Computer Methods and Programs in Biomedicine 257 (2024) 108448

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Concept-based AI interpretability in physiological time-series data: Example of abnormality detection in electroencephalography

Alexander Brenner [a,*], Felix Knispel [b], Florian P. Fischer [c], Peter Rossmanith [d], Yvonne Weber [c], Henner Koch [c], Rainer Röhrig [b], Julian Varghese [a], Ekaterina Kutafina [e]

[a] Institute of Medical Informatics, University of Münster, Münster, Germany
[b] Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany
[c] Department of Epileptology and Neurology, Medical Faculty, RWTH Aachen University Hospital, Aachen, Germany
[d] Theoretical Computer Science, Department of Computer Science, RWTH Aachen University, Aachen, Germany
[e] Institute for Biomedical Informatics, Faculty of Medicine, University Hospital Cologne, University of Cologne, Cologne, Germany

ARTICLE INFO

Keywords:
Artificial neural networks
Decision support systems
Electroencephalography
Explainable artificial intelligence
Supervised learning
TCAV

ABSTRACT

Background and Objective: Despite recent performance advancements, deep learning models are not yet adopted in clinical practice on a wide scale. The intrinsic intransparency of such systems is commonly cited as one major reason for this reluctance. This has motivated methods that aim to provide explanations of model functioning. Known limitations of feature-based explanations have led to an increased interest in concept-based interpretability. *Testing with Concept Activation Vectors* (TCAV) employs human-understandable, abstract concepts to explain model behavior. The method has previously been applied to the medical domain in the context of electronic health records, retinal fundus images and magnetic resonance imaging.
Methods: We explore the usage of TCAV for building interpretable models on physiological time series, using an example of abnormality detection in electroencephalography (EEG). For this purpose, we adopt the XceptionTime model, which is suitable for multi-channel physiological data of variable sizes. The model provides state-of-the-art performance on raw EEG data and is publically available. We propose and test several ideas regarding concept definition through metadata mining, using additional labeled EEG data and extracting interpretable signal characteristics in the form of frequencies. By including our own hospital data with analog labeling, we further evaluate the robustness of our approach.
Results: The tested concepts show a TCAV score distribution that is in line with the clinical expectations, i.e. concepts known to have strong links with EEG pathologies (such as epileptiform discharges) received higher scores than the neutral concepts (e.g. sex). The scores were consistent across the applied concept generation strategies.
Conclusions: TCAV has the potential to improve interpretability of deep learning applied to multi-channel signals as well as to detect possible biases in the data. Still, further work on developing the strategies for concept definition and validation on clinical physiological time series is needed to better understand how to extract clinically relevant information from the concept sensitivity scores.

1. Introduction

Due to significant performance improvements of deep learning systems, their usage could become increasingly viable in clinical decision support systems, where medical data is evaluated automatically in order to provide additional guidance to medical professionals. Still, artificial intelligence remains underused in medical healthcare, which may lead to potential risks of missed opportunities and heightened costs in public health [1]. Clinicians have been reluctant to adopt such methods in real-world clinical practice due to a number of concerns. One of such concerns is a lack of explainability of complex deep learning models [2,3]. This goes hand in hand with the unclear accountability for decisions based on AI recommendations or decisions [4,5]. Traceability and interpretability are key factors for legal requirements, as introduced in the legislation for “high-risk” artificial intelligence and further investigated by the European Commision [6]. Both, network size and

* Corresponding author.
E-mail address: alexander.brenner@uni-muenster.de (A. Brenner).

https://doi.org/10.1016/j.cmpb.2024.108448
Received 10 June 2024; Received in revised form 2 September 2024; Accepted 29 September 2024
Available online 30 September 2024

architecture, as well as input feature complexity contribute to the intransparency of a network's decision processes. This problematic characteristic of deep learning is increasingly being addressed in a research effort towards more interpretable machine learning far beyond the medical domain. Many already well-established interpretability methods, such as saliency maps [7], strive to explain model behavior by calculating importance scores of individual input features without altering the underlying model. As we have discussed in our previous work, these feature attribution methods need to be used in conjunction with interpretable input features, which comes at the risk of restricting model performance due to poor feature presentation [8]. Since saliency maps focus on local feature importance, they do not consider the network's overall understanding of the data.

A recent alternative approach of providing insight into model explanations is the usage of human-understandable, high-level *concepts*. Testing with Concept Activation Vectors (TCAV) [9] is a post hoc interpretability method, where the user defines a set of concepts of interest beforehand. The authors illustrated TCAV by investigating the sensitivity of a model to the concept of *stripes* when classifying whether an image depicts a *zebra*. The concept of stripes is defined by providing TCAV with images of isolated stripes. This approach has been used in the context of medical data as well: The authors of the original TCAV paper have defined clinically relevant concepts for images of diabetic retinopathy [9], others have used TCAV in cardiac MRI images [10]. Structured temporal data in the form of electronic health records have been used in conjunction with TCAV as well [11].

In the context of biosignal data, the potential of the TCAV approach still needs to be explored. Biosignals typically refer to measurable indicators derived from the human body with generally high temporal, but low spatial information (typically, physiological time-series). They are an integral part of various healthcare domains, enabling healthcare professionals to monitor, diagnose, and treat a wide range of medical conditions. To our best knowledge, at the moment of submission only one study applied TCAV on a transformer model for the classification of electroencephalography (EEG) data [12].

Here, we also focus on the EEG that is used as a diagnostic and monitoring tool for various neurological disorders such as epilepsy and sleep disorders. EEG interpretation is a complex and time-consuming task that requires specialized expertise. The subjective nature of manual EEG analysis can lead to inter-rater variability and diagnostic errors. By leveraging machine learning techniques, the research community aims to develop objective and automated classification models. These could assist healthcare professionals in making more accurate and reliable assessments, as well as reduce analysis time. While machine learning models have achieved promising classification accuracies in the detection of abnormal EEGs [8,13-16], the nature of these models renders the classification process untraceable for humans. We therefore propose the usage of TCAV to mitigate the issues of lacking interpretability. In comparison to the original application of TCAV to image data, additional challenges such as different record lengths and temporally restricted signal patterns must be taken into account.

While exploring a specific use-case, we are also tackling a more general research question: how to construct the concept-testing process in the context of medical biosignal analysis? We make use of a state-of-the-art architecture designed for time-series classification. Further, we employ time-labeled EEG samples for the definition of concepts that represent certain abnormal signal patterns, such as epileptiform discharges. Other concepts are derived from the available metadata and from the spectral content of the data. The application of TCAV in this context allows us to investigate whether the trained model exhibits sensitivities to concepts that are typically considered to be clinically relevant.

2. Methods

The Temple University Hospital (TUH) EEG Corpus [17] has established a comprehensive collection of EEG records from hospital patients. For the TUH Abnormal EEG Corpus (TUAB), a standard benchmarking subset of the database, trained specialists categorized records into normal and abnormal EEGs [13]. To answer our research question, we performed the following workflow, consisting of four stages (Fig. 1):

1. **Model training:** The train/test split of the TUAB set is used for training and evaluation of the deep learning model. Note that, instead of training a new model from scratch, any readily-trained neural network model for this task could be used in the subsequent stages. We include the model training to allow for increased control and transparency of the process.
2. **Definition of concepts of interest:** Concepts are defined by a batch of exemplary data. Relevant concepts are extracted from labeled samples, metadata, or generated synthetically. These samples are not part of the model training.
3. **Learning concept activation vectors:** Given the concept examples and the trained model, concept activation vectors (CAVs) are computed for a selection of network layers. These serve as an abstract representation of the concept in the latent space of a specific layer.
4. **Testing for concept sensitivity:** Given the CAVs, the model and a batch of test samples with known target classes (abnormal or normal), TCAV is applied to compute the concept's influences on the decision process. The final results are concept sensitivity scores that are aggregated from the set of test samples and thus give a global explanation on how important the concepts are for the classification of the tested target class.

