation are then conducted on this 3D-image.

3.4.1 Feature extraction: time-frequency analysis

Creating Morlet wavelets [84] is a frequently used method for time-frequency analysis. While there are several kinds of Morlet wavelets, we make use of the complex-valued Morlet wavelets defined as the product of a complex sine wave and a Gaussian window. Subsequently, a time-frequency map can be created through wavelet convolution, where the Morlet wavelet is convolved with the time series signal. The wavelet convolution enables the extraction of instantaneous power and phase at any time point. Our final time-frequency map holds the duration of the extracted interval on the x-axis $(0...w)$, and the respective frequency (0.5 ...12.5 Hz) on the y-axis (see Fig. 3, bottom). The color corresponds to the EEG power, i.e. amplitude of the oscillations. We reduce the image size of the time-frequency maps down to 128×128 in order to reduce the complexity of computation.

3.4.2 Feature extraction: 3D-image construction

The EEG data used in this paper was acquired with 19 electrodes of which we exclude the reference electrode Cz. Consequently, each interval results in 18 time-frequency maps. To maintain spatial information of the EEG signals, the respective time-frequency maps are concatenated forming a 3D multi-channel image. This 3D-image, denoted by Inp_{3D}, is used as the input for the 3D-CNN. Its structure is [18, 128, 128, 3] where the 1st dimension corresponds to the number of electrodes and the 4th to the images' color channels (see Fig. 3, bottom right).

3.4.3 Seizure detection: 3D-CNN

A common seizure detection approach is the use of 2D-CNNs to classify the time-frequency maps of univariate EEG signals. However, we aim to incorporate all dimensions – spectral, spatial, and temporal – into one model. Classifying univariate signals ignores the locations of electrodes over the scalp. Hence, we propose to use a 3D-CNN to simultaneously extract EEG features from spectral, temporal, and spatial dimensions by performing 3D convolutions on the 3D-images.

We use three consecutive hidden layers, each followed by a pooling layer. The number of filters in the convolution layer is set to 32, 64, and 64 with a kernel size of [3, 3, 3], [3, 3, 3], and [2, 2, 2]. Furthermore, L2 weight regularization with a regularization parameter of 1e-2 is applied. The set of hidden layers is followed by a flatten layer, a fully connected layer with 32 nodes, and a dropout layer with a rate of 0.2. An output layer with one node and logistic activation function is used to perform final seizure detection. The remaining layers use rectifier activation functions.

3.5 Post-hoc explanation: the proposed explanation module

In the following we elaborate on the explainability of both seizure detection methods, which clearly arises from incorporating our proposed explanation module.

The raw EEG signal representation of the monitoring system is traditionally used in clinical settings by medical experts to detect seizures. Therefore, the main idea is to display the feature contributions (in this work SHAP values) in an explanation module that is leaned to the aforementioned monitoring system.

To compute the SHAP values of our classifiers' predictions, we choose two post-hoc explainers from the SHAP package introduced in [16]: the first is SHAP DeepExplainer – an adoption of the DeepLIFT algorithm – and is incorporated into *Detector1D*. The second explainer is SHAP GradientExplainer – an implementation of expected gradients – and is integrated into *Detector3D*. Note that positive SHAP values increase the probability of the predicted class, while negative SHAP values decrease the probability.

The proposed explanation module is depicted in Fig. 5 (bottom) and is based on a grid of size $[w \times 3]$ holding the length w of the extracted interval on the x-axis, and the frequency bands δ, θ and α on the y-axis. This part of our notation appears in the uppermost row of Table 3. While the x-axis constitutes the explanation in the temporal dimensions, the y-axis explains the spectral dimensions. Thus, each of the 45 cells represents a spectral-temporal EEG region holding the respective SHAP value and is colored in red – inspired by the idea of heatmaps. The more intense the red, the more this spectral-temporal region contributes to the final prediction. The explanation in the spatial dimensions arises from obtaining the feature contributions per electrode.

We propose a workflow to validate the classifier's predictions, where users contrast the returned explanation patterns with the electrographic patterns of the raw EEG signal. Hence, XAI4EEG highlights both agreeing and

Table 3 Notation: post-hoc explanation

Notation	Description
explanation module	Proposed visual construct based on a grid of size $[w \times 3]$ Holding the length w of the interval on the x-axis, and the frequency bands δ, θ and α on the y-axis (see Fig. 5)
$SHAP_{1D}$	Tensor containing the SHAP values obtained by SHAP DeepExplainer
Electrode-wise explanation	Vsual explanation for Detector1D by mapping $SHAP_{1D}$ to multiple explanation modules. Explained dimensions: spectral, spatial and temporal
$Explanation_{1D}$	Visual explanation for Detector1D by mapping $SHAP_{1D}$ to an explanation module. Explained dimensions: spectral and temporal
$SHAP_{3D}$	Tensor containing the SHAP values obtained by SHAP GradientExplainer
$Explanation_{3D}$	Visual explanation for Detector3D by mapping $SHAP_{3D}$ to an explanation module. Explained dimensions: spectral and temporal.

disagreeing patterns, increasing trust in the predictions. The relevant regions highlighted by the explanation module match, at best, the regions on the basis of which the medical expert makes its decision. To illustrate the proposed workflow, an electrode's raw EEG signal representation of an interval (in this case a ground truth seizure) (see Fig. 5, top) is depicted above the explanation module. The spectral-temporal EEG region highlighted by the explanation module match the occurring electrographic discharges in the raw EEG signal that constitutes the onset of a seizure.

In the following we thoroughly describe the flow of patterns to map the computed SHAP values to the explanation module.

3.5.1 Post-hoc explanation: Detector1D

In *Detector1D*, we use a set of explanation modules, each visualizing the feature contributions of one electrode. We refer to this explanation as *electrode-wise explanation*. As a result, an explanation in the (a) spectral, (b) spatial and (c) temporal dimensions is provided. This part of our notation appears in the middle rows of Table 3.

SHAP DeepExplainer computes a SHAP value for each element of Inp_{1D} resulting in a tensor $SHAP_{1D}$ with dimensions $[w \times 54]$ containing the SHAP values. From the medical perspective, the computed contribution of each element (i.e. feature) of Inp_{1D} enables an analysis of decision-relevant electrodes (spatial dimensions), frequency bands (spectral dimensions) and subsequences (temporal dimensions). In order not to overload the visual explanations, we exclude negative values of $SHAP_{1D}$.

First, we design the *electrode-wise explanation* (see Fig. 4, top) by slicing $SHAP_{1D}$ to obtain 18 subsets of size $[w \times 3]$, each containing the feature contributions of one of the 18 electrodes, i.e. the electrodes' decision-relevant regions in the power spectrum for each of the w subsequences. Each of the 18 subsets is then mapped to an explanation module with a grid of size $[w \times 3]$, with the time interval on the x-axis and the three frequency bands on the y-axis. The grid cells are colored in red according to the corresponding SHAP value of $SHAP_{1D}$. One of these explanation modules allows to interpret the feature contributions of one electrode in the spectral and temporal EEG dimensions. As a result, displaying all of these explanation modules adds the explanation in the spatial dimension (see Fig. 4, top middle).

Thereafter, to provide an overall view of the feature contributions across all electrodes, for each subsequence the maximum SHAP value at the frequency band level is extracted across all electrodes. This results in a subset of $SHAP_{1D}$ of size $[w \times 3]$. The elements of the subset are then mapped to the explanation module resulting in *Explanation1D* (see Fig. 4, top right). Each of the grid's $w \times 3$ cells is colored in red where the color intensity corresponds to the subset's SHAP value.

