Original Paper

Automated Detection of Abnormal Pediatric Electroencephalograms Using Machine Learning: Retrospective Study

Leonardo Poggi[1,2,3], MSc; Nils Flaschel[1,2], PhD; Sven Steffen Becker[1,2], PhD; Felix Nensa[1,2], Prof Dr; Judith Kohnke[1,2], MSc; Sandra Greve[4,5], Dr med; Luisa Paul[4,5,6], Dr med; Christian Dohna-Schwake[4,5], Prof Dr Med; Ursula Felderhoff-Müser[4,5], Prof Dr Med; Adela Della Marina[4,5], Dr med; Christopher Johannes Stein[7,8], Prof Dr; René Hosch[1,2*], PhD; Nora Bruns[4,5*], PD Dr

[1]Institute of Diagnostic and Interventional Radiology and Neuroradiology, Essen University Hospital, Essen, North Rhine-Westphalia, Germany
[2]Institute for Artificial Intelligence in Medicine (IKIM), Essen University Hospital, Essen, North Rhine-Westphalia, Germany
[3]Department of Neurology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden-Wurttemberg, Germany
[4]Department of Pediatrics I, Essen University Hospital, Essen, North Rhine-Westphalia, Germany
[5]C-TNBS, Centre for Translational Neuro- and Behavioural Sciences,, Essen University Hospital, Essen, North Rhine-Westphalia, Germany
[6]Department of Pediatric Cardiology/Congenital Cardiology, University Hospital Heidelberg, Heidelberg, Baden-Wurttemberg, Germany
[7]Theoretische Physik and CENIDE, University of Duisburg-Essen, Duisburg, North Rhine-Westphalia, Germany
[8]Catalysis Research Center and Atomistic Modeling Center, Department of Chemistry, Technical University of Munich, Garching, Bavaria, Germany
[*]these authors contributed equally

Corresponding Author:
Leonardo Poggi, MSc
Department of Neurology
Medical Faculty Mannheim
Heidelberg University
Theodor-Kutzer-Ufer 1-3
Mannheim, Baden-Wurttemberg, 68167
Germany
Phone: 49 621383 5588
Email: leonardo.poggi@medma.uni-heidelberg.de

Abstract

Background: Electroencephalography (EEG) is widely used in medicine as a technique for the assessment of several brain-related pathologies. However, EEG interpretation requires specialized expertise and remains a labor-intensive task, particularly for prolonged recordings in intensive care, where timely expert review is often challenging.

Objective: The aim of this study is to address this problem by using AI methods to automate EEG interpretation and compare the performance of different modeling approaches.

Methods: A gradient boosting–based model, inspired by the work of Albaqami et al, was implemented and compared to deep learning (DL)–based models using convolutional neural networks to classify images generated from raw EEG signals. The models were trained and tested on pediatric EEGs acquired at the University Hospital in Essen.

Results: The DL model showed higher observed sensitivity (78.3%) and F_1-score (80%) compared to the gradient boosting model (sensitivity: 65.2%; F_1-score: 73.8%), indicating improved detection of abnormal EEGs. However, the gradient boosting approach achieved higher specificity (93% vs 89.6%), suggesting a stronger ability to recognize normal EEGs. Overall, both methods yielded promising results, with classification accuracies reaching up to 85.3%.

Conclusions: Both approaches show potential for automated EEG interpretation. Future improvements, including larger and more diverse datasets, refined DL architectures, and integration of explainable AI techniques, could further enhance performance.

(JMIR Med Inform 2026;14:e96601) doi: 10.2196/96601

XSL•FO
RenderX

KEYWORDS

EEG interpretation; electroencephalography; machine learning; deep learning; pediatric intensive care; AI

Introduction

The electrical activity of the human brain is impacted by a large number of cerebral and noncerebral diseases [1]. Electroencephalography (EEG) is widely used in medicine for the assessment of such pathologies. However, EEG interpretation requires specialized expertise and remains a labor-intensive task, particularly for prolonged recordings in intensive care, where timely expert review is often challenging [2].

In critical care, EEG recordings are frequently conducted for hours or even days and, at the same time, require close to real-time interpretation in order to perform neuroprotective or rescue measures in case of deterioration. In this context, time-compressed mathematical transformations of the raw EEG, referred to as quantitative EEG, facilitate the detection of changes but still require human interpretation [3]. Especially in pediatric critical care patients, EEG interpretation is further complicated by age-dependent physiological differences in electrocortical activity, underlying acute and chronic diseases [4-6]. Critical illness-related metabolic and hemodynamic imbalances and neuroactive medications such as sedation or antiepileptic drugs add up, causing dynamic changes in electrocortical activity [7]. Automated real-time abnormality detection could advance recognition of present or new-onset abnormalities and help to direct diagnostic and neuroprotective measures.

Machine learning (ML) and deep learning (DL) provide methods that are suitable to address these problems. In recent years, AI has become an important part of health care, offering solutions that help medical staff to make diagnoses faster and with higher accuracy [8]. In the context of the automated interpretation of EEGs, some progress has already been made. The TUH Abnormal EEG Corpus [9] is a dataset provided by Temple University that consists of 2993 EEGs acquired from adult patients. Each EEG was either labeled as normal (no signs of pathology) or as abnormal (pathological EEG) by expert neurologists. This dataset has been used by different research groups as a basis for building and testing novel AI pipelines dedicated to the EEG interpretation task [10-12]. Another example is represented by Standardized Computer-based Organized Reporting of EEG-Artificial Intelligence, an AI model capable of fully interpreting routine clinical EEGs [13]. This model was trained on a big multicenter dataset comprising a total of 30,493 fully annotated normal and abnormal clinical EEGs.