2.1. Data

All data from public sources were taken from the TUH database that is openly available for research purposes. The TUH study has been ethically approved and patient consent was obtained for data use [17]. The database provides several annotated subsets, while the TUH EEG Corpus (TUEG) comprises all of the available clinical EEGs. One of the subsets, the TUAB dataset, has been thoroughly analyzed by various researchers, where a wide range of machine learning methods have been evaluated in the detection of abnormal EEGs. In particular, deep learning methods have shown promising accuracy in the distinction of normal and abnormal records. The current TUAB dataset (v3.0.0) provides 2993 EEG sessions, where each recording was labeled by trained specialists as normal or abnormal (details on the labeling process can be derived from the original TUH publication [13]). Abnormalities, such as transient spikes and sharp waves or repetitive epileptiform discharges, are for example exhibited by patients with epilepsy. Additionally, the dataset is split into a train-set (2717 samples) and a test-set (276 samples). The class distribution is balanced for both sets. We cut the data into crops of 60 s extracted from each record. For the train set, we employ a 30 s overlap and use up to eight minutes of record time to increase training size. For the test set, we only use the first 60 s to ensure comparability to previous studies [13,15]. For generating concept samples, we extracted 2613 sample crops from the TUH EEG Events (TUEV) subset (v2.0.0) and further selected 563 unrelated samples from the TUEG dataset (v2.0.0) [18], see the git repository [19] for the complete sample list.

Additionally, we used our own data from the University Hospital RWTH Aachen (UKA). An experienced epileptologist viewed through video EEGs from 45 patients (25 males with an average age of 45, 20 females with an average of 42) recorded in the epilepsy monitoring unit and annotated the records with similar labels as used in the TUEV dataset. This study was approved by the ethics committee of the Faculty of Medicine at the RWTH Aachen University (EK 335/21, CTC-A 21–254). The participants provided their written informed consent, and the study was conducted according to the Declaration of Helsinki.

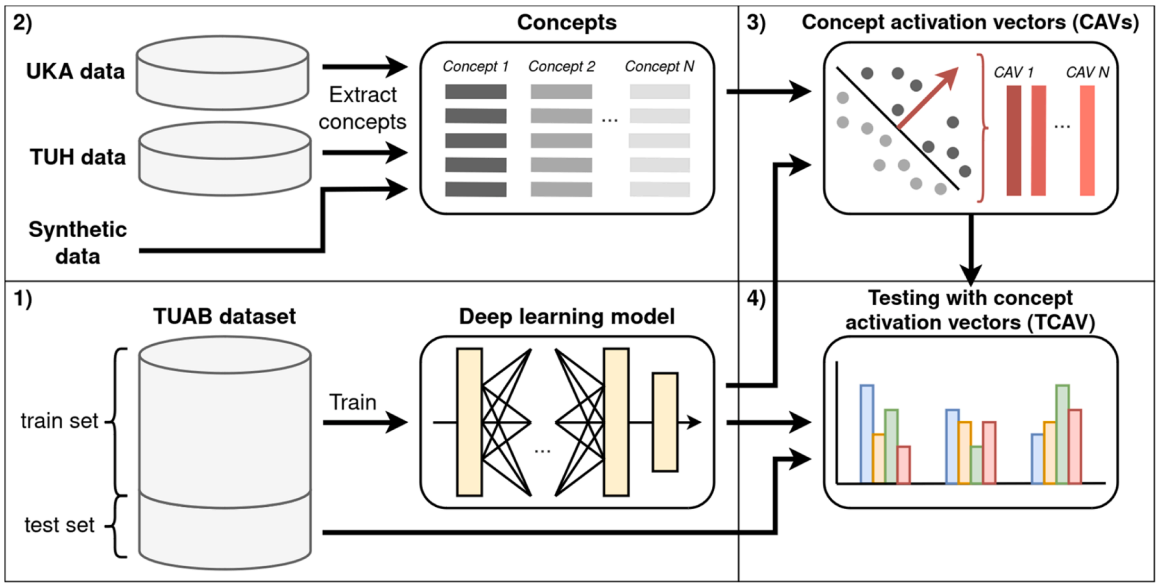

Fig. 1. Overview of the experiment workflow. 1) Deep learning model trained on TUAB data, 2) Concept extraction from labeled EEGs from the University Hospital RWTH Aachen (UKA), the Temple University Hospital (TUH) and simulated data, 3) Computation of CAVs, 4) Testing for concept sensitivity.

We used 19 channels to be compatible with most of the available records, the overview of available channels in the TUH dataset is described in [20]. All EEGs were transformed to match the Temporal Central Parasagittal (TCP) montage, as suggested by the TUH authors. Applying the montage resulted in 20 channels per sample. Further, we resampled each file to a uniform 150 Hz.

2.2. Deep learning model for normal-abnormal EEG classification

For the selection of a suitable deep learning model, we considered different architectures. Our requirements for models were that they are capable of reading raw time-series of variable lengths and achieve competitive classification performance. While ChronoNet has shown among the highest classification scores according to the original report, Khan et al. reported results that lag behind after an open-source re-implementation and evaluation [21]. Alhussein et al. reported even higher accuracy, but lacked details on pre-training and published code [16]. While Schirrmeister et al. made their implementation available, the method aggregates the predictions of short crops of the signal [14]. Further, as we encounter signal crops of different length when working with the annotations of the TUEV dataset, we need the model to allow for arbitrary input lengths. Therefore, we decided to use an alternative architecture that has been specifically designed for multi-channel time-series data. We implemented the XceptionTime model in PyTorch using the library *tsai* [22]. The architecture was initially designed and evaluated on surface electromyography (sEMG) signals [23]. Following the code notation from *tsai,* we categorized the network layers into a backbone holding a series of XceptionTime modules with residual connections, as well as the network head consisting of Adaptive Average Pooling and Convolution (1×1) layers (Fig. 2). Each XceptionTime module in the backbone concatenates the output of three sets of Depth-wise Separable Convolution, each with a different kernel size (11, 21 or 41). With the stride set to one and additional padding, the output length of the transferred data in the backbone remains unchanged. The first Adaptive Average Pooling layer of the network head transforms the output of the backbone to a fixed length of 50, allowing the model to be applied to signals of variable lengths. Stacked Convolution layers finally transform the channel dimension to the number of classes, while the length of the input signal is reduced to one in the second Adaptive Average Pooling layer. The model was configured to use 20-channel EEG samples (according to the TCP montage) as input and provides a binary classification of abnormality as output. We trained the model for a maximum of 100 epochs on the training set of the preprocessed TUAB dataset. The Adam optimization algorithm was used together with the one cycle learning policy and early stopping based on internal validation (10 % of the training data, grouped by subject).

2.3. Application of concept-based interpretability

We implemented the concept-testing process using the TCAV functionality provided in the Python library *captum* [24]. To determine the relevance of a certain concept, TCAV is typically used by comparing the concept of interest C to a "random" concept, constructed from randomly selected samples of data that are unrelated by nature, but of the same type and format. Kim et al. [9] sample from the large number of different types of images in the ImageNet dataset to construct random concepts with no shared pattern or structure [25]. For a selected layer l of the network, linear classifiers are trained on the embedding space

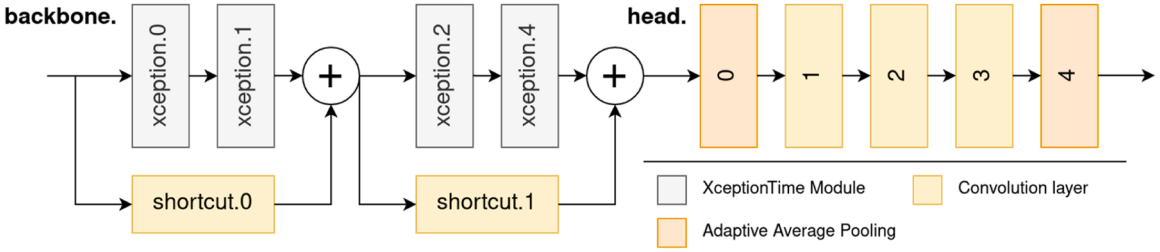

Fig. 2. Simplified view of the utilized model architecture based on XceptionTime [23].

representations using the target and control concepts to construct a separating hyperplane between them. The *concept activation vectors* (CAVs) v_C^l are then extracted as the normalized directions of the latent space pointing towards the target concepts. These directions indicate how the internal activations change as the concept becomes more or less present. Using the directional derivative for the target class k, the sensitivity towards changes in the direction of C for given input data x is quantified by

$$S_{C,k,l}(x) = \nabla h_{l,k}(f_l(x)) \cdot v_C^l,$$

where $f_l(x)$ defines the activations for the input x at layer l and $h_{l,k}$ denotes the function that maps those activations through the remaining network to predict class k.