3.5.2 Post-hoc explanation: Detector3D

In *Detector3D*, one explanation module is used visualizing the feature contributions in the spectral and temporal dimensions. Here, an *electrode-wise explanation* is not realizable.

By default, SHAP GradientExplainer produces a baseline visual explanation on the basis of the computed SHAP values and highlights important areas of the 3D multichannel image by coloring image pixels with either red (positive SHAP value) or blue (negative SHAP value).

The raw output of SHAP GradientExplainer is a tensor $SHAP_{3D}$ of size $[1 \times 1 \times 3 \times 30 \times 30 \times 32]$ containing the SHAP values, where the 4th and 5th dimension are the height and width of the baseline visual explanation. Note that the size of $SHAP_{3D}$ depends on the used data set, the preprocessing and feature extraction steps, and the model

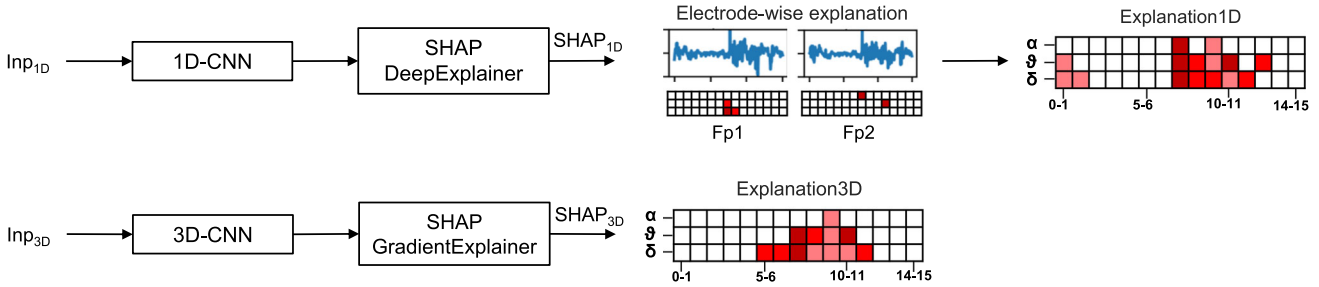

Fig. 4 Seizure detection and post-hoc explanation steps: novel representation of calculated SHAP values obtained from the used SHAP explainers, each explaining the prediction of one of the two classifiers

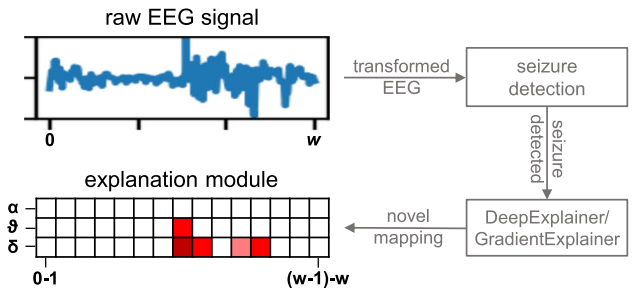

Fig. 5 The proposed explanation module visualizes calulcated SHAP values and is leaned to the raw EEG signal representation of the monitoring system

architecture. We then map $SHAP_{3D}$ to the explanation module and denote the resulting explanation as *Explanation3D*. As for *Detector1D*, we exclude negative values from $SHAP_{3D}$. The aforementioned mapping is done by placing a grid of size $[w \times 3]$, i.e. the explanation module, over the baseline visual explanation of SHAP GradientExplainer. Each of the grid's columns represents a time span with 1-second duration (i.e. a subsequence), while each of the three rows represents a frequency band delta, theta and alpha. Thereafter, the 4th and 5th dimension of $SHAP_{3D}$ are sliced resulting in $w \times 3$ subsets. For each of the subsets the SHAP values are summed up and mapped to the corresponding cell of the explanation module, as visualized in Fig. 4, bottom right. The cells are colored in red where higher intensity represents higher SHAP values. This part of our notation appears in the lower rows of Table 3.

3.6 The operationalization of XAI4EEG

The conceptual framework of XAI4EEG is operationalized using the streamlit package. The resulting user interface is depicted in Fig. 6, , where we aim to emulate the hybrid characteristic of XAI4EEG.

We locate the electrodes' raw EEG signals in the top center (see Fig. 6b), each holding the duration of the extracted interval I on the x-axis and the signal on the y-axis scaled from $-100\,\mu V$ to $+100\,\mu V$. In contrast to the standard way of showing the EEG time series, i.e. a row-wise plot, we propose to arrange the EEG time series according to the position of the electrodes on the scalp, i.e. as per international 10-20 standard localization system. Thus, enabling experts to identify spatial patterns and relations of occurring seizures. The electrode placement of the international 10-20 system on the scalp is depicted in Fig. 7 whereby the electrodes are allocated to the four different lobes of the brain: the frontal, the parietal, the occipital, and the temporal lobe.

We display the preprocessed and transformed EEG data of both seizure detection methods besides the raw EEG signals. Inp_{1D} that comprises the power spectrum of the 18 electrodes subdivided into three frequency bands (described in Sect. 3.3.2) is pictured in tabular form (Fig. 6a). The generated time-frequency maps per electrode – Inp_{3D} – (described in Sect. 3.4.1) are also positioned according to the international 10-20 standard localization system (Fig. 6, c).

Thereunder, we display the predictions of *Detector1D* and *Detector3D* (see Fig.6d, g) and the corresponding visual explanations (see Fig. 6e, f, h). To enable an intuitive user experience, we add the corresponding raw EEG signal below each of the explanation modules (Fig. 6h). Note, that the *electrode-wise explanation* is composed of 18 explanation modules, each for one electrode. But for reasons of space in Fig. 6, only the explanation of electrode Fp1 and Fp2 is depicted.

4 Evaluation

In this section we first describe the data set used to evaluate XAI4EEG. Thereafter, we report on the performance indicators of both deep learning models.