While the combination of AI-driven methodologies and EEG recordings acquired in pediatric intensive care units (PICUs) as input data has been explored in several studies [14-16], little work has been done on the AI-assisted interpretation of such recordings [17,18]. This study aims to address this research gap by automating and comparing the task of interpreting EEGs acquired in PICUs using multiple ML and DL methods. We propose an EEG classification pipeline that relies on a classical gradient boosting approach inspired by the work of Albaqami et al [11] as well as a DL approach. In the latter one, the EEGs are converted to images and classified using a modified ResNet-50 architecture [19].

Methods

Ethical Considerations

The EEG measurements used in this work were recorded in the Department for Pediatrics I (Klinik für Kinderheilkunde I) of the University Hospital in Essen. All analyses were approved by the local Ethics Committee of the Medical Faculty of the University of Duisburg-Essen (20-9444-BO, 21-10264-BO). Written informed consent was waived according to local legislation for the analysis of retrospective anonymized data. All methods were conducted in accordance with relevant guidelines and regulations, including the principles outlined in the Declaration of Helsinki.

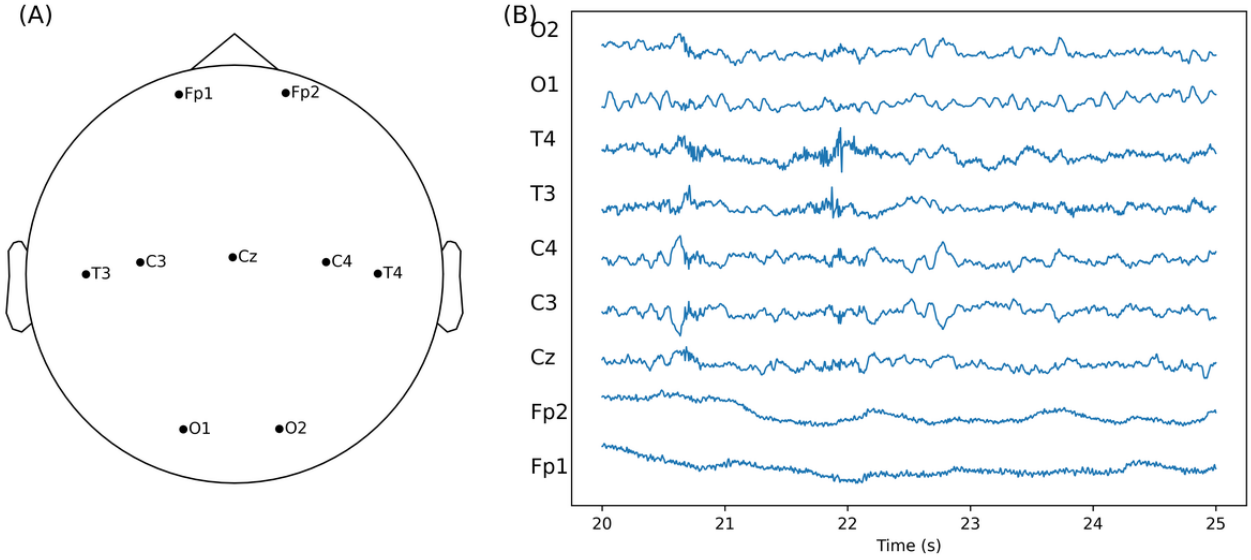
Dataset Collection

The dataset consisted of 916 multichannel routine EEG measurements acquired on patients between 0 and 18 years. The EEG recordings were interpreted as part of clinical routine by a board-certified pediatric neurologist with additional certification in epileptology and classified as normal or abnormal. No additional independent review or formal consensus adjudication was performed as part of the present retrospective study. In total, 571 normal measurements were recorded on healthy children who did not assume neuroactive medications and had no cerebral disease, damage, or developmental delay. Some of these data were already used by Greve et al [6] for a study on amplitude-integrated EEGs. The remaining 345 abnormal EEGs were acquired in the PICU, from critically ill children with manifest or highly suspected neurological disease [5].

All measurements were acquired with EEG-1200 Neurofax systems from the company Nihon Kohden, with a set sampling frequency for all EEGs of 200 Hz. A minimum of 11 and a maximum of 25 electrodes, placed according to the 10-20 system [20], were used for the recordings. The subset of electrodes common to all EEGs contains the following 9 electrodes: Fp1, Fp2, Cz, C3, C4, T3, T4, O1, and O2 (Figure 1A; image generated with MNE-Python [21]). An example of a 5-second portion of an EEG is displayed in Figure 1B.

XSL•FO
RenderX

Figure 1. (A) Placement of the 9 electrodes common to all measurements. Image generated with MNE-Python. (B) A 5-second portion of an electroencephalography. The 9 displayed signals are acquired by the electrodes common to all measurements in the dataset.