To generate the random counterpart for our use-case, we randomly selected samples from the TUEG set. To get the most independent representation, we only selected subjects that are not present in the TUAB dataset and included only one record per subject. Similar to the test set, we cut down each record to the first minute. An alternative to comparing concepts of interest to random control concepts, is to compare different target concepts to one another. To do so, the original paper proposes a one vs. rest approach that is referred to as *relative TCAV*. This way, it is argued that particularly concepts that are closely related can be used for fine-grained comparisons. This approach may be more generally applicable in the domain of biosignals due to a lack of comprehensive databases of diverse signal data and the difficulty of generating realistic, but random physiological signals.

The result of TCAV is the concept sensitivity score that indicates how much the concept represented by the CAV affects the model's output with respect to the target class. We stick to the original definition by Kim et al. [9] where the score denotes the fraction of target class inputs $x \in X_k$ whose activation vector was positively influenced by the concept, i.e.

$$TCAV_{C,k,l} = \frac{\left|\{x \in X_k : \ S_{C,k,l}(x) > 0\}\right|}{|X_k|}.$$

TCAV scores therefore range between 0 and 1 and are computed for individual network layers. The target class for concept sensitivity testing was set to the "abnormal" EEG class. In all experiments we set the size of each concept batch to 75 samples, from which one-third were used for validation. The CAVs for all experiments were extracted using the last four network layers of the trained model, referred to as the network head. The layers were numbered consecutively with higher numbers referring to later stages of the network and smaller latent spaces respectively. To best reduce random effects, we constructed concept sets for training CAVs by resampling randomly 100 times. The resulting TCAV score was averaged across the randomized repetitions. A statistical t-test was performed to determine whether scores for concepts of interest are consistently higher than those for their respective control concepts. Alpha was set to 0.05 and Bonferroni correction was used ($p < \alpha/m$ with $m = 2$).

2.4. TCAV experiments

We analyzed different approaches to define concepts. The experiments included the use of additional sources of labeled EEG data, categorization by metadata, extraction of signal frequencies and the generation of synthetic signal data.

2.4.1. Experiment 1 a,b: concepts defined through annotated clinically relevant EEG events

In version (a) of the experiment we made use of the TUEV dataset. The TUEV dataset contains EEGs where segments of the signal are annotated with one of six classes: (1) spike and sharp wave (SPSW), (2) generalized periodic epileptiform discharges (GPED), (3) periodic lateralized epileptiform discharges (PLED), (4) eye movement (EYEM), (5) artifact (ARTF) and (6) background (BCKG). We refer to the classes by their aforementioned abbreviations to comply with the notation of Harati et al. [18]. The events SPSW, GPED, and PLED are considered to be relevant for the assessment of abnormality, while the events EYEM, ARTF and BCKG represent signal activity that can generally occur during signal recording, independent of the presence of neurological diseases [26]. We define six concepts analogously to the labels of the TUEV corpus. To generate representative concept samples from the TUEV dataset, we iterated through the provided start and end timestamps, as well as marked channels, for each event type. We defined blocks of continuous event activity by considering event annotations throughout channels without disruptions. Given these markings, we then extracted windows containing such labeled periods of EEG activity (see Fig. 3). With this procedure we generated concept samples that guaranteed containment of event activity.

For version (b) of the experiment, our own data from the UKA was annotated with the same labels as used in TUEV, except for GPED and PLED, since these types are rarely recorded in the epilepsy monitoring unit. This experiment allowed us to have more control of the signal and labeling quality. Further, having a different data source allows us to test the algorithm for generalizability. The concept samples were generated analogously as described above.

2.4.2. Experiment 2 a,b: frequency bands

To evaluate how more abstract concepts such as frequency band power can be used with TCAV, we prepared two experiment variants. Five clinically relevant frequency bands (Delta: 1–3 Hz, Theta: 4–7 Hz, Alpha: 8–13 Hz, Beta: 14–30 Hz, Gamma: >30 Hz) were selected as target concepts. In the first variant (a) we applied band-pass-filtering to real EEG records from the TUEG database to only contain the above mentioned frequency ranges. In the alternative version (b), instead of using real EEG records, we generated simulated signal data. Given e.g. the concept of the alpha band, each sample was constructed by randomly choosing a fixed frequency between 8 Hz and 13 Hz and generating a sine wave for each channel. To add variation to the data, the starting phase was randomized per sample. Further, data augmentation, including amplitude scaling, magnitude warping, time warping and the addition of noise was applied channel-wise.

2.4.3. Experiment 3: concepts from metadata

We used samples from the TUEG dataset to construct concept samples from demographic metadata. Using age and sex information encoded in the header of the record files, we categorized a subset of records into males and females, while another subset was categorized into elderly ($\geq$ 60 years) and young ($\leq$ 30 years) subjects.

2.5. Code and implementation details

The model parameters and checkpoints, as well as the code for all experiments are available online [19]. The online repository gives more details on how to run each script.

3. Results

3.1. Implementation of the deep learning model

The proposed implementation of the deep learning model achieved a classification accuracy of 85.1 % on the dedicated test set. This result comes close to the state-of-the-art models as reported in Roy et al. [15] and Kiessner et al. [27]. Table 1 compares the performance of the classification models from other reports.

3.2. Concepts representation

The concepts for explaining the network are represented by exemplary sets of signal data matching the models input requirements. Fig. 4 shows examples for the concept sets used in our Experiment 1b. While

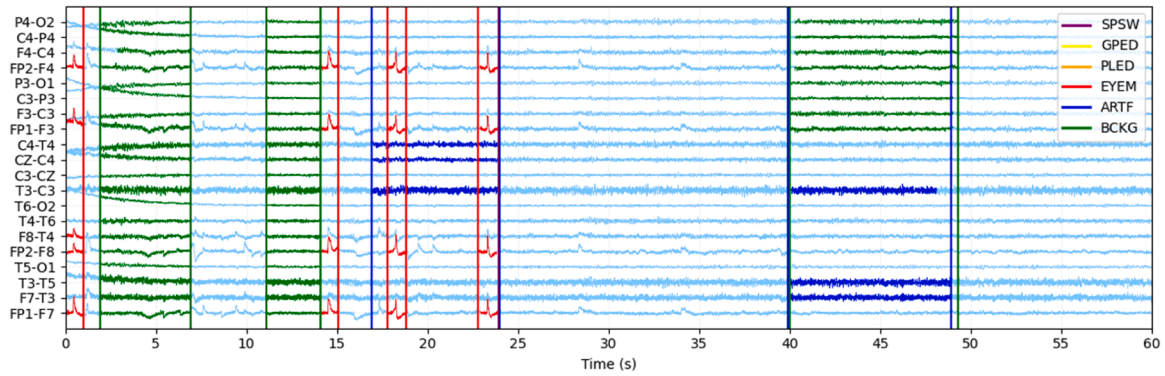

Fig. 3. Extraction of concepts from the TUEV corpus. The image shows an exemplary EEG with event annotations. Blocks of continuous event activity are defined by undisruptive marking throughout channels. During sampling, the concept is generated by randomly selecting an event block (indicated by vertical lines) and cropping the signal data accordingly. The colors mark the event classes: spike and sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifact (ARTF) and background (BCKG).

Table 1
Comparison of classification accuracy on the TUAB dataset. All scores in %, n.a.: not available. *Usage of longer and/or more time segments. **Usage of additional training data. [R]Publicly available re-implementation.

Authors	Year	Input / Methods	Accuracy	Sensitivity	Specificity
de Diego et al. [13]	2017	Wavelet features / MLP	78.8	n.a.	n.a.
Schirrmeister et al. [14]	2017	Raw Signal / BD-Deep4	85.4*	75.1*	94.1*
Brenner et al. [28]	2018	Wavelet features / MLP	79.8	72.4	86.0
Roy et al. [15]	2019	Raw Signal / ChronoNet	85.3, (86.6*), (81[R])	n.a.	n.a.
Alhussein et al. [16]	2019	Raw Signal / AlexNet + MLP	89.1**	80.2**	96.7**
Knispel et al. [8]	2022	Frequency features / CNN	81.4	n.a.	n.a.
Khan et al. [21]	2022	Raw Signal/ Hybrid Model (LSTM+CNN)	85.0	n.a.	n.a.
Kiessner et al. [27]	2023	Raw Signal / BD-TCN	86.5**	77.1**	94.4**
Proposed method	2024	Raw Signal / XceptionTime	85.1	85.7	84.7

the presented snippets mainly represent their respective event label, they can also overlap, as shown previously in Fig. 3.