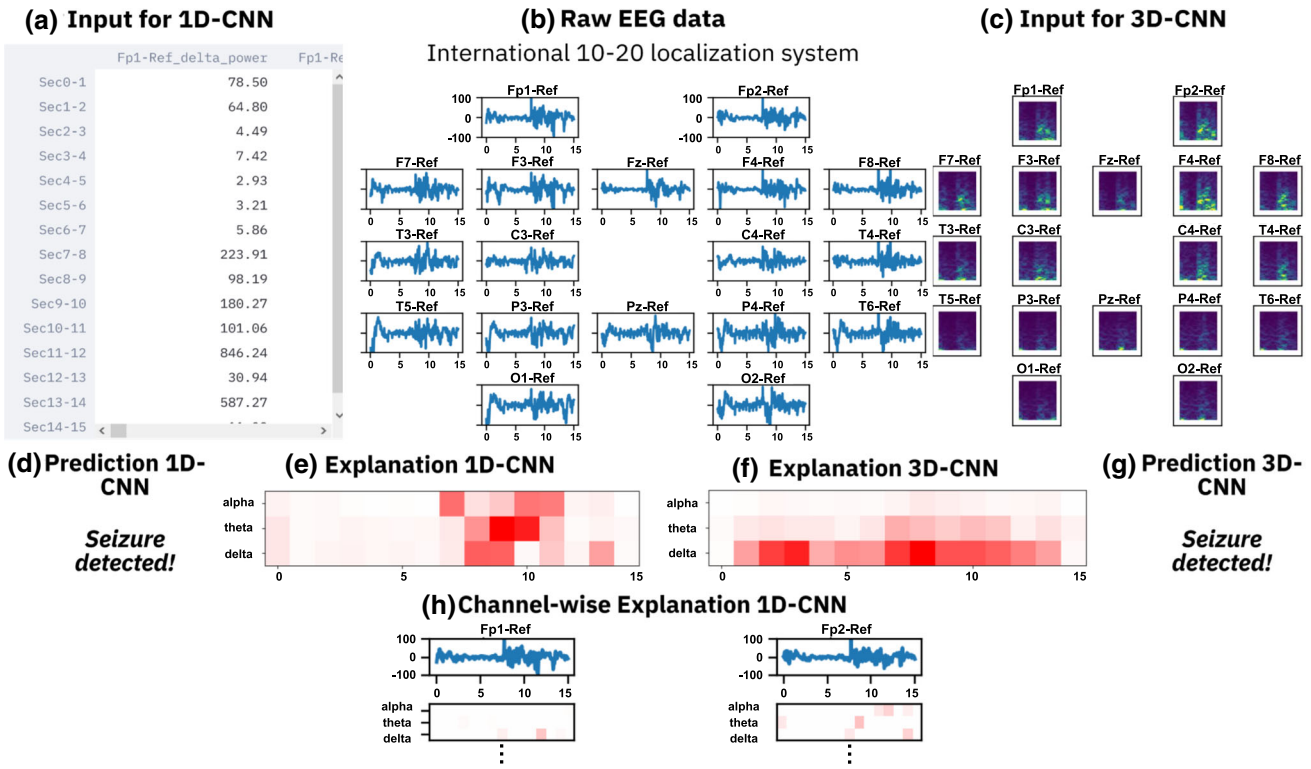

Fig. 6 Operationalized interface: **a** computed power spectrum of the subsequences displayed in tabular form, **b** electrodes' raw EEG signals arranged according to their position on the scalp, **c** generated time-frequency maps per electrode, **d** 1D-CNN prediction, **e** visual explanation for 1D-CNN, **f** visual explanation for 3D-CNN, **g** 3D-CNN prediction, **h** electrode-wise explanation

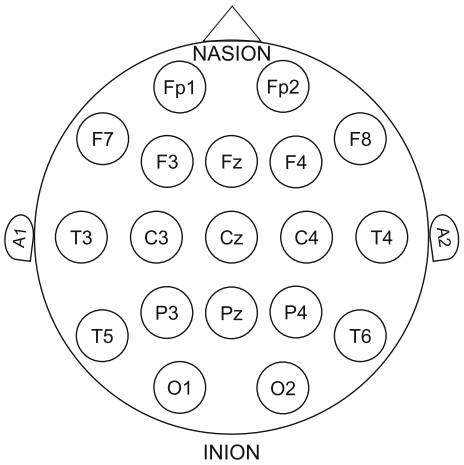

Fig. 7 Placement of electrodes on the scalp according to the 10–20 international localization system. Fp = frontopolar, F = frontal, T = temporal, C = central, O = occipital, P = parietal. Odd numbers = left hemisphere, even numbers = right hemisphere

4.1 Evaluation data

We used an EEG data set presented in [85] comprising neonatal EEG recordings with seizure annotations from three experts. Multi-channel EEG was recorded from 79 neonates admitted to the NICU at Helsinki University Hospital between 2010 and 2014 whereby the median recording duration was 74 min. The raw EEG data was acquired with a sampling frequency of 256 Hz and the electrodes were placed according to the international 10-20 standard referenced at midline, of which 18 electrodes are considered in this paper. Three experts independently annotated the presence of seizures in the EEG data resulting in 1379 seizures in total marked by the experts, of which 889 (65 percent) were annotated by all three experts. A minimum seizure duration of 10 s was defined as criterion for annotating a seizure.

4.2 Determining the interval length *w*

The International Federation of Clinical Neurophysiology assumes a minimum seizure duration of 5 s in case of normal background EEG and 10 s in case of abnormal background EEG [86]. The interval length *w* used in the ML-based seizure detection literature varies from 5 and 10 s [69], 8 s [36] up to 32 s [87]. The choice of length depends on both specific engineers' requirements and characteristics of the used data set, e.g. the background EEG. Given the minimum seizure duration of 10 s requested by the data set's authors for annotating seizures and the fact that most neonates have an abnormal

background EEG, an interval length w of 15 s was chosen in this work, as shown in Fig. 3 (left). This results in $w = 15$ subsequences of length 1 s. The annotations of each of the three experts are given in seconds. Hence, an interval I holds 45 annotations in total, three of which annotate one of the 15 subsequences. We apply a majority vote for each subsequence to receive a single final annotation.

4.3 Seizure detection results

The main focus of this work is to incorporate our explanation module that visualizes feature contributions in all EEG dimensions into an application-aware approach for a hybrid seizure detection. Hence, for reasons of computational time, we decided to randomly select a subset of neonates rather than processing the whole data set. In order to consider neonates with both normal and abnormal EEG background, e.g. asphyxia or cerebral infarction, the following 12 were included in this subset: No. 03, 04, 10, 19, 27, 28, 34, 38, 48, 50, 58, 66. Among these neonates, 6 show normal patterns only, while the remaining 6 repeatedly suffer from seizures. The first 55 min of any neonatal EEG recording were extracted resulting in 2640 intervals each lasting 15 s. We follow our approach that we described in Sect. 3 to label the intervals. As a result, 2290 intervals are labelled as normal and 350 as seizures. Hence, the class distribution is skewed, with the minority class seizure covering 13.3% of the intervals.

We splitted the data set into a training set and a hold-out set. For model training we applied k-fold cross validation (k=10) using the training set. For each iteration the best model hyperparameter values are choosen as per validation loss and used to calculate the performance indicators based on the hold-out set that is never shown to our classifiers during training. The predictive performance of our models depend substantially on the selection of appropriate hyperparameter values. Therefore, we employed a grid search for hyperparameter selection.

The training was performed on a NVIDIA Quadro RTX 6000 GPU. While the 1D-CNN was trained for 200 epochs with a batch size of 64, the 3D-CNN was trained for 30 epochs with a batch size of 32. We used a learning rate of 1^{-3} for both models. After each fold, the models were evaluated using the hold-out set, and the performance indicators were arithmetically averaged. The performance indicators of both models are shown in Table 4.

Table 4 Performance indicators of our models. Values are based on unseen hold-out set and are averaged over 10 folds

Performance-Ind.	1D-CNN		3D-CNN	
	mean (%)	SD (%)	mean (%)	SD (%)
Balanced Accuracy	90.06	0.02	89.07	0.02
Sensitivity	82.57	0.04	86.00	0.05
Specificity	97.55	0.01	92.14	0.03
Precision	84.24	0.06	63.04	0.05

From the medical perspective, it is mandatory that ground truth seizures do not remain unnoticed since patients are dependent on immediate treatment. Hence, high sensitivity is a desired goal in seizure detection. The sensitivity value of the 1D-CNN and 3D-CNN is 86.00 and 82.57%, respectively. Moreover, a low false alarm rate (i.e. high specificity) is desirable and is measured by the number of falsely detected seizures in a given period of time, i.e. false positives (FP) per hour. The specificity value of the 1D-CNN and 3D-CNN is 97.55 and 92.14%, respectively. The precision determines how many of the intervals classified as belonging to seizures are originally seizures. The precision value of the 1D-CNN and 3D-CNN is 84.24 and 63.03%, respectively. Thus, the 3D-CNN misclassifies ground truth normal intervals as seizures (FP) more frequently than the 1D-CNN.