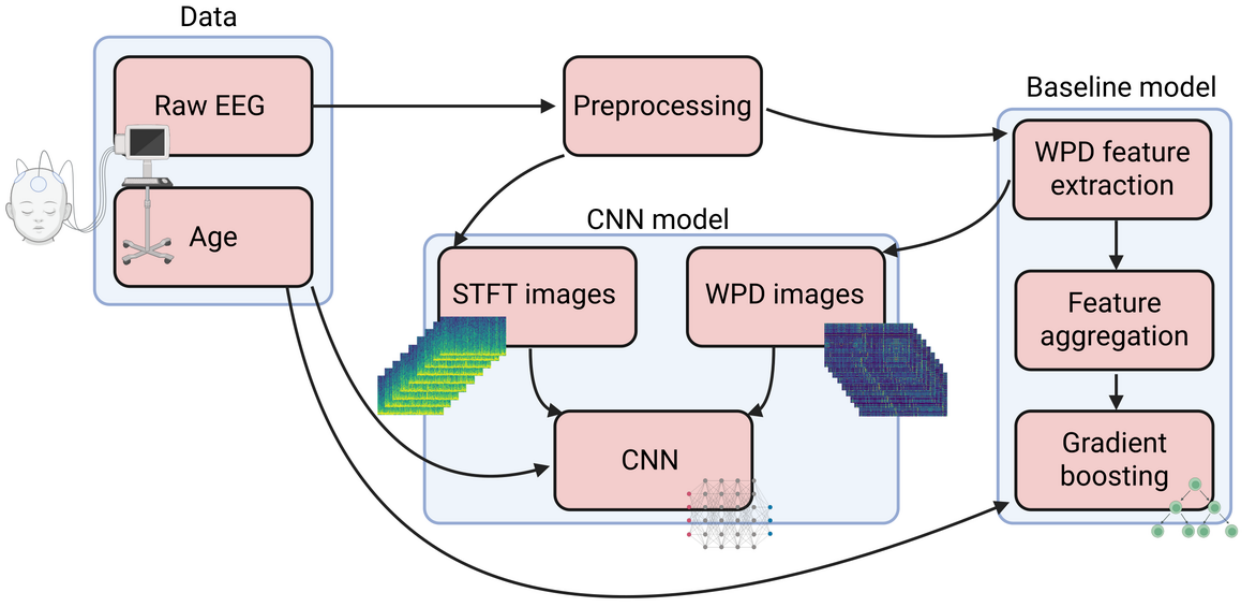


EEG Classification Pipeline

Overview

In Figure 2 (created in BioRender [22]), a diagram of the EEG classification pipeline implemented for this study is presented. The pipeline comprises 2 classification models. The baseline model is inspired by the work of Albaqami et al [11], a classification model based on gradient boosting, that was trained on the Temple University Hospital EEG Abnormal Corpus dataset [9]. The model consisted of 3 major steps: a feature extraction step that uses the wavelet packet decomposition (WPD) method, a feature aggregation step for dimensionality reduction, and a gradient boosting classification of the features.

Figure 2. Diagram of the implemented EEG classification pipeline. The raw data undergo several common preprocessing steps, after which it is prepared for the baseline and CNN model. For the baseline model, WPD features are extracted from the signals and are compressed by means of a feature aggregation process. Subsequently, the patient's age is concatenated to the dataset of aggregated features, and gradient boosting–based classification models are trained. For the DL approach, 2 kinds of images are generated from the signals. The first type of images is STFT spectrograms directly obtained from the preprocessed signals. The second type is constructed using the extracted WPD features, prior to the feature aggregation step. The patient's age is used as a separate input in the CNN architecture. Created in BioRender. CNN: convolutional neural network; EEG: electroencephalography; STFT: short-time Fourier transform; WPD: wavelet packet decomposition.



The second model implemented for this study aimed to classify images generated from the EEG signals, using a ResNet50-based convolutional neural network (CNN) architecture [19]. Prior to the feature aggregation process of the baseline model, images were generated from the extracted WPD features. Additionally, short-time Fourier transform (STFT) spectrograms were generated directly from the preprocessed data. Both WPD and STFT images were subsequently fed separately into the same CNN architecture.

XSL•FO
RenderX

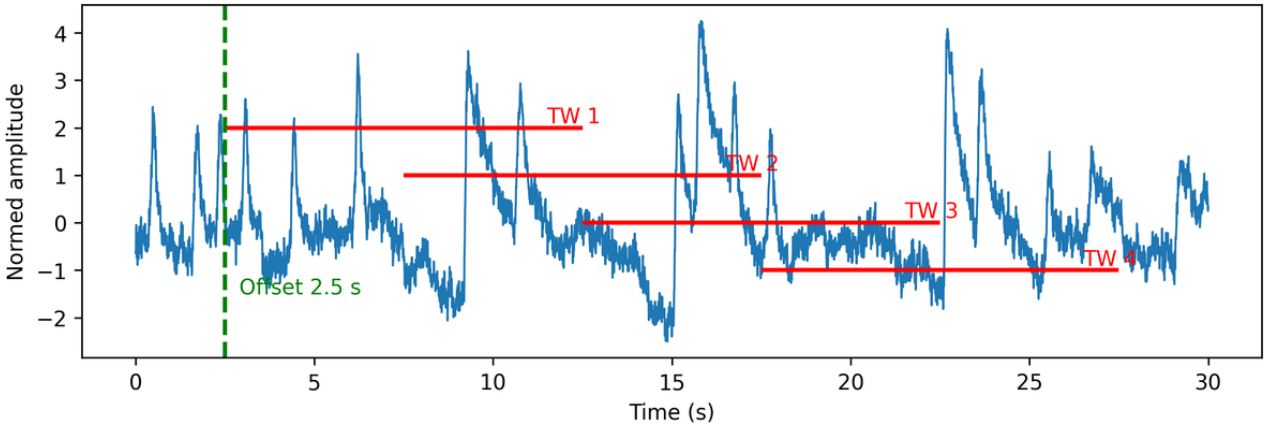
Preprocessing

All data underwent the same preprocessing steps. The raw EEG measurements were retrieved from the database, and all the sensitive metadata (name, birthday, and identification number of the patients) were removed. From the raw data, the 9 channels common to all EEGs present in the database were selected: Fp1, Fp2, Cz, C3, C4, T3, T4, O1, and O2. To minimize the effect of artifacts generated from power line interference, 50 Hz (power line frequency in Europe) was filtered out of the selected signals using a digital notch filter. Additionally, a digital band-pass filter was used to allow frequencies f in the range f_{low}=0.3 Hz≤f≤f_{high}=70 Hz. In the last step, each signal was scaled by means of the StandardScaler method provided by the Python library Scikit-learn (version 1.5.2) [23]. No further manual artifact rejection has been performed. The resulting preprocessed signals were subsequently used for both the WPD feature extraction and the STFT transformation described in the following sections.