3.3. TCAV results

Fig. 5 shows the TCAV scores for the experimental concept sets: the extracted events from the TUEV set and our own event annotations, the frequency band representatives and the concepts sampled with utilization of metadata. We only view TCAV scores that passed the statistical significance testing. CAVs that not passed the test were marked with an asterisk "*". In Experiment 1a, the concepts SPSW, GPED and PLED had significant TCAV scores across all tested network layers. The concepts EYEM, ARTF and BCKG, in contrast, had no significant influence in the earlier network layers, but yielded few high scores in the final layers of the neural network. We observed a similar trend for the data from Experiment 1b, where only SPSW yielded consistent high scores. Viewing the frequency band concepts from Experiment 2, we observed only significant TCAV scores for the theta and delta bands for both variants of concept generation. For Experiment 3, the sex-related concepts did not show significant influence across layers, while the concept of elderly patients' EEGs showed significant scores in the earlier network layers and the concept of young patients' EEGs had a peak in the last layer.

Fig. 6 shows a comparison of sensitivity scores with the relative TCAV approach. The relative TCAV scores for the events from the TUEV dataset in Experiment 1a were comparable to the regular TCAV

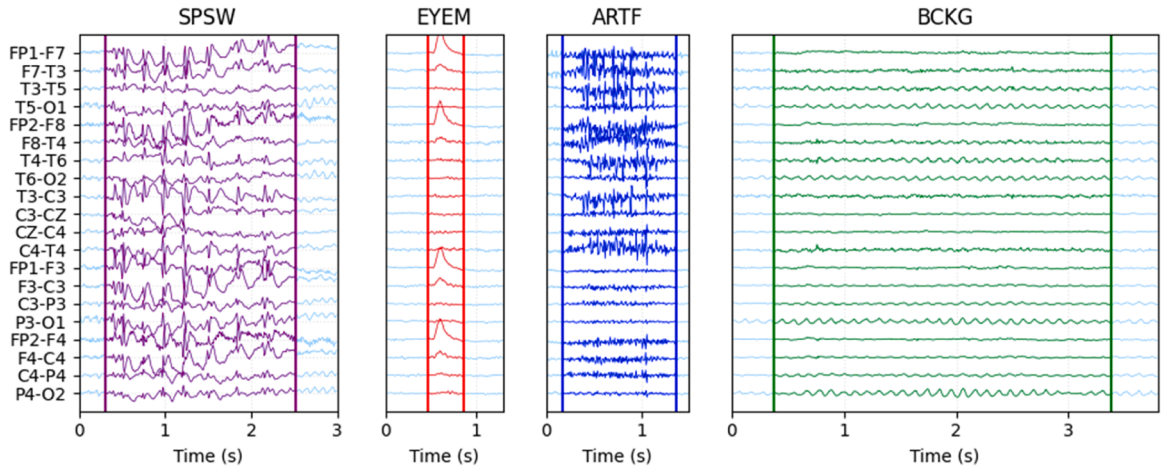

Fig. 4. Concept examples for experiment 1b. The colored parts mark the original event annotations. The images show examples for the following labels: spike and sharp wave (SPSW), eye movement (EYEM), artifact (ARTF) and background (BCKG).

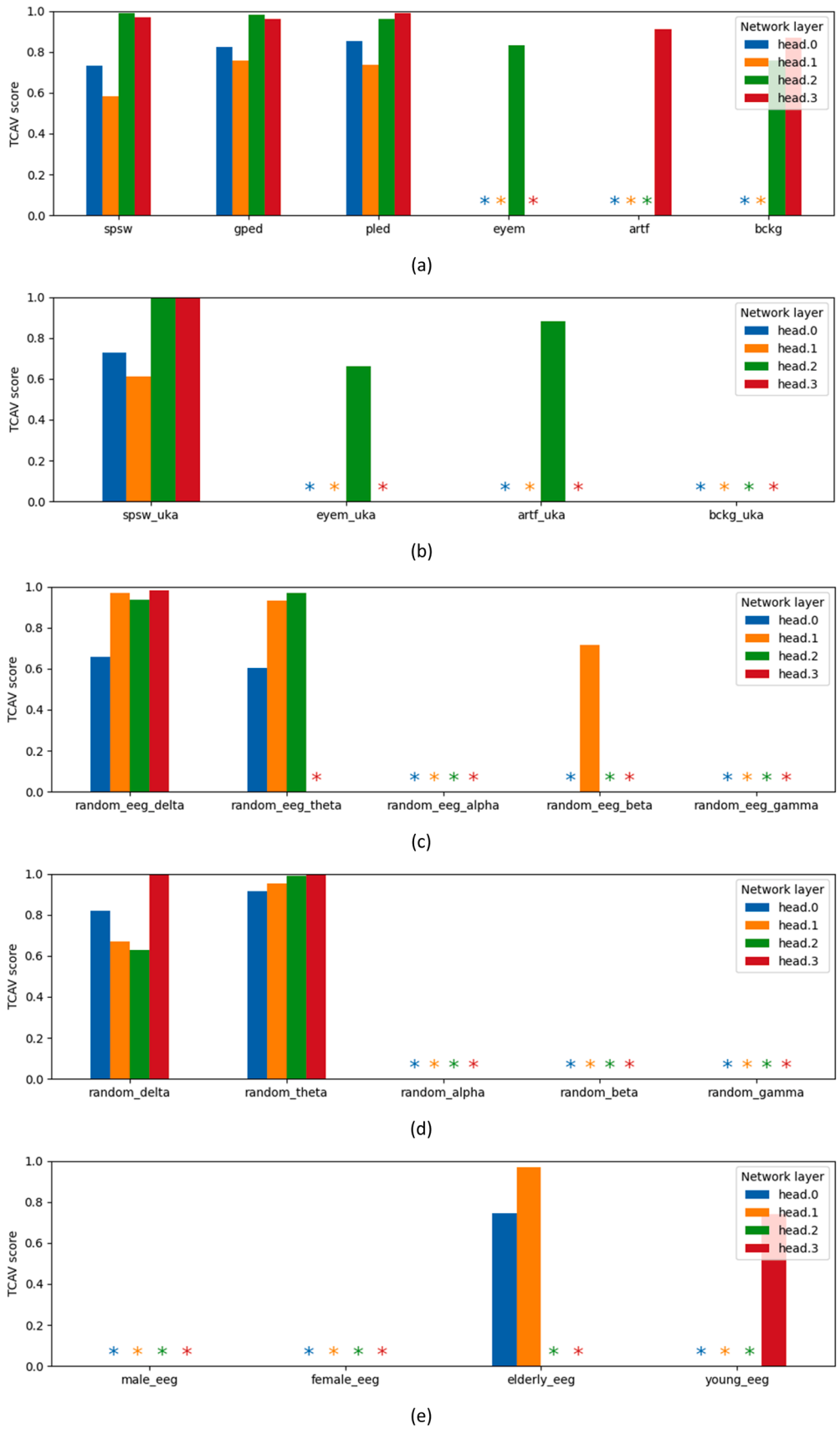

Fig. 5. TCAV scores given the TUAB dataset and the XceptionTime classification model. The top image (a) compares the event-based concepts derived from the TUEV dataset. The next row (b) compares analogously defined concepts from our own data. The middle images (c, d) compare the scores of the artificially generated signals representing relevant frequency bands. The bottom image (e) compares concepts generated using meta information. All CAVs were computed with random TUEG samples as counterparts. The "*" marks CAVs omitted after statistical testing.