5 User study

We performed a user study to evaluate the usefulness of XAI4EEG. Specifically, we aimed to study the effectiveness of the proposed explanation module for validating the model predictions. Since an evaluation of XAI4EEG with the intended audience in NICUs is challenging, we followed a human-grounded evaluation principle that is proposed in [14] by recruiting laypersons instead of application experts. This principle still maintains the core of the target application while allowing a bigger sample size and causing low expenses.

Under the medical malpractice law, a clinical decision-maker faces liability for grave treatment outcomes when he/she does not follow the standard of care, and (1) rejects a correct algorithmic prediction or (2) follows an incorrect algorithmic prediction [88]. Thus, it is indispensable to understand the reasoning behind the predictions and to ensure patient caretaking is further benefiting from advances in ML. By transferring the increasing need for validation to our problem setting, we aim to mimic a medical workflow in seizure detection, where the study participants were requested to complete a *validation task* that we defined as follows:

- *Validation task*: Validate the model predictions considering raw and transformed EEG data and the visual explanations.

The main goal of the user study is to study whether our proposed explanation module is more advantageous than the use of feature contribution plots implemented in the SHAP package, in terms of time-efficiency and human interpretability during the validation task. The authors of [89] highlight the need to compare the performance of various explanation schemata in the clinical context. We argue that explanations with a high level of human interpretability lead to reduced time for validating the predictions, a fact that is substantial for neonatal seizure detection where immediate treatment is essential. To that end, the *validation time* was measured and defined as follows:

- *Validation time*: The time it took the participant to complete the validation task.

Interpretability has a subjective nature and participants may experience various levels of interpretability. Thus, the *validation task* was followed by three 5-point Likert-scale [90] questions measuring participants' feedback about the *validation task*. Participants used a clearly labeled bipolar 5-point Likert scale to rate their response to each of the three questions, which correspond to three dimensions of interest: (1) confidence, (2) trust and (3) interpretability. The questions and response options appear in Table 5.

5.1 Participants

We recruited 28 participants where they were students and graduates with no specific knowledge of ML, EEG or neonatal seizures. The age of the participants ranged from 21 to 33 ($\bar{x} = 25.21$, $\sigma = 2.61$). The group was composed of 8 women and 20 men. Note that these participants are laypersons in the subject of neonatal seizures. All participants used the tool for the first time and had never worked with the data set before. Prior to the user study, we described the underlying problem setting of this work to the participants, and then we demonstrated the basic components of XAI4EEG.

Table 5 Questions on user self-assessment

Question Text	Answer Options [1...5]
$\mathbf{Q_1}$: How confident did you feel in validating the predictions?	very insecure ...very confident
$\mathbf{Q_2}$: How much do you trust the predictions of the models?	very little ...very much
$\mathbf{Q_3}$: How do you perceive the interpretability of the visual explanations?	very low ...very high

5.2 User study design

The study was subdivided into two A-B test setups: S_1 and S_2. The participants were randomly assigned to S_1 and S_2 resulting in a sample size of $N = 14$ for a paired test, respectively. S_1 was set to research whether our proposed explanation module leads to substantially lower validation time compared to the use of feature contributions plots implemented in the SHAP package. Hypothesis H1 was formulated and two concrete tasks as modifications of the defined validation task were specified (see Table 6, top).

In $T1_1$, for *Detector1D* the SHAP force plot [91] (implemented in SHAP DeepExplainer) is chosen to represent the feature contributions contained in $SHAP_{1D}$. Since this plot expects to process a tensor with one row, but $SHAP_{1D}$ is a tensor with dimensions $[15 \times 54]$, we transposed $SHAP_{1D}$ to a tensor with dimensions $[1 \times 54]$ by selecting the maximum SHAP value of each electrodes' frequency bands over the temporal dimension to allow for the visualization of feature contributions with the force plot. We refer to this as $ForcePlot_{1D}$. Furthermore, the electrode-wise explanation is disabled. For *Detector3D* we chose the SHAP image plot implemented in SHAP GradientExplainer highlighting image pixels with either blue or red, and refer to this plot as $ImagePlot_{3D}$.

S_2 evaluates the usefulness of the hybrid characteristic of XAI4EEG resulting from interacting with *Detector1D* and *Detector3D*. In particular, we want to investigate whether the hybrid characteristic lead to a substantially lower validation time compared to interacting with only one of the two proposed explainable seizure detection methods. Therefore, hypothesis H2 was formulated and two concrete tasks were formulated (see Table 6, bottom). In $T2_1$, only *Detector1D* is provided to the participants, while *Detector3D* is disabled.

We randomly selected 20 intervals as data for our user study, considering balanced class distribution. While 10 of these were shown in $T1_1$ and $T2_1$, the remaining 10 were presented in $T1_2$ and $T2_2$, respectively. To emulate the fact that clinical diagnosis is done – more often than not – under time pressure, we introduce a time constraint of 30 s for completing the tasks. After each task, the participants were asked to complete the aforementioned questionnaire with 5-point Likert scales (see Table 5).

5.3 User study results

The results obtained in the user study are statistically evaluated in this section.

Validation time evaluation Each of the 28 participants – assigned to S_1 or S_2 – works on two tasks, each encompassing 10 intervals. Since the participants were asked to

Table 6 Hypotheses and tasks performed in the user study

Setup 1	
$\mathbf{H1_{Null}}$	With the proposed explanation module the validation time is not substantially lower than with $ForcePlot_{1D}$ and $ImagePlot_{3D}$.
$\mathbf{H1_{Alternative}}$	... is substantially lower.
$T1_1$	Complete the validation task with $ForcePlot_{1D}$ and $ImagePlot_{3D}$.
$T1_2$	Complete the validation task with the proposed explanation module.
Setup 2	
$\mathbf{H2_{Null}}$	With both explainable seizure detection methods, the validation time is not substantially lower than with *Detector1D* only.
$\mathbf{H2_{Alternative}}$	... is substantially lower.
$T2_1$	Complete the validation task with *Detector1D*
$T2_2$	Complete the validation task with both explainable seizure detection methods.

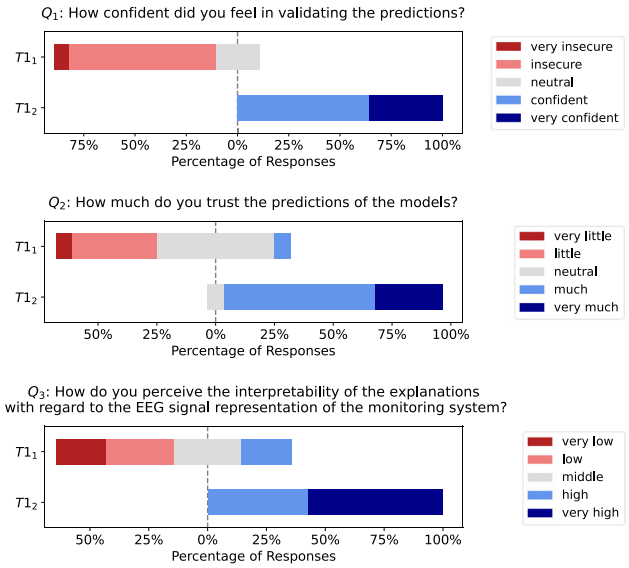

Fig. 8 Visualization of the Likert scale data that we obtained from the users in S_1. The subjective impressions on confidence (Q_1), trust (Q_2), and interpretability (Q_3) the users perceived interacting with $ForcePlot_{1D}$ and $ImagePlot_{3D}$ ($T1_1$), and with the proposed explanation module ($T1_2$) are faced

measure the above validation time, each participant records a set of 10 timestamps per task, one for each of 1...10 intervals. We refer to this set of 10 timestamps as *timestamps-set*. Thereafter, we calculate the mean of each *timestamps-set* and refer to this as *timestamps-set-mean*. Hence, in each of the four tasks we receive 14 *timestamps-set-mean* values, one for each of the 14 participants. Note that although the validation times we obtained from the participants can be evaluated quantitatively, it bases on the subjective assessment whether and when the predictions were successfully validated.