Baseline Model

A slightly modified version of the EEG classification model designed by Albaqami et al [11] was implemented and used as a baseline. In the present study, differently from Albaqami et al [11], the EEGs are segmented into 10-second long time windows. In addition, the data are augmented by the use of 4 offset values (0, 2.5, 5, and 7.5 seconds) that define the starting point from which the segmentation process starts (Figure 3). Additionally, for the present study, the patient's age was normalized and concatenated to the dataset.

Figure 3. A portion of a preprocessed electroencephalography signal, segmented into 10-second long TW starting from an offset value of 2.5 seconds. TW: time window.



The optimal settings for the baseline model used for the final evaluations were determined by means of a grid search. Two gradient boosting algorithms (extreme gradient boosting [XGBoost] [24] and categorical boosting [CatBoost] [25]) were chosen, and 3 common parameters were varied during the training phase. For each combination of the parameters and algorithm, the validation accuracy was measured. The parameters included in the grid search are listed in Table S1 in Multimedia Appendix 1.

CNN Model

CNN Architecture

In Figure 4 (created in BioRender [26]), a diagram of the CNN architecture implemented for this study is presented. The CNN block consists of a pretrained ResNet-50 architecture [19] that was modified by increasing the number of input channels to 9 and copying the weights and biases of the 3 original channels sequentially to the newly created channels. The output of the CNN block is reshaped by means of a global average pooling operation, after which it passes through a dropout layer and 3 fully connected layers before the final classification is made. The age of the patients is scaled in the range [0, 1] by means of the MinMax method provided by the Python library Scikit-learn (version 1.5.2) [23]. Subsequently, this feature is concatenated to the second fully connected layer (Fc2).

XSL•FO
RenderX

Figure 4. CNN-based architecture is used to classify the images. The input data consist of a multichannel image together with the corresponding patient's age. The image is fed to a modified ResNet-50 block, adapted to receive 9-channel images. After a dropout layer and a first fully connected layer (Fc1), the scaled age is concatenated into a second fully connected layer (Fc2). After a third fully connected layer (Fc3), the sample is classified as normal or abnormal. Created in BioRender. CNN: convolutional neural network.

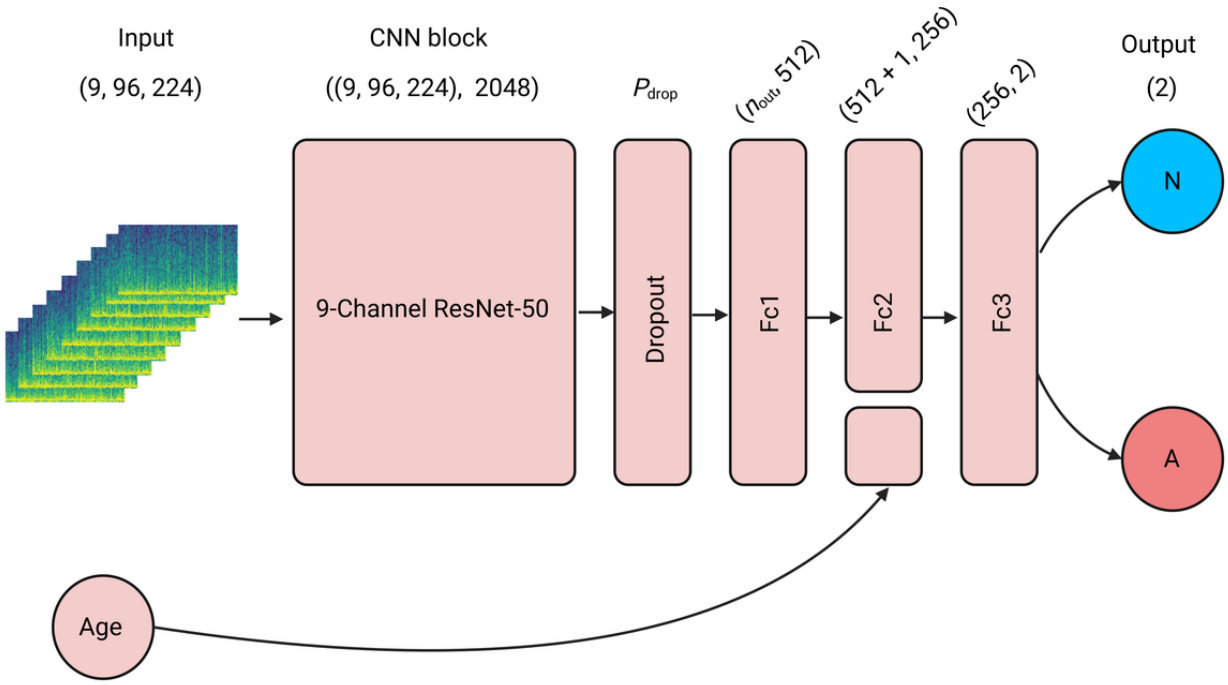


Image Generation

The CNN architecture described earlier was designed to accept 9-channel images that originated from the selected EEG channels. After the common preprocessing steps described earlier, 2 different EEG-to-image conversion pipelines were investigated: a WPD-based representation and an STFT-based representation.

WPD-Based Image Generation

For the WPD representation, wavelet packet features were extracted from the preprocessed EEG signals using the same WPD procedure as described for the baseline model, but without applying the subsequent feature aggregation step. For each EEG recording, the resulting coefficients were arranged into a 9-channel image, where each channel corresponded to one of the selected EEG electrodes and the pixel values represented the extracted WPD features for the respective time windows. To increase the amount of the training data, images were generated from WPD features extracted at different offset values: 0, 2.5, 5, and 7.5 seconds.