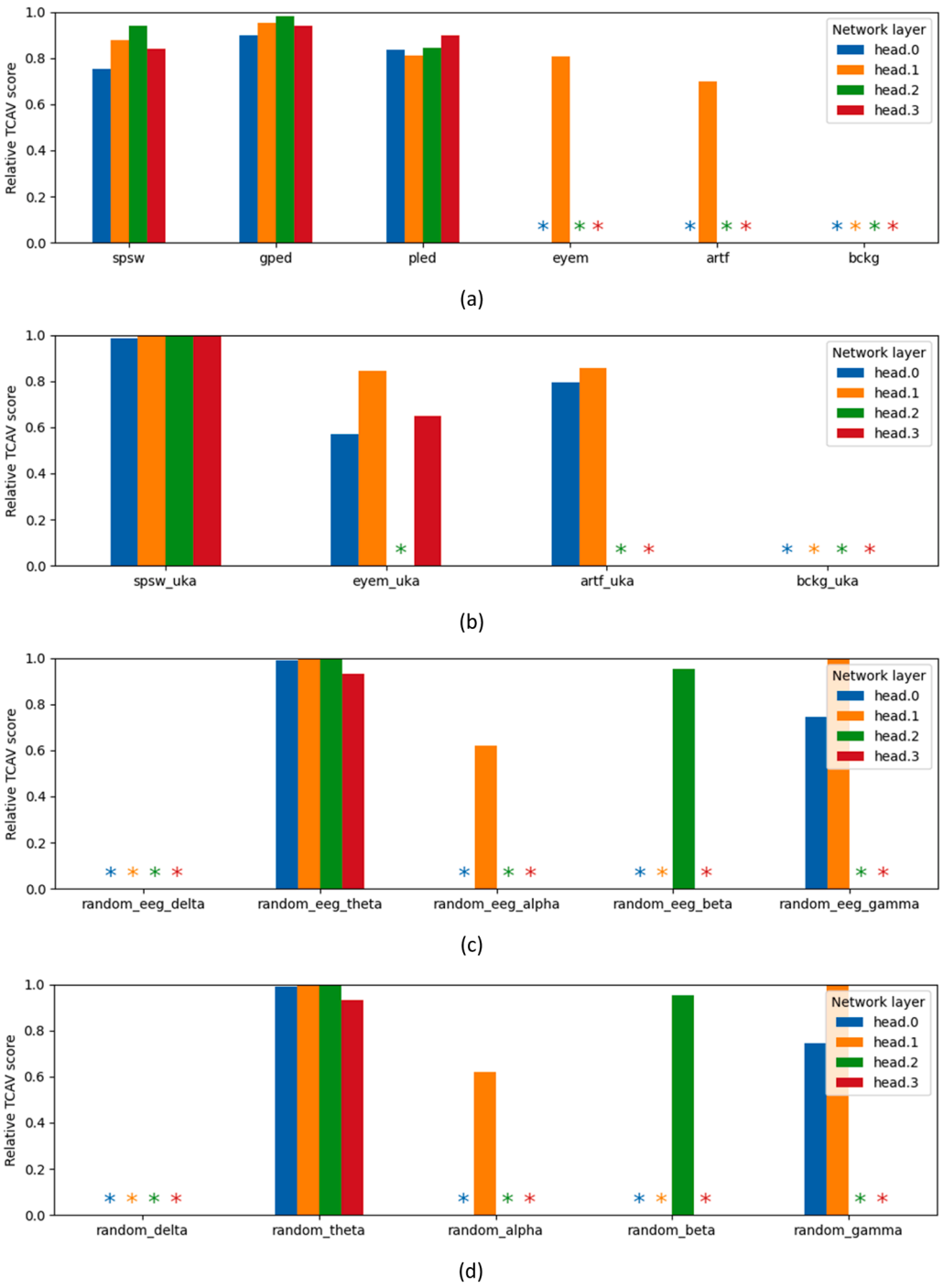

Fig. 6. Relative TCAV scores given the TUAB dataset and the XceptionTime classification model. The top image (a) compares the event-based concepts derived from the TUEV dataset. The next row (b) compares analogously defined concepts from our own data. The bottom images (c, d) compare the scores of the artificially generated signals representing relevant frequency bands. The "*" marks CAVs omitted after statistical testing.

approach. The concepts SPSW, GPED and PLED again were significant in all layers, for the remaining concepts only EYEM and ARTF had a significant score in one layer. For experiment 1b results were also comparable, however, EYEM had a relevant influence on three out of four tested network layers. Experiment 2 yielded different results when comparing the relative TCAV approach to the aforementioned results. While the theta band had significant scores across all network layers, the delta band scores were omitted after statistical testing. Further, for the alpha, beta and gamma frequency ranges high scores were computed for some of the tested layers. Experiment 3 was skipped due to low accuracy of the linear CAV classifier (near the random baseline).

In order to better understand occasionally occurring inconsistencies between the scores of individual layers, we plotted the performance of the internal linear classifiers used to define the CAVs in Fig. 7. While TCAV scores in the later layers of the network may represent more direct influences on the prediction than lower layers since the latent representations are closer to the final output, we observed lower accuracy of the CAV classifiers in the final layers. Lower accuracy results in higher noise and variance in the CAVs and thus a higher instability of the TCAV scores.

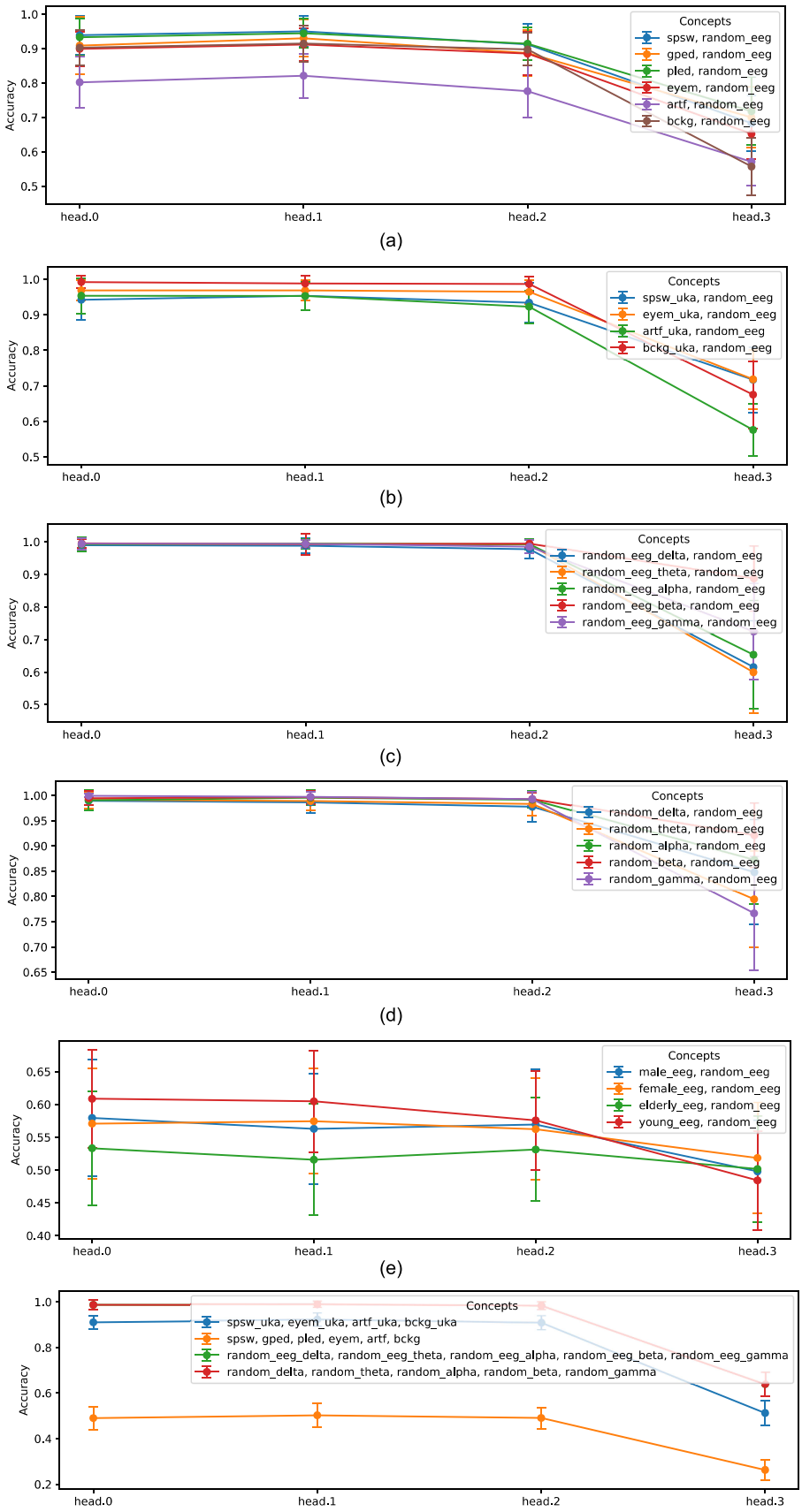

Fig. 7. Classifier accuracies for CAVs at different layers. Each curve shows the mean accuracy (std) of the linear classifiers used to define the CAVs.