Before applying a one-sided paired *t*-test [92], we checked the difference of pairs to approximately follow a normal distribution by means of Shapiro-Wilk normality test [93]. We did not find extreme outliers in the difference of pairs. Both hypotheses are tested with a significance level of $\alpha = 0.05$ with Bessel's correction. In S_1, the resulting critical value is $c_{H1} = 1.55$, while the observed difference between both tasks is $|\bar{x}_{T2_{H1}} - \bar{x}_{T1_{H1}}| = 11.01$, where all values are given in seconds. We reject $H1_{Null}$ with p-value <0.001, i.e. the proposed visual explanations were found to lead to a substantially lower validation time. In S_2, the resulting critical value is $c_{H2} = 2.42$, while the observed difference between both tasks is $|\bar{x}_{T2_{H2}} - \bar{x}_{T1_{H2}}| = 1.94$, where all values are given in seconds. We accept $H2_{Null}$, i.e. the hybrid characteristics was not found to lead to a substantially lower validation time.

Questionnaire evaluation The Likert scale answers for S_1 are visualized in Fig. 8. Users' self-assessment on the perceived confidence (Q_1) reveals that the majority of the users felt insecure completing $T1_1$ ($ForcePlot_{1D}$ and $ImagePlot_{1D}$). In contrast, when validating the predictions supported by the explanation module ($T1_2$) the users mainly stated that they felt confident and none of them has felt insecure. In addition, the explanation module has positively impacted the level to which the users trust the predictions. In $T1_2$ the majority reported to trust the predictions "much", while in $T1_1$ merely one user stated this feeling. In regard to the third dimension of our interest, in $T1_2$ the majority of the users perceived the interpretability as "very high". By contrast, in $T1_1$ the majority experienced the interpretability as "low" or rather "middle". Only 3 users found the interpretability of $ForcePlot_{1D}$ and $ImagePlot_{1D}$ "high".

The answers for S_2 are shown in Fig. 9. During $T2_2$ supported by *Detector1D* and *Detector3D* the number of users that felt "confident" increased by half compared to $T2_1$ (*Detector1D* only). The trust in the model predictions the users experienced has increased in $T2_2$. The number of users who reported to perceive "much" and "very much" trust has doubled. Yet, one user stated to have little trust. The perceived interpretability of the explanations could not be increased in $T2_2$. While in $T2_1$ only one user stated to perceive "low" interpretability, in $T2_2$ two did. One user even stated to have experienced very low interpretability in $T2_2$.

6 Discussion

In this section, we will assess the performance indicators of the proposed deep learning models and discuss the outcomes of the user study.

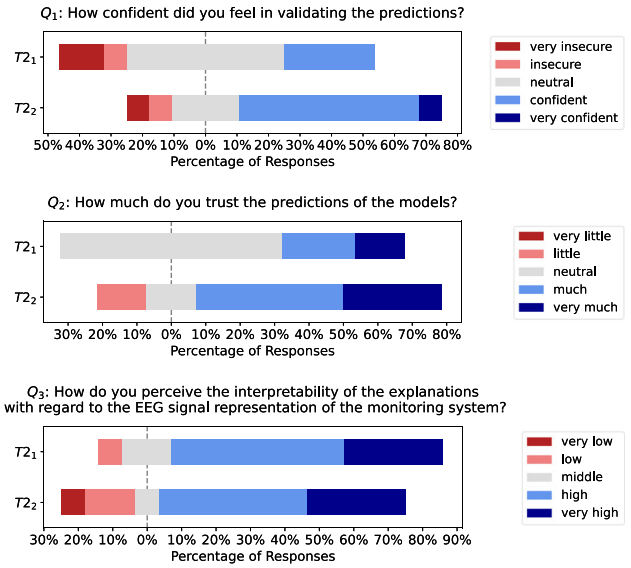

Fig. 9 Visualization of the Likert scale data that we obtained from the users in S_2. The subjective impressions on confidence (Q_1), trust (Q_2), and interpretability (Q_3) the users perceived interacting with *Detector1D* ($T2_1$), and with both explainable seizure detection approaches ($T2_2$) are faced

When comparing the performance of both models, the almost three times higher false alarm rate of the 3D-CNN strikes as compared to the 1D-CNN. This may be due the fact that we did not remove artifacts from the neonatal EEG recordings. The authors of [94] notice a significant decrease in false alarms when artifacts are removed from neonatal EEG, since artifacts show similar characteristics as seizures. However, this does not impair the fundamental concept of XAI4EEG, i.e. the novel mapping of SHAP values to the explanation module.

The authors of [95, 96] highlight the ambiguity of experienced experts in visual inspection of multi-channel EEG for neonatal seizure detection. Our experiments do confirm that the two learning methods returned disagreeing explanation patterns in several instances (i.e. intervals). Our explanation module cannot eliminate the disagreement but highlights it for inspection by the medical expert. This constitutes an advantage of a hybrid detection incorporating an identical visual explanation schema where one learning module may miss to capture and therefore to present decision-relevant factors, while the other learning module does and therefore augments the explanation. In addition, this constitutes a way for evaluating an explanation [97]. The extent to which this has an impact on users when validating the algorithmic predictions must be investigated in future work.

In the first (S_1) of the two setups of our user study, we studied the effectiveness of our explanation module as compared to SHAP force plot and SHAP image plot. To this end, the authors of [89] underline the usefulness of assessing how various explanation schemata impact clinical-decision makers. The explanation module is specific to neonatal seizures, since the integrated band-pass filter retains a frequency band of 0.5 Hz ...12.5 Hz where neonatal seizures primarily emerge. To be remarked, the explanation module used to visualize the SHAP values can, however, be transferred to any interval lengths and higher frequency bands without loss of generality.

Moreover, we argue that our proposed novel mapping of SHAP values to the explanation module permits generalization to other problem domains, especially promising where information is comprised in time and frequency. Besides EEG data, this could be the analysis of electrocardiogram (ECG) signals, e.g. for the detection of left ventricular hypertrophy [98]. Beyond the medical domain, another target domain could be ML-enabled predictive maintenance in automotive applications, e.g. the detection of engine [99] and gearbox [100] faults from vibration signals. In that context, the authors of [101] have already highlighted the need for interpretability.

Since our preprocessing and feature extraction steps maintain spectral, spatial, and temporal dimensions of the EEG signal, we do not ignore the fact of multi-channel signal processing [68] and can therefore incorporate feature contributions from all dimensions into our explanation module. Hence, from the medical perspective, our proposed novel mapping of SHAP values enables an analysis of decision-relevant electrodes (spatial dimensions), frequency bands (spectral dimensions) and subsequences (temporal dimensions).