STFT-Based Image Generation

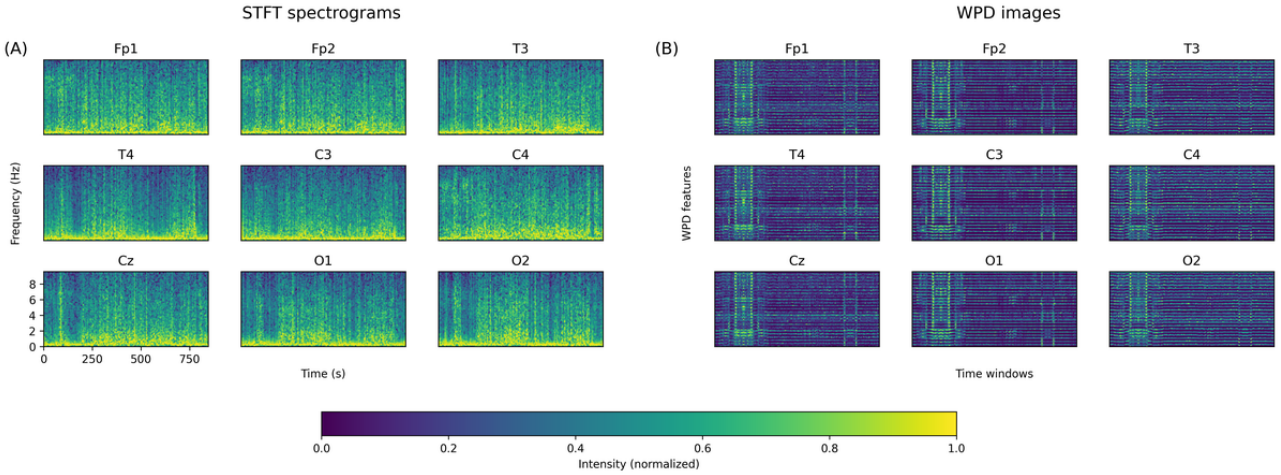
For the STFT representation, the preprocessed EEG signals were transformed directly using the STFT [27]. The STFT was computed using a Hann window, with a window and fast Fourier transform length of 2000 samples (10 seconds at a sampling frequency of 200 Hz) and a hop length of 1000 samples (5 seconds), resulting in a frequency resolution of 0.1 Hz. The magnitude of the STFT coefficients was logarithmically transformed. To match the dimensionality of the WPD images, the STFT representations were limited to 96 frequency bins. Given the frequency resolution of 0.1 Hz, these bins corresponded to frequencies ranging from 0 to 9.5 Hz. In addition, here, the data were augmented by generating STFT representations at different offset values: 0, 2.5, 5, and 7.5 seconds.

Both STFT and WPD images were reshaped to a common width of 256. In 879 of the 916 (96%) EEGs, the resulting images had a width that was smaller than 256 pixels. For those images, symmetric padding was applied. Subsequently, a random crop operation was performed to reach a final width of 224 pixels. For the remaining 37 (4%) EEGs, which resulted in images with a width larger than 256, a random crop was applied without prior padding. Finally, the images were normalized in the range [0, 1].

The resulting images therefore had a common spatial dimension and consisted of 9 channels corresponding to the 9 selected EEG electrodes. An example of a set of WPD images and STFT spectrograms generated from the 9 channels of an abnormal EEG acquired on a 6-year-old patient is presented in Figure 5.

XSL•FO
RenderX

Figure 5. (A) STFT spectrograms generated from an abnormal EEG recording acquired on a 6-year-old patient. (B) WPD images generated from an abnormal EEG recording acquired on a 6-year-old patient. The images have a height of 96 pixels (corresponding to the extracted WPD features) and a width in pixels equal to the number of extracted time windows. EEG: electroencephalography; STFT: short-time Fourier transform; WPD: wavelet packet decomposition.



The prepared WPD and STFT images of the training dataset were presented separately to the CNN architecture described earlier. For each image representation, a grid search was performed to investigate the effect of selected CNN hyperparameters on validation accuracy (Table S2 in Multimedia Appendix 1). The investigated hyperparameters included batch size, dropout probability, weight decay, and learning rate. Dropout and weight decay were included as regularization strategies to reduce the risk of overfitting. In each run, the CNN was trained for a total of 100 epochs, and the set initial learning rate was reduced by a factor of 10 after the first 70 epochs. The validation performances obtained with the WPD and STFT representations were subsequently compared. Since the STFT-CNN configurations outperformed the corresponding WPD-CNN configurations in 176 runs out of the total 192 (92%) grid search runs, the STFT representation was selected for the final CNN model. The WPD-CNN was therefore not evaluated on the test dataset. The best-performing STFT-CNN hyperparameter configuration was subsequently used to train new final CNN models, as described in the CNN Model Performance section.

Model Training and Evaluation

The labeled data were split into train and test datasets with a ratio of 80/20. The split was performed at the EEG measurement level rather than at the patient level. The splits were stratified with respect to the target label and age to conserve the original ratio between normal and abnormal measurements and the distribution of patients' ages in both datasets. Since multiple EEG measurements from the same patient were present in the abnormal dataset, this procedure could result in recordings from the same patient being assigned to different subsets. A retrospective analysis identified 24 patients with recordings present in both the training and test datasets. Consequently, 32 of the 69 (46%) abnormal recordings in the test dataset originated from patients who also contributed recordings to the training dataset. The training data were further split into 3 stratified folds that were used for cross-validation training. The test dataset was not used for model training and was used only for the final evaluations of the models.

The best-trained models given by the grid searches were used to create classification predictions on the augmented test data. Each EEG gave rise to 3 vectors (1 for each cross-validation fold model) containing 4 prediction probabilities (1 for each offset value) $P=(P_0, P_{2.5}, P_5, P_{7.5})$, resulting in a total of 12 prediction probabilities. Thus, the final prediction probability was given by the average of the 12 vector components. A measurement was classified as normal if the average prediction probability was smaller than 0.5 and as abnormal if the probability was greater than or equal to 0.5. To assess and compare the performance of the models, the following performance metrics were measured: accuracy, specificity, sensitivity, and F_1-score. In addition, from the prediction's probabilities, the receiver operating characteristic (ROC) curves were computed together with their areas under the curve (AUC). To gain a more in-depth understanding of the importance of age for the automatic EEG interpretation, the predictions made on the test data were grouped according to the age groups described before, and all performance metrics were calculated on those subsets separately as well as on the whole test dataset.