4. Discussion

4.1. Classification performance

Given the publicly available TUAB dataset, we have trained and evaluated a deep learning model for the distinction of abnormal and normal EEGs based on 60 s records. Utilizing the XceptionTime architecture we achieved a classification accuracy of 85.1 %, which is en par with the reported performance of the ChronoNet model by Roy et al. that achieved 85.3 % using the first minute of each training sample and 86.6 % with the enlarged training set [15]. While the source code of ChronoNet is not publicly available, Khan et al. reported 81 % accuracy using a re-implementation [21]. Similarly, the other comparable methods summarized only slightly outperformed the proposed model when using additional data. Gemein et al. generally argued that accuracy scores for the current binary EEG classification could saturate near 90 % due to imperfect inter-rater agreement of the clinical labels [27,29].

4.2. Interpretation of the TCAV scores

TCAV gives an overview of the general sensitivity of the classification model towards concepts that can be viewed by a set of examples. In our analysis we conducted the TCAV analysis via different experiments using synthetically generated data and additional sources of labeled EEG data (Table A1). The results overall imply that the presented deep learning model generally pays attention to epileptic discharges as well as theta band activity when classifying abnormality. The control concepts, in contrast, mostly did not have significant effect in the TCAV analysis, which indicates that these types of signal activity were not erroneously caught up as an influencing feature by the model.

4.2.1. Experiment 1

In the first test, the event types EYEM, ARTF and BCKG from the TUEV dataset were not significantly relevant to the model's classification of abnormality in most layers. This result is in line with expectations, as these concepts represent general patterns that may occur during EEG recording but are unrelated to abnormality. Still, we observed significant scores in some of the later network layers. While these scores can potentially point out spurious correlations, they may be influenced by other biases. These result from the fact that the concept snippets can include overlapping event activity. Further, we observed lowest accuracy in the internal linear separation of the aforementioned concepts from their random counterpart that increases noisiness in the CAVs. In contrast, SPSW, GPED and PLED, which are typically found in abnormal epileptic activity, had significant concept sensitivity scores across all tested network layers, confirming the medical expectation.

Using our own labeled data as concepts produced comparable results, the SPSW concept was significant across layers, while the control concepts had overall low scores. This result indicates stability and generalizability of the approach, since machine learning can tend to be sensitive to sources of data, e.g. different recording devices [30].

4.2.2. Experiment 2

In the second test, both the band-pass filtered EEGs and the synthetically generated concept samples were chosen to represent activity of clinically relevant frequency bands. The concept of randomly generated theta waves had the overall highest TCAV scores and exhibited significant scores throughout the tested network layers. Theta activity is generally related to drowsiness and may be an abnormal sign in awake patients. However, it can also be linked to medications [31]. In previous work [8], we analyzed the impact of EEG frequency band powers utilizing feature extraction via Welch's method and feature attribution methods. We found that a strong signal in the theta frequency ranges was important to the model's decision when compared to other frequency ranges, which is in line with our current results. While there is no need for manual feature definition when applying TCAV, the results indicate that the method still can be tested for concepts defined equivalently to manually defined features.

4.2.3. Experiment 3

From the TCAV scores of the sex concepts, it can be extracted that neither females nor males presented a relevant concept for abnormality. In the comparison between elderly and young records, however, an influence was observable. Viewing the sex and age distribution of the TUAB dataset (see [13]), we observed that the elderly group is slightly more represented in the train set in comparison to younger individuals, while males and females are equally distributed (Fig. A1). That is, the scores may indicate this imbalance. However, it has to be noted that the accuracy of the CAV classifier was very low and thus the concepts may not have been stably encoded (see Fig. 7(d)). For future experiments we therefore recommend to consider the linear separability of the concepts in the latent space of the network by e.g. applying permutation tests similar to Schrouff et al. [32].

4.3. Limitations

4.3.1. Biosignals-specific limitation

Since the interpretation when using TCAV relies on concept generation with exemplary samples, both the availability of such data and the domain knowledge needed to decide for how concepts should be characterized may be seen as a limitation. Generally, this process can be subjective and may not capture all relevant concepts. If important concepts are not included, however, the explanations may miss crucial factors contributing to the network's decisions. Further, we observed that it is useful to compare multiple concepts to one another, since a single TCAV score of one concept may not be well interpretable without context. While we have shown one potential way of creating concepts via synthetic data in experiment 2a, it is generally non-trivial to simulate representative physiological data with certain characteristics. Here, both domain-knowledge, as well as technical knowledge about signal processing may be needed. Another alternative for defining concept labels is to use metadata, as presented in Experiment 3. In practice, however, metadata is often available in the form of written clinical reports that need additional processing with e.g. natural-language processing techniques to extract concept labels. When including different data sources for the definition of concepts, also demographic shifts and external influences, such as differences in measuring devices, have to be taken into account, since machine learning models may be sensitive to subtle changes [30].

4.3.2. General TCAV-related limitations

Adebayo et al. found that post-hoc explanations applied to deep neural networks generally may not be well suited for finding spurious correlations to signals patterns or artifacts that are unknown to the user [33]. Schrouff et al. combined TCAV with Integrated Gradients to enable local attribution [32]. Still, more fine-grained explanations that highlight subtle dependencies, such as specific interactions or relationships between concepts and their localization in the input, remain challenging [34]. While the noisiness of model gradients is handled by aggregating results over multiple samples and repeated randomized runs, CAVs can also be applied for local analysis by computing concept sensitivities for a single sample. However, analyzing single samples may not generate stable results, so it is recommended to apply TCAV only for the analysis of general model influences given a batch of input examples [9]. The presented method for concept-based explanation is limited to neural networks. Further, it is recommended to use TCAV on deep neural networks, since it is generally reported that concepts in the latent space of deeper layers are more separable as compared to concepts in shallower networks [9,35]. TCAV relies on linear separability that could be a limitation when comparing complex concept patterns that may be closely related. Crabbe et al. addressed this problem by comparing clustered regions in the latent space of the network instead of linear

CAVs [36]. As displayed in Fig. 7, the CAV classifier accuracy is sub-optimal in some cases (e.g. male/female EEG, elderly/young EEG), increasing noise in the results. Schrouff et al. [32] proposed not to consider such cases in the analysis and to test for whether a concept has been "significantly" encoded in each layer by using permutation tests on the CAV classifiers. Alternatively, other approaches to separating the concept representations could be evaluated, such as done by Crabbe et al. [36]. In general, it is important to note that TCAV does not provide hard guarantees on model behavior, nor does the method allow for *full* explanations of model behavior. The utility of TCAV fully depends on the concept data and labeling provided.

4.4. Clinical applicability and translational aspects

To address our research question, we used three concept generation strategies in our experiments: labeled signal excerpts, synthetically generated data, and meta-information. Overall, we observed consistent results for testing with similar concepts that have been generated with different approaches. More specifically, frequency band filtered EEGs and corresponding synthetically generated frequencies led to a similar distribution of TCAV scores. The same applies to concepts extracted from the TUAB event labels and our own hospital data that have been labeled accordingly. We started our work independently from the research of Madsen et al. [12], but recognized overlap in the experiments as they also extracted event-based concepts from the TUEV dataset. Madsen et al. extracted 60-second windows from the TUEV records to generate representative samples for the event labels. However, this approach may bias the results as events are often short in time and long time windows may cover multiple different labels (see Fig 3.). In contrast, our pipeline allows for variable-length concept samples that only consider the time period of actual event activity. Madsen et al. further included concepts for the representation of certain frequency band activity in specific brain regions. Their experiment yielded expected concept sensitivity in the classification of left fist movement from EEG, confirming lateralization in cortical activities. The results underline the potential of the usage of abstract concepts such as frequency band activity, which is in line with our observations.