As a final remark, the concept of XAI4EEG constitutes a prototype that has to be further customized, improved and validated before an implementation into established medical workflows and routines is feasible [53].

7 Conclusion

In this work, we introduced XAI4EEG: an application-aware approach for an explainable and hybrid deep learning-based detection of seizures in multivariate EEG time series. In XAI4EEG, we combined deep learning models and domain knowledge on seizure detection, namely (a) frequency bands, (b) location of EEG leads and (c) temporal characteristics. From the technical perspective, XAI4EEG encompasses EEG data preparation, two deep learning models, and the proposed explanation module. Intuitive post-hoc explainability arises from a novel flow of patterns that maps the feature contributions to the explanation module that are obtained by two SHAP explainers, each explaining the predictions of one of the models. As a result, the generated visual explanations leverage an identification of decision-relevant regions in

the (a) spectral, (b) spatial and (c) temporal EEG dimensions that are all crucial for seizure detection. From the medical perspective, XAI4EEG promotes clinical experts and decision-makers in validating the algorithmic predictions by providing a full-fledged explanation, before a final clinical decision must be met.

In an initial user study, we reported on the effectiveness of the explanation module and show that it leads to a substantially lower time for validating the predictions compared to selected feature contribution plots implemented in the SHAP package. Furthermore, the explanation module leads to increased interpretability, trust in predictions, and confidence in validation. Although interacting with both proposed explainable seizure detection methods in XAI4EEG did not result in a significant decrease of validation time, users stated to feel more confident and experienced an increase in trust. Moreover, while a single detection method may fail to capture decision-relevant factors in some instances, a further method could do so, thus augmenting the explanation.

7.1 Limitations and future work

We did not use the entire data set instead we selected a subset of neonatal recordings, and did not remove signal artifacts. This may have affected our classifiers' performance. While our explanation module can, however, be incorporated into other EEG-related problem domains, the proposed novel mapping of SHAP values cannot. Although the underlying idea remains the same, it is specific to the used data set, the data preparation steps, the classifier, and the explainer. In our user study, the set of volunteers was small, gender distribution was uneven, all participants were younger than a typical medical expert and had no medical expertise. Thus, our experimental findings cannot be generalized towards usability of the method by medical experts. We rather see this evaluation as preliminary, before recruiting medical experts for interaction with XAI4EEG.

Future work could be user studies with medical experts/staff, to verify the usefulness of XAI4EEG for clinical decision making. This would also allow to examine the clinical importance of the observed decreased validation time. Moreover, one could study the global feature contributions of our learning modules. Evaluating to what extent XAI4EEG is suited to fit medical education [57], i.e. for prospective medical experts in NICUs, is also a point to be addressed in future work. Furthermore, there is scope to improve the performance of our seizure detection methods, e.g. by processing the whole data set.

Author Contributions DR: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing – Original Draft, Visualization, Supervision, Writing - Review & Editing. AT: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing – Original Draft, Visualization, Supervision, Writing - Review & Editing. MS: Conceptualization, Methodology, Visualization, Writing - Review & Editing.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors have no relevant financial or non-financial interests to disclose.

Data availability Code and data availability: www.ml-and-vis.org/xai4eeg.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Popel M, Tomkova M, Tomek J et al (2020) Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. Nat Commun. https://doi.org/10.1038/s41467-020-18073-9
2. McKinney SM, Sieniek M, Godbole V et al (2020) International evaluation of an ai system for breast cancer screening. Nature 577(7788):89–94. https://doi.org/10.1038/s41586-019-1799-6
3. Wu N, Phang J, Park J et al (2020) Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imag 39(4):1184–1194. https://doi.org/10.1109/TMI.2019.2945514
4. Assael YM, Shillingford B, Whiteson S, et al (2016) Lipnet: end-to-end sentence-level lipreading arXiv:1611.01599
5. Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8):832. https://doi.org/10.3390/electronics8080832
6. Theissler A, Pérez-Velázquez J, Kettelgerdes M et al (2021) Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliab Eng Syst Saf 215(107):864. https://doi.org/10.1016/j.ress.2021.107864
7. Gruner T et al (2020) Evaluation of machine learning for sensorless detection and classification of faults in electromechanical drive systems. Procedia Computer Sci. https://doi.org/10.1016/j.procs.2020.09.170
8. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access