Statistical Analysis

Variables were reported as mean and SD if normally distributed and otherwise as median and IQR. The Shapiro-Wilk test was used to assess normality [28]. The age was not distributed normally and exhibited a median of 6 (IQR 1-13) years. The 95% CIs for the mean age were estimated using a bootstrapping approach with 1000 iterations [29]. The resulting 95% CI ranged from 6.6 to 7.4 years.

To assess statistical significance between the models' performances, the McNemar test was used to compare the final classifications, the DeLong test was used to compare the ROC-AUC values, and paired bootstrap analyses were used to compare specificity, sensitivity, and F_1-score. For each paired bootstrap analysis, the difference was calculated as the performance metric of the first model listed in the comparison minus that of the second model. All statistical analyses were conducted using Python (NumPy version 1.26.4 and SciPy

XSL•FO
RenderX

version 1.14.1; Python Software Foundation). *P* values ≤.05 were considered statistically significant.

Results

Patient Characteristics

In Figure 6A [30], the duration distribution of the labeled data is presented. The measurements last on average 16.6 (SD 3.3) minutes, with the error given by the SD. This value is consistent with the typical duration of routine EEGs [31]. In Table 1, descriptive statistics of the duration distributions such as mean, median, SD, and IQR are listed.

Figure 6. (A) Duration distribution of the labeled electroencephalography (EEG) considered for the analysis. (B) Age distribution of the labeled EEGs. In the background, the distribution of the EEGs grouped according to the scheme proposed by Eisermann et al is shown.

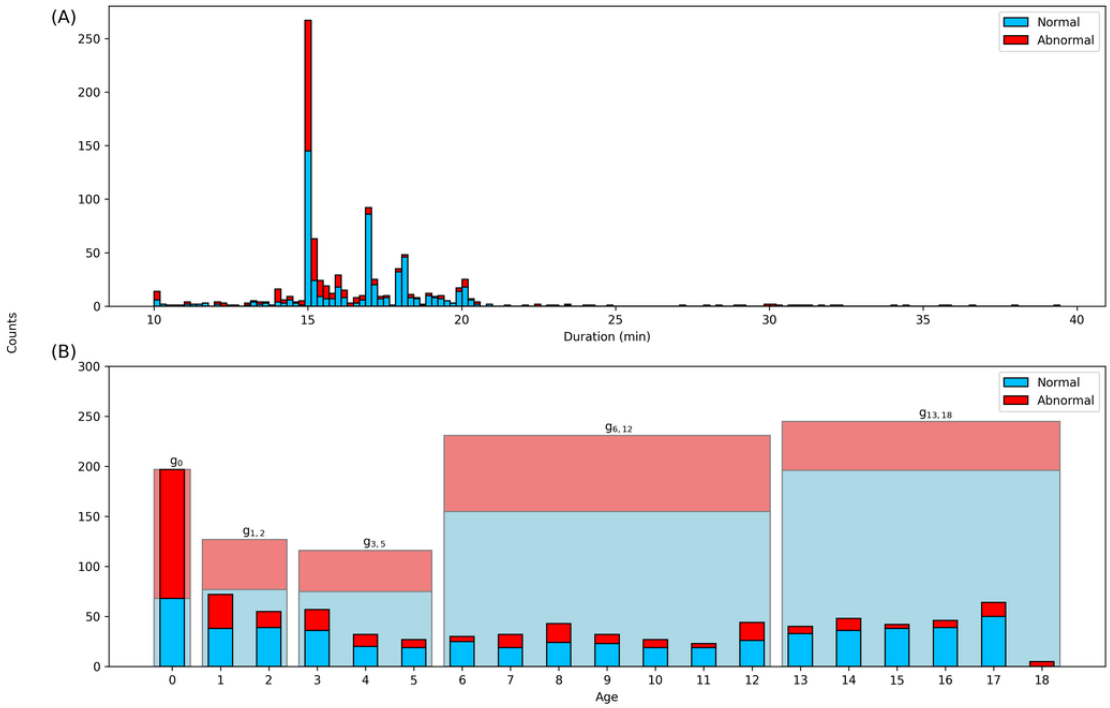


Table 1. Descriptive statistics of the duration of the electroencephalography measurements.

Label	Values, n	Mean (SD) (minutes)	Median (IQR) (minutes)
Normal	571	16.6 (2.2)	17.0 (15.0-18.0)
Abnormal	345	16.7 (4.6)	15.1 (15.0-16.4)
All	916	16.6 (3.3)	15.6 (15.0-18.0)

Each patient in the normal dataset contributed only 1 EEG measurement. In contrast, the abnormal dataset comprised 345 recordings from 224 unique patients. Of these, 61 patients contributed more than 1 recording, accounting for 182 recordings in total (median 2, IQR 2-21 recordings per patient). Consequently, the 916 EEG measurements do not correspond to 916 unique patients.

The age distribution of the data is presented in Figure 6B. This feature exhibited a median of 6 (IQR 1-13) years. The 95% CIs for the mean age ranged from 6.6 to 7.4 years. With 197 EEGs, 0-year-old patients accounted for the highest number of recordings. To study the effect of age on the performance of the trained models, the patients were arranged into 5 separate age groups according to the grouping scheme for pediatric EEGs proposed by Eisermann et al [30]. The distribution between normal and abnormal EEGs in these groups is shown in Figure 6B. In Table 2, the number of normal and abnormal measurements as well as their total sum is listed for each age group.

XSL•FO
RenderX

Table 2. Grouping of the dataset according to the scheme proposed by Eisermann et al [30].