Further exploration work is necessary to build the best practices of using the concept-based interpretability in medical biosignals. The most probable direction to proceed is to explore large existing datasets with concept-supporting labeling to better understand similarities of a model's decision process to human decision-making, and then continue with improving the concept definition towards specific clinical needs. With the observations made, we consider different implications for scientific research and for setting up models for integration into clinical practice. With TCAV the model can not only be checked for concepts that are known to be important for a certain distinction, but further be used for systematic screening of biases of the model behavior. In our example, we have shown that our trained model was sensitive to known EEG pathologies that include transient spikes and sharp waves, as well as periodic epileptiform discharges. Madsen et al. found concept sensitivity for the same labels for the classification of seizures [12]. But moreover, unexpected or clinically/physiologically irrelevant sensitivity towards known patterns that are generally not associated with the classification task, may give indications for rechecking data balance, e.g. in terms of age or sex distributions, and quality. For example, Wu et al. defined a process to filter out spurious correlations using TCAV [37]. One advantage of TCAV is that it is not necessary to be able to precisely define differences between concepts that are tested against each other, but sufficient to know which label they belong to. In this way, e.g. the concept of sex can be analyzed by passing male and female records into TCAV, respectively. These capabilities could be used to evaluate pre-trained models regarding their sensitivity towards certain domains or unwanted behavior. As models are increasingly being trained on large but intransparent datasets, a reliable and actionable assessment of model behavior becomes more important. TCAV gives insights into the linear separability of the concepts given their representation in the latent space of the models layers (Table A2). This characteristic may be used to screen existing models trained on large databases that could potentially be used for transfer learning utilizing their conserved knowledge. In this way e.g. a network trained for the more general task of abnormality detection may give a solid foundation for detecting more specific abnormal patterns, such as epileptiform discharges. The general question of whether deep learning is sensitive to the same concepts as humans, however, cannot be answered straightforward yet. While there may exist a real impact factor on the classification that is not considered in human decision-making, there is always the risk that a pattern has been learned due to data or correlation biases as described above. Since machine learning models always have a dependency on the data they were trained on, finding out true causal influences remains challenging.

5. Conclusions

Our study presents a pipeline for employing concept-based explanations for deep-learning classification of complex biosignals. Using the publicly available dataset of labeled EEGs with a dedicated test set, we trained an open source reproducible classifier for distinguishing between abnormal and normal EEGs. By employing the XceptionTime architecture, we achieved a competitive classification accuracy of 85.1 %. Importantly, the classifier uses raw data as opposed to human-engineered features. Further, we proposed several alternatives to construct clinically meaningful concepts based on other labeled data, synthetic data and the available metadata. The TCAV algorithm applied on those concepts demonstrated the feasibility of analyzing a deep learning model for biosignals regarding its sensitivity to clinically relevant concepts. Furthermore, the results were consistent when comparing concepts from the public dataset and our own similarly labeled hospital data, indicating that the process is generalizable.

In our experimental setting, TCAV demonstrated concept sensitivity scores that were high in case of abnormal signal patterns, e.g. epileptiform discharges, and mostly negligible for concepts unrelated to abnormality, such as sex. These results are generally in line with the medical expectation. However, in several measurements unexpectedly high TCAV scores were observed, possibly indicating patterns and potential biases that the classifier caught up in the training process or being the result of the discussed above technical limitations. While further work is necessary to uncover the full potential of the concept-based interpretability methods in the context of medical biosignals, the presented project lays the bases through building the analogies to the relatively well-established research in the image analysis domain. Our results demonstrate a potential direction towards adapting the concept-based explainability approach from image analysis to the domain of biosignals analysis.

CRediT authorship contribution statement

Alexander Brenner: Writing – original draft, Methodology, Conceptualization, Software, Formal analysis, Visualization. **Felix Knispel:** Writing – review & editing, Methodology, Software. **Florian P. Fischer:** Data curation, Writing – review & editing, Validation. **Peter Rossmanith:** Writing – review & editing, Supervision. **Yvonne Weber:** Writing – review & editing, Supervision, Validation. **Henner Koch:** Writing – review & editing, Validation. **Rainer Röhrig:** Writing – review & editing, Supervision. **Julian Varghese:** Writing – review & editing, Supervision, Resources. **Ekaterina Kutafina:** Writing – original draft, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the ZM-Fond (ZM200013). We acknowledge support from the Open Access Publication Fund of the University of Münster.

Appendix

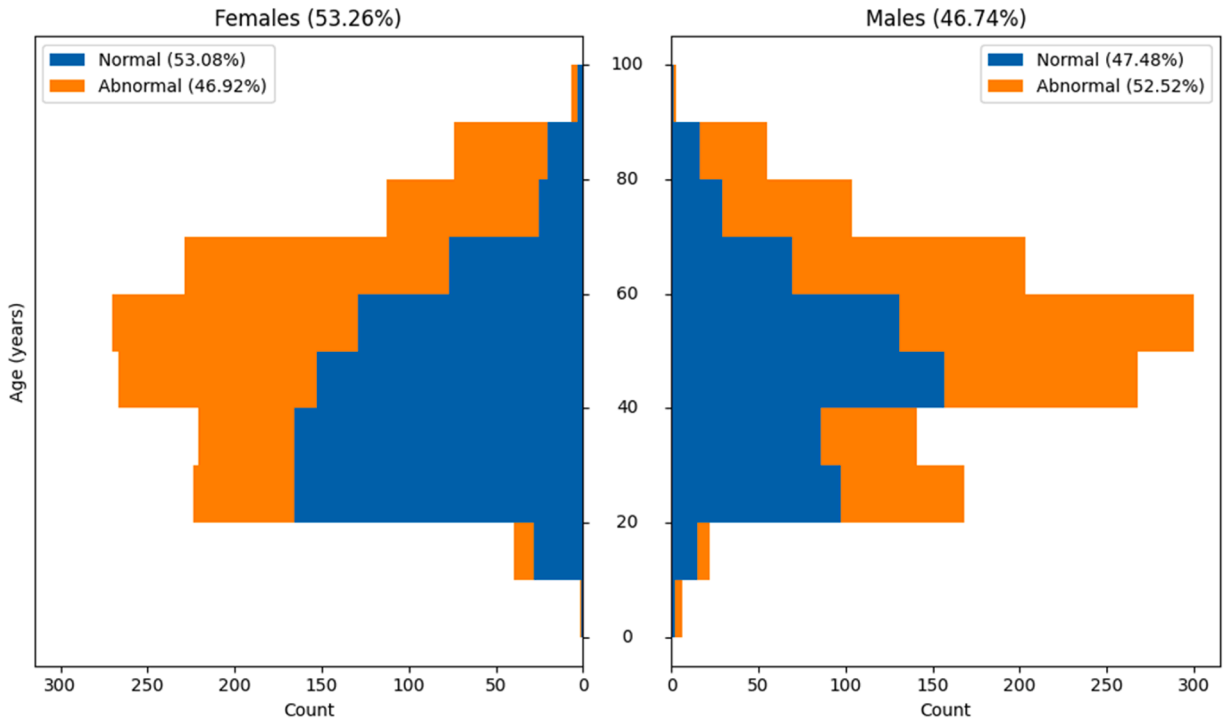

Fig. A1. Histogram of age distribution in the TUAB train set (v3.0.0). The samples are categorized by sex.

Table A1
Overview of the data sources taken for the generation of representative concept samples. While the number of available samples may differ between concept groups, random sub-sampling of 75 samples was used in each experimental run.

Data source	Concept class	Number of samples
TUEG	Random	184
	Male	100
	Female	100
	Old	100
	Young	79
TUEV	SPSW	80
	GPED	505
	PLED	718
	EYEM	269
	ARTF	357
	BCKG	664
UKA	SPSW	80
	EYEM	106
	ARTF	134
	BCKG	123

Table A2
Sizes of the latent space representations of the final network layers from the used XceptionTime model.

Network layer	head.0	head.1	head.2	head.3
Output shape	[384, 50]	[192, 50]	[96, 50]	[32, 2]

References

[1] U. Pagallo, S. O'Sullivan, N. Nevejans, A. Holzinger, M. Friebe, F. Jeanquartier, C. Jean-Quartier, A. Miernik, The underuse of AI in the health sector: opportunity costs, success stories, risks and recommendations, Health Technol. 14 (2024) 1–14, https://doi.org/10.1007/s12553-023-00806-7.

[2] S. Reddy, Explainability and artificial intelligence in medicine, Lancet Digit. Health 4 (2022) e214–e215, https://doi.org/10.1016/S2589-7500(22)00029-2.

[3] S. Tonekaboni, S. Joshi, M.D. McCradden, A. Goldenberg, What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use, in: Proc. 4th Mach. Learn. Healthc. Conf., PMLR, 2019: pp. 359–380.