6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
9. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Digital Signal Process 73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011
10. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps arXiv:1312.6034
11. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B et al (eds) Computer Vision - ECCV 2014. Springer, Cham, pp 818–833
12. Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery Queue 16(3):31–57
13. Ribeiro MT, Singh S, Guestrin C (2016) Why Should I Trust You? In: Krishnapuram B, Shah M, Smola A et al (eds) KDD2016 Association for Computing Machinery Inc (ACM), New York, NY, pp 1135–1144 https://doi.org/10.1145/2939672.2939778
14. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning arXiv:1702.08608
15. Theissler A, Vollert S, Benz P, et al (2020) ML-ModelExplorer: an explorative model-agnostic approach to evaluate and compare multi-class classifiers In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, pp 281–300 https://doi.org/10.1007/978-3-030-57321-8_16
16. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions In: Proc 31st Int Conf on NeurIPS Curran Associates Inc, Red Hook, NY, USA, pp 4768-4777
17. Vollert S, Atzmueller M, Theissler A (2021) interpretable machine learning: a brief survey from the predictive maintenance perspective In: IEEE International Conference on Emerging Technologies and Factory Automation (ETFA 2021) IEEE
18. Francesco P, Fabio G, Edoardo P et al (2021) Artificial intelligence and healthcare: forecasting of medical bookings through multi-source time-series fusion. Information Fusion 74:1–16. https://doi.org/10.1016/j.inffus.2021.03.004
19. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med. https://doi.org/10.1038/s41591-018-0300-7
20. Varshney KR, Alemzadeh H (2017) On the safety of machine learning: cyber-physical systems, decision sciences, and data products. Big Data 5(3):246–255
21. Oh SL, Hagiwara Y, Raghavendra U et al (2020) A deep learning approach for parkinson's disease diagnosis from eeg signals. Neural Comput Appl 32(15):10927–10933. https://doi.org/10.1007/s00521-018-3689-5
22. Buettner R, Frick J, Rieg T (2019) High-performance detection of epilepsy in seizure-free EEG recordings: a novel machine learning approach using very specific epileptic EEG sub-bands In: Krcmar H, Fedorowicz J, Boh WF, et al (eds) Proc 40th Int Conf Information Systems, ICIS 2019, Munich, Germany, December 15–18, 2019 Association for Information Systems
23. Sun J, Cao R, Zhou M et al (2021) A hybrid deep neural network for classification of schizophrenia using eeg data. Sci Rep 11(1):4706. https://doi.org/10.1038/s41598-021-83350-6
24. Zhou M, Tian C, Cao R et al (2018) Epileptic seizure detection based on eeg signals and cnn. Front Neuroinform 12:95. https://doi.org/10.3389/fninf.2018.00095
25. Ngugi AK, Bottomley C, Kleinschmidt I et al (2010) Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. Epilepsia 51(5):883–890. https://doi.org/10.1111/j.1528-1167.2009.02481.x
26. Devinsky O, Vezzani A, O'Brien TJ et al (2018) Epilepsy. Nat Rev Dis Primers. https://doi.org/10.1038/nrdp.2018.24
27. Annegers JF, Hauser WA, Lee JR et al (1995) Incidence of acute symptomatic seizures in rochester, minnesota, 1935–1984. Epilepsia 36(4):327–333. https://doi.org/10.1111/j.1528-1157.1995.tb01005.x
28. Rennie J, Boylan G (2007) Treatment of neonatal seizures. Arch Dis Child Fetal Neonatal Ed 92(2):F148-50. https://doi.org/10.1136/adc.2004.068551
29. Panayiotopoulos CP (2010) Neonatal epileptic seizures and neonatal epileptic syndromes. In: Panayiotopoulos CP (ed) A clinical guide to epileptic syndromes and their treatment. Springer, London. https://doi.org/10.1007/978-1-84628-644-5_8
30. Cowan LD (2002) The epidemiology of the epilepsies in children. Ment Retard Dev Disabil Res Rev 8(3):171–181. https://doi.org/10.1002/mrdd.10035
31. Cowan F, Rutherford M, Groenendaal F et al (2003) Origin and timing of brain lesions in term infants with neonatal encephalopathy. Lancet 361(9359):736–742. https://doi.org/10.1016/S0140-6736(03)12658-X
32. Volpe JJ (ed) (2018) Volpe's neurology of the newborn 6th edn Elsevier, Philadelphia https://doi.org/10.1016/C2010-0-68825-0
33. Padiyar S, Nusairat L, Kadri A et al (2020) Neonatal seizures in the us national inpatient population: prevalence and outcomes. Pediatr Neonatol 61(3):300–305. https://doi.org/10.1016/j.pedneo.2019.12.006
34. Boylan G, Burgoyne L, Moore C et al (2010) An international survey of eeg use in the neonatal intensive care unit. Acta Paediatr 99(8):1150–1155. https://doi.org/10.1111/j.1651-2227.2010.01809.x
35. Mizrahi EM, Kellaway P (1998) Diagnosis and management of neonatal seizures. Lippincott-Raven, Philadelphia
36. Temko A, Thomas E, Marnane W et al (2011) Eeg-based neonatal seizure detection with support vector machines. J Clin Neurophysiol 122(3):464–473. https://doi.org/10.1016/j.clinph.2010.06.034
37. Tsuchida TN, Wusthoff CJ, Shellhaas RA et al (2013) American clinical neurophysiology society standardized eeg terminology and categorization for the description of continuous eeg monitoring in neonates: report of the american clinical neurophysiology society critical care monitoring committee. J Clin Neurophysiol 30(2):161–173. https://doi.org/10.1097/WNP.0b013e3182872b24
38. Srinivasakumar P, Zempel J, Trivedi S et al (2015) Treating eeg seizures in hypoxic ischemic encephalopathy: a randomized controlled trial. Pediatrics 136(5):e1302–e1309. https://doi.org/10.1542/peds.2014-3777
39. Shellhaas RA (2015) Continuous long-term electroencephalography: the gold standard for neonatal seizure diagnosis. Semin Fetal Neonatal Med 20(3):149–153. https://doi.org/10.1016/j.siny.2015.01.005
40. Duncan JS, Sander JW, Sisodiya SM et al (2006) Adult epilepsy. Lancet 367(9516):1087–1100. https://doi.org/10.1016/S0140-6736(06)68477-8
41. Mitra J, Glover JR, Ktonas PY et al (2009) A multistage system for the automated detection of epileptic seizures in neonatal electroencephalography. Clin Neurophysiol 26(4):218–226. https://doi.org/10.1097/WNP.0b013e3181b2f29d
42. Nagaraj SB, Stevenson NJ, Marnane WP et al (2014) Neonatal seizure detection using atomic decomposition with a novel dictionary. IEEE Trans Biomed Eng 61(11):2724–2732. https://doi.org/10.1109/TBME.2014.2326921
43. Ansari AH, Cherian PJ, Caicedo A et al (2019) Neonatal seizure detection using deep convolutional neural networks. Int J Neural Syst 29(4):1850011. https://doi.org/10.1142/S0129065718500119

44. Cho KO, Jang HJ (2020) Comparison of different input modalities and network structures for deep learning-based seizure detection. Sci Rep 10(1):122. https://doi.org/10.1038/s41598-019-56958-y
45. Gómez C, Arbeláez P, Navarrete M et al (2020) Automatic seizure detection based on imaged-eeg signals through fully convolutional networks. Sci Rep 10(1):21833. https://doi.org/10.1038/s41598-020-78784-3
46. Wagholikar KB, Sundararajan V, Deshpande AW (2012) Modeling paradigms for medical diagnostic decision support: a survey and future directions. J Med Syst 36(5):3029–3049. https://doi.org/10.1007/s10916-011-9780-4
47. Naderpour M, Lu J, Zhang G (2014) An intelligent situation awareness support system for safety-critical environments. Decis Support Syst 59:325–340. https://doi.org/10.1016/j.dss.2014.01.004
48. Alarcon G, Binnie C, Elwes R et al (1995) Power spectrum and intracranial eeg patterns at seizure onset in partial epilepsy. Electroencephalogr Clin Neurophysiol 94(5):326–337. https://doi.org/10.1016/0013-4694(94)00286-T
49. Zhang Y, Guo Y, Yang P et al (2020) Epilepsy seizure prediction on eeg using common spatial pattern and convolutional neural network. IEEE J Biomed Health Inform 24(2):465–474. https://doi.org/10.1109/JBHI.2019.2933046
50. Karafin M, St Louis EK, Zimmerman MB et al (2010) Bimodal ultradian seizure periodicity in human mesial temporal lobe epilepsy. Seizure 19(6):347–351. https://doi.org/10.1016/j.seizure.2010.05.005
51. Greenberg MD (2009) Medical malpractice and new devices: defining an elusive standard of care. Health Matrix 19(2):423–445
52. Wiens J, Saria S, Sendak M et al (2019) Do no harm: a roadmap for responsible machine learning for health care. Nat Med 25(9):1337–1340. https://doi.org/10.1038/s41591-019-0548-6
53. Hacker P, Krestel R, Grundmann S et al (2020) Explainable ai under contract and tort law: legal incentives and technical challenges. Artif Intell Law 28(4):415–439. https://doi.org/10.1007/s10506-020-09260-6
54. Rory S, Ankur T, Ehsan R et al (2019) Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 126(4):552–564. https://doi.org/10.1016/j.ophtha.2018.11.016
55. Holzinger A (2020) Explainable ai and multi-modal causability in medicine. i-com 19(3):171–179. https://doi.org/10.1515/icom-2020-0024
56. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x
57. Holzinger A, Langs G, Denk H et al (2019) Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 9(4):e1312. https://doi.org/10.1002/widm.1312
58. Temko A, Lightbody G (2016) Detecting neonatal seizures with computer algorithms. J Clin Neurophysiol 33(5):394–402. https://doi.org/10.1097/WNP.0000000000000295
59. Fellous JM, Sapiro G, Rossi A et al (2019) Explainable artificial intelligence for neuroscience: behavioral neurostimulation. Front Neurosci 13:1346. https://doi.org/10.3389/fnins.2019.01346
60. Pintelas E, Liaskos M, Livieris IE et al (2021) A novel explainable image classification framework: case study on skin cancer and plant disease prediction. Neural Comput Appl 33(22):15171–15189. https://doi.org/10.1007/s00521-021-06141-0
61. Ieracitano C, Mammone N, Hussain A et al (2021) A novel explainable machine learning approach for eeg-based brain-computer interface systems. Neural Comput Appl. https://doi.org/10.1007/s00521-020-05624-w
62. Hu X, Yuan S, Xu F et al (2020) Scalp eeg classification using deep bi-lstm network for seizure detection. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2020.103919
63. Li Y, Yu Z, Chen Y et al (2020) Automatic seizure detection using fully convolutional nested lstm. Int J Neural Syst. https://doi.org/10.1142/S0129065720500197
64. Abdelhameed AM, Daoud HG, Bayoumi M (2018) Deep convolutional bidirectional lstm recurrent neural network for epileptic seizure detection In: 16th IEEE Int New Circuits Syst Conf (NEWCAS) IEEE, pp 139–143 https://doi.org/10.1109/NEWCAS.2018.8585542
65. O'Shea A, Lightbody G, Boylan G, et al (2017) Neonatal seizure detection using convolutional neural networks In: 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP) IEEE, pp 1–6 https://doi.org/10.1109/MLSP.2017.8168193
66. Yuan Q, Zhou W, Zhang L et al (2017) Epileptic seizure detection based on imbalanced classification and wavelet packet transform. Seizure 50:99–108. https://doi.org/10.1016/j.seizure.2017.05.018
67. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition arXiv:1409.1556
68. Denemark T, Fridrich J, Comesaña P (2016) Improving selection-channel-aware steganalysis features. Electron Imag 2016:1–8. https://doi.org/10.2352/ISSN.2470-1173.2016.8.MWSF-080
69. Wei X, Zhou L, Chen Z et al (2018) Automatic seizure detection using three-dimensional cnn based on multi-channel eeg. BMC Med Inform Decis Mak 18(5):111. https://doi.org/10.1186/s12911-018-0693-8
70. Zhang B, Wang W, Xiao Y et al (2020) Cross-subject seizure detection in eegs using deep transfer learning. Comput Math Methods Med. https://doi.org/10.1155/2020/7902072
71. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778 https://doi.org/10.1109/CVPR.2016.90
72. Mansour M, Khnaisser F, Partamian H (2020) An explainable model for eeg seizure detection based on connectivity features arXiv:2009.12566
73. Zhang X, Yao L, Dong M et al (2020) Adversarial representation learning for robust patient-independent epileptic seizure detection. IEEE J Biomed Health Inform 24(10):2852–2859. https://doi.org/10.1109/JBHI.2020.2971610
74. Uyttenhove T, Maes A, van Steenkiste T, et al (2020) Interpretable epilepsy detection in routine, interictal eeg data using deep learning In: Alsentzer E, McDermott MBA, Falck F, et al (eds) Proc Machine Learning Health NeurIPS Workshop, Proceedings of Machine Learning Research, vol 136 PMLR, pp 355–366
75. Dissanayake T, Fernando T, Denman S et al (2021) Deep learning for patient-independent epileptic seizure prediction using scalp eeg signals. IEEE Sens J 21(7):9377–9388. https://doi.org/10.1109/JSEN.2021.3057076
76. Valentin G, Tomas T, Marina Z et al (2021) Interpreting deep learning models for epileptic seizure detection on eeg signals. Artif Intell Med 117(102):084. https://doi.org/10.1016/j.artmed.2021.102084
77. Theissler A (2017) Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. Knowl-Based Syst 123:163–173. https://doi.org/10.1016/j.knosys.2017.02.023