Group name	Minimum age (years)	Maximum age (years)	Normal, n	Abnormal, n	Total, n
g_0	0	0	68	129	197
$g_{1,2}$	1	2	77	50	127
$g_{3,5}$	3	5	75	41	116
$g_{6,12}$	6	12	155	76	231
$g_{13,18}$	13	18	196	49	245

Baseline Model Performance

All combinations of the grid search parameters resulted in a total of 54 models trained by means of a 3-fold cross-validation. The XGBoost algorithm outperformed the CatBoost algorithm in all runs except one. The model achieving the highest validation accuracy was trained with the XGBoost algorithm set with the following parameters: max_depth=6, $n_{estimators}$=500, and η=0.05. Those settings were used for the final evaluations on the test dataset (see the Model Comparison section for a thorough overview of the models' performances). Additionally, to study the effect of age on the classification's performance, this feature was removed from the input data, and an XGBoost model was trained using the best parameters found by the grid search.

CNN Model Performance

All combinations of the parameters defined by the grid search resulted in a total of 192 trained models for each type of input images (WPD or STFT). In 176 of the 192 (92%) grid search runs, the CNN trained with the STFT images outperformed the same CNN trained with the WPD images when comparing the validation accuracies. The highest measured validation accuracy was provided by a model trained with STFT images and initialized with a batch size of 8, a dropout probability of 0, weight decay set to 0.01, and a learning rate of 0.001. Those settings were used for the final evaluations of the CNN model (from now on STFT-CNN). To reach convergence for the final model, the number of epochs was increased to 500. In addition, the initial learning rate was decreased by a factor of 10 after 300 and 400 epochs. This final model was trained in 3-fold cross-validation. Furthermore, for this approach, to study how age affects the classification's performance, a further model was trained in 3-fold cross-validation with the same best parameters found by the grid search but excluding the age feature from the input data.

Model Comparison

Models Trained With Age Feature

The main classification results are summarized in Table 3. When trained including the age feature, the STFT-CNN and baseline models are capable of correctly classifying EEG recordings from the test dataset with a similar degree of accuracy of approximately 85% and 83%, respectively. The STFT-CNN model exhibits higher sensitivity and F_1-score (approximately 13 and 6 percentage points higher, respectively) than the baseline model, thereby indicating better abilities in identifying abnormal EEG recordings. On the contrary, the baseline model is better able to correctly classify normal EEG recordings and shows a higher specificity (around 93%) when compared to the STFT-CNN model (around 90%). Those behaviors are also confirmed by the confusion matrices of the 2 models displayed in Figure 7. The STFT-CNN correctly identifies 9 more abnormal measurements than the baseline model, which, conversely, is able to spot 4 more normal measurements than the STFT-CNN model. This behavior is further illustrated by the ROC curves computed from the prediction's probabilities performed at different thresholds (Figure 8A) [30]. For false positive rates (ie, 1–specificity), values smaller than 50%, higher sensitivity values are measured with the STFT-CNN model.

Table 3. Overview of the classification's performance of the short-time Fourier transform (STFT)-convolutional neural network (CNN) and baseline models trained with and without the patient's age as an input feature[a].

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	F_1-score (%)	ROC[b]-AUC[c] (%)
STFT-CNN+age	*85.3* [d]	89.6	*78.3*	*80.0*	*89.6*
Baseline+age	82.6	*93.0*	65.2	73.8	87.4
STFT-CNN–age	84.2	88.7	76.8	78.5	89.2
Baseline–age	79.9	92.2	59.4	68.9	87.2

[a]The performances were computed on the test dataset that was kept separate from the training process.

[b]ROC: receiver operating characteristic.

[c]AUC: area under the curve.

[d]Values in italics format indicate the best performance for each metric.

XSL•FO
RenderX

Figure 7. Confusion matrices of the STFT-CNN and baseline models computed on the test dataset. Both models were trained with and without the age feature. CNN: convolutional neural network; STFT: short-time Fourier transform.