[4] F. Funer, W. Liedtke, S. Tinnemeyer, A.D. Klausen, D. Schneider, H.U. Zacharias, M. Langanke, S. Salloch, Responsibility and decision-making authority in using clinical decision support systems: an empirical-ethical exploration of German prospective professionals' preferences and concerns, J. Med. Ethics (2023), https://doi.org/10.1136/jme-2022-108814.

[5] Stellungnahme der Zentralen Kommission zur Wahrung ethischer Grundsätze in der Medizin und ihren Grenzgebieten (Zentrale Ethikkommission) bei der Bundesärztekammer "Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz, Dtsch. Ärztebl. Online (2021), https://doi.org/10.3238/arztebl.zeko_sn_cdss_2021.

[6] K. Stöger, D. Schneeberger, A. Holzinger, Commun ACM 64. Medical artificial intelligence: the European legal perspective, 2021, pp. 34–36, https://doi.org/10.1145/3458652.

[7] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity Checks for Saliency Maps. Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 2018, in: https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html. accessed January 16, 2024.

[8] F. Knispel, A. Brenner, R. Röhrig, Y. Weber, J. Varghese, E. Kutafina, Consistency of feature importance algorithms for interpretable EEG abnormality detection, Stud. Health Technol. Inform. 296 (2022) 33–40, https://doi.org/10.3233/SHTI220801.

[9] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV), in: Proc. 35th Int. Conf. Mach. Learn., PMLR, 2018, pp. 2668–2677, in: https://proceedings.mlr.press/v80/kim18d.html. accessed April 16, 2023.

[10] A. Janik, J. Dodd, G. Ifrim, K. Sankaran, K. Curran, Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset, Med. Imaging 2021 Image Process. (2021) 111, https://doi.org/10.1117/12.2582227.

[11] D. Mincu, E. Loreaux, S. Hou, S. Baur, I. Protsyuk, M. Seneviratne, A. Mottram, N. Tomasev, A. Karthikesalingam, J. Schrouff, Concept-based model explanations for electronic health records, in: Proc. Conf. Health Inference Learn, Association for Computing Machinery, New York, NY, USA, 2021, pp. 36–46, https://doi.org/10.1145/3450439.3451858.

[12] A.G. Madsen, W.T. Lehn-Schiøler, Á. Jónsdóttir, B. Arnardóttir, L.K. Hansen, Concept-Based Explainability for an EEG Transformer Model. 2023 IEEE 33rd Int. Workshop Mach. Learn. Signal Process, MLSP, 2023, pp. 1–6, https://doi.org/10.1109/MLSP55844.2023.10285992.

[13] L. de Diego, S. Isabel, Automated interpretation of abnormal adult electroencephalograms, (2017). https://scholarshare.temple.edu/handle/20.500.12613/1767 (accessed May 14, 2023).

[14] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp 38 (2017) 5391–5420, https://doi.org/10.1002/hbm.23730.

[15] S. Roy, I. Kiral-Kornek, S. Harrer, ChronoNet: a Deep Recurrent Neural Network for Abnormal EEG Identification, Eds., in: D. Riaño, S. Wilk, A. ten Teije (Eds.), Artif. Intell. Med., Springer International Publishing, Cham, 2019, pp. 47–56, https://doi.org/10.1007/978-3-030-21642-9_8.

[16] M. Alhussein, G. Muhammad, M.S. Hossain, EEG pathology detection based on deep learning, IEEE Access 7 (2019) 27781–27788, https://doi.org/10.1109/ACCESS.2019.2901672.

[17] I. Obeid, J. Picone, The temple university hospital EEG data corpus, Front. Neurosci. 10 (2016). https://www.frontiersin.org/articles/10.3389/fnins.2016.00196. accessed August 14, 2023.

[18] A. Harati, M. Golmohammadi, S. Lopez, I. Obeid, J. Picone, Improved EEG event classification using differential energy, IEEE Signal Process. Med. Biol. Symp. SPMB IEEE Signal Process. Med. Biol. Symp. 2015 (2015), https://doi.org/10.1109/SPMB.2015.7405421.

[19] A. Brenner, Concept-based AI interpretability in physiological time-series data: code repository, GitHub (n.d.) (2024). https://github.com/alex-bre/tcav-in-eeg. accessed March 6.

[20] S. Ferrel, V. Mathew, M. Refford, V. Tchiong, T. Ahsan, I. Obeid, J. Picone, The temple university hospital EEG corpus: electrode location and channel labels, (2022). https://isip.piconepress.com/publications/reports/2020/tuh_eeg/electrodes/.

[21] H.A. Khan, R. Ul Ain, A.M. Kamboh, H.T. Butt, S. Shafait, W. Alamgir, D. Stricker, F. Shafait, The NMT scalp EEG dataset: an open-source annotated dataset of healthy and pathological EEG recordings for predictive modeling, Front. Neurosci. 15 (2022). https://www.frontiersin.org/articles/10.3389/fnins.2021.755817. accessed May 15, 2023.

[22] I. Oguiza, tsai - A state-of-the-art deep learning library for time series and sequential data, (2022). https://github.com/timeseriesAI/tsai.

[23] E. Rahimian, S. Zabihi, S.F. Atashzar, A. Asif, A. Mohammadi, XceptionTime: independent Time-Window xceptiontime architecture for hand gesture classification. ICASSP 2020-2020 IEEE Int. Conf. Acoust. Speech Signal Process, ICASSP, IEEE, 2020, pp. 1304–1308.

[24] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: a unified and generic model interpretability library for PyTorch, (2020). https://doi.org/10.48550/arXiv.2009.07896.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conf. Comput. Vis. Pattern Recognit, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[26] W.O. Tatum IV, Handbook of EEG Interpretation, Springer Publishing Company, 2021.

[27] A.-K. Kiessner, R.T. Schirrmeister, L.A.W. Gemein, J. Boedecker, T. Ball, An extended clinical EEG dataset with 15,300 automatically labelled recordings for pathology decoding, NeuroImage Clin. 39 (2023) 103482, https://doi.org/10.1016/j.nicl.2023.103482.

[28] A. Brenner, E. Kutafina, S.M. Jonas, Automatic recognition of epileptiform EEG abnormalities, Build. Cont. Knowl. Oceans Data Future Co-Creat. EHealth (2018) 171–175, https://doi.org/10.3233/978-1-61499-852-5-171.

[29] L.A.W. Gemein, R.T. Schirrmeister, P. Chrabąszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, T. Ball, Machine-learning-based diagnostics of EEG pathology, Neuroimage 220 (2020) 117021, https://doi.org/10.1016/j.neuroimage.2020.117021.

[30] L. Plagwitz, T. Vogelsang, F. Doldi, L. Bickmann, M. Fujarski, L. Eckardt, J. Varghese, The Necessity of Multiple Data Sources for ECG-Based Machine Learning Models, Stud, Health Technol. Inform. 302 (2023) 33–37, https://doi.org/10.3233/SHTI230059.

[31] A. Maxion, A.J. Gaebler, D. Albiez, K. Mathiak, J. Zweerings, E. Kutafina, EEG Spectral Changes Linked to Psychiatric Medications: Computational Pipeline For Data Mining and Analysis, in: Chall. Trust. AI Added-Value Health, IOS Press, 2022, pp. 957–958, https://doi.org/10.3233/SHTI220639.

[32] J. Schrouff, S. Baur, S. Hou, D. Mincu, E. Loreaux, R. Blanes, J. Wexler, A. Karthikesalingam, B. Kim, Best of both worlds: local and global explanations with human-understandable concepts, (2022). https://doi.org/10.48550/arXiv.2106.08641.

[33] J. Adebayo, M. Muelly, H. Abelson, B. Kim, Post hoc explanations may be ineffective for detecting unknown spurious correlation, (2022). https://doi.org/10.48550/arXiv.2212.04629.

[34] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From "where" to "what": towards human-understandable explanations through concept relevance propagation, (2022). https://doi.org/10.48550/arXiv.2206.03208.

[35] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, (2018). https://doi.org/10.48550/arXiv.1610.01644.

[36] J. Crabbé, M. van der Schaar, Concept activation regions: a generalized framework for concept-based explanations, Adv. Neural Inf. Process. Syst. 35 (2022) 2590–2607.

[37] S. Wu, M. Yuksekgonul, L. Zhang, J. Zou, Discover and cure: concept-aware mitigation of spurious correlation, (2023). https://doi.org/10.48550/arXiv.2305.00650.