78. Kitayama M, Otsubo H, Parvez S et al (2003) Wavelet analysis for neonatal electroencephalographic seizures. Pediatr Neurol 29(4):326–333. https://doi.org/10.1016/S0887-8994(03)00277-7
79. Delorme A, Makeig S (2004) Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. J Neurosci Methods 134(1):9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009
80. Welch P (1967) The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 15(2):70–73. https://doi.org/10.1109/TAU.1967.1161901
81. Ahirwal MK, Londhe N (2012) Power spectrum analysis of eeg signals for estimating visual attention. Int J Comput Appl 42:34–40
82. van Vugt MK, Sederberg PB, Kahana MJ (2007) Comparison of spectral analysis methods for characterizing brain oscillations. J Neurosci Methods 162(1–2):49–63. https://doi.org/10.1016/j.jneumeth.2006.12.004
83. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
84. Teolis A (1998) Computational signal processing with wavelets. Applied and Numerical Harmonic Analysis Birkhäuser Boston, Boston, MA https://doi.org/10.1007/978-1-4612-4142-3
85. Stevenson NJ, Tapani K, Lauronen L et al (2019) A dataset of neonatal eeg recordings with seizure annotations. Sci Data. https://doi.org/10.1038/sdata.2019.39
86. de Weerd AW, Despland PA, Plouin P (1999) Neonatal eeg the international federation of clinical neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 52:149–157
87. O'Shea A, Lightbody G, Boylan G et al (2020) Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture. Neural Net 123:12–25. https://doi.org/10.1016/j.neunet.2019.11.023
88. Price W, Nicholson II, Gerke S, Cohen IG (2019) Potential liability for physicians using artificial intelligence. JAMA 322(18):1765–1766. https://doi.org/10.1001/jama.2019.15064
89. Lauritsen SM, Kristensen M, Olsen MV et al (2020) Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 11(1):3852. https://doi.org/10.1038/s41467-020-17431-x
90. Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 22(140):1–55
91. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK, Newman SF, Kim J, Lee SI (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2(10):749–760. https://doi.org/10.1038/s41551-018-0304-0
92. Student (1908) The probable error of a mean. Biometrika 6(1):1 https://doi.org/10.2307/2331554
93. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591. https://doi.org/10.2307/2333709
94. de Vos M, Deburchgraeve W, Cherian PJ et al (2011) Automated artifact removal as preprocessing refines neonatal seizure detection. Clin Neurophysiol 122(12):2345–2354. https://doi.org/10.1016/j.clinph.2011.04.026
95. Stevenson NJ, Clancy RR, Vanhatalo S et al (2015) Interobserver agreement for neonatal seizure detection using multichannel eeg. Ann Clin Transl Neurol 2(11):1002–1011. https://doi.org/10.1002/acn3.249
96. Stevenson NJ, Lauronen L, Vanhatalo S (2018) The effect of reducing eeg electrode number on the visual interpretation of the human expert for neonatal seizure detection. Clin Neurophysiol 129(1):265–270. https://doi.org/10.1016/j.clinph.2017.10.031
97. Stiglic G, Kocbek P, Fijacko N et al (2020) Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. https://doi.org/10.1002/widm.1379
98. Jothiramalingam R, Jude A, Patan R et al (2021) Machine learning-based left ventricular hypertrophy detection using multi-lead ecg signal. Neural Comput Appl 33(9):4445–4455. https://doi.org/10.1007/s00521-020-05238-2
99. Meng-Hui W, Kuei-Hsiang C, Wen-Tsai S et al (2010) Using enn-1 for fault recognition of automotive engine. Expert Syst Appl 37(4):2943–2947. https://doi.org/10.1016/j.eswa.2009.09.041
100. Heidari BH, Abdolreza O (2014) Application of wavelet energy and shannon entropy for feature extraction in gearbox fault detection under varying speed conditions. Neurocomputing 133:437–445. https://doi.org/10.1016/j.neucom.2013.12.018
101. Andreas T, Judith P-V, Marcel K et al (2021) Predictive maintenance enabled by machine learning: use cases and challenges in the automotive industry. Reliab Eng Syst Saf 215(107):864. https://doi.org/10.1016/j.ress.2021.107864