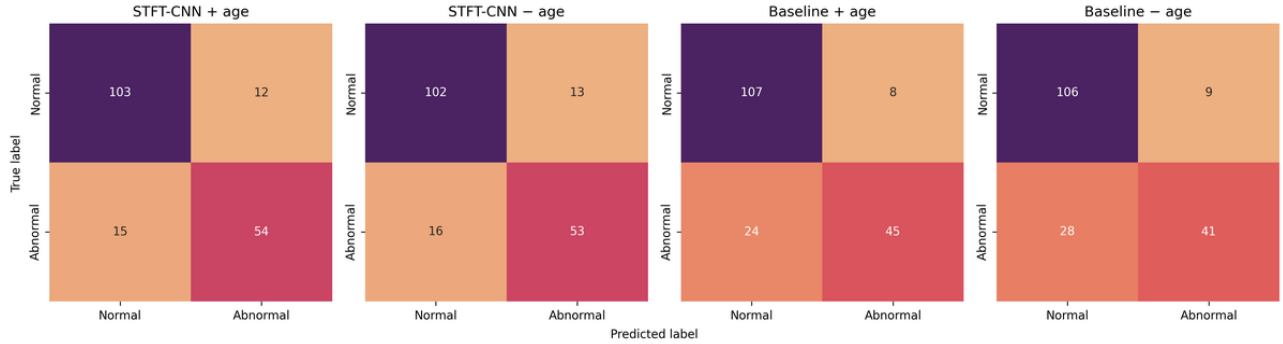
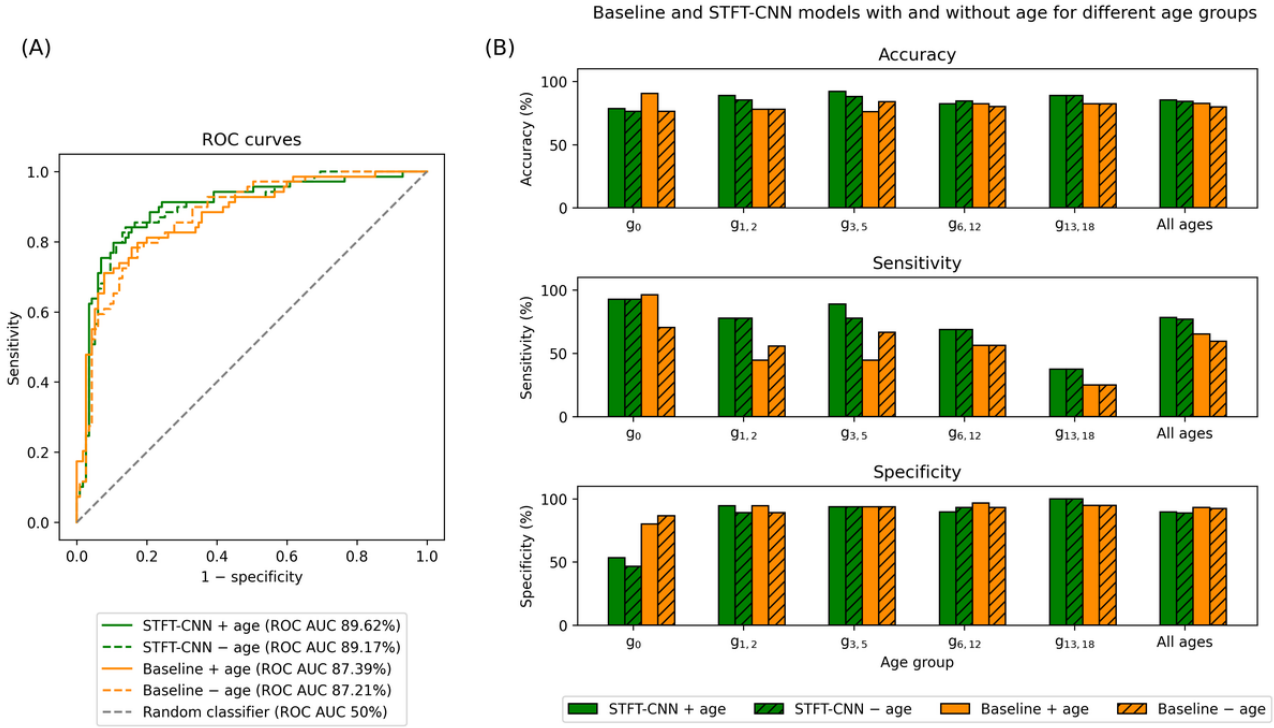


Figure 8. (A) ROC curves of the STFT-CNN (green) and baseline (yellow) models trained with (continuous lines) and without (dashed lines) the age feature. The prediction's probabilities and corresponding thresholds were computed on the test dataset. (B) Accuracy, sensitivity, and specificity measured on separate age groups. The test dataset was split into 5 age groups according to the grouping scheme proposed by Eisermann et al, and the performance metrics were computed for all groups for both STFT-CNN (green bins) and baseline models (yellow bins) trained with (plain color) and without (hatched bins) age. AUC: area under the curve; CNN: convolutional neural network; ROC: receiver operating characteristic; STFT: short-time Fourier transform.



Despite these differences in the individual performance metrics, the statistical analyses did not reveal significant differences between the 2 models for most performance metrics. The McNemar test showed no significant difference in the final classifications performed by the models (P=.46), and the DeLong test indicated no significant difference between their ROC-AUC values (P=.14). The paired bootstrap analysis showed that the observed differences in F_1-score (baseline–STFT-CNN: –0.06, 95% CI –0.15 to 0.02; P=.16) and specificity (baseline–STFT-CNN: 0.03, 95% CI –0.02 to 0.10; P=.31) were compatible with 0. In contrast, the difference in sensitivity was statistically significant (baseline–STFT-CNN: –0.13, 95% CI –0.25 to –0.02; P=.03), indicating higher sensitivity for the STFT-CNN model.

Models Trained Without Age Feature

The baseline model trained without the age feature shows a reduction in all performance metrics compared with the same model type trained including age (Table 3). In particular, a decrease of approximately 3%, 1%, 6%, and 5% is measured on accuracy, specificity, sensitivity, and F_1-score, respectively. This decline in performance suggests that the age feature is an important component for interpreting EEG recordings by means of the baseline model. A further clue is provided by the feature importance scores of the baseline model trained including the patient's ages. For the 3 cross-validation folds, the age feature is ranked at the 41st, 79th, and 57th position with a gain score of 14.8, 10.6, and 12.3, respectively. Therefore, with a total of 1729 input features, age belongs to the top 2%, 5%, and 3% of the most important features of the models for the 3 different folds, respectively. Despite the measured drop in performance,

XSL•FO
RenderX

the statistical tests showed no statistically significant difference between the models: the McNemar test for accuracy (P=.30), the DeLong test for ROC-AUC (P=.18), and paired bootstrap analysis for specificity (baseline with age–baseline without age: 0.01, 95% CI –0.03 to 0.05; P=.82), sensitivity (baseline with age–baseline without age 0.06, 95% CI –0.03 to 0.15; P=.26), and F_1-score (baseline with age–baseline without age: 0.05, 95% CI –0.02 to 0.12; P=.18). The STFT-CNN model trained without the age feature performed equally well as the original STFT-CNN model. As shown in Table 3, all performance metrics exhibit no significant difference, and the ROC curves computed from the predicted probabilities of both models overlap and cover almost the same area (Figure 8A).

When comparing the baseline and the CNN models trained without the age feature, the differences in the final classification performance were not statistically significant according to the McNemar test (P=.23). Similarly, the difference in specificity was not significant according to the paired bootstrap analysis (baseline without age–STFT-CNN without age: 0.03, 95% CI –0.03 to 0.09; P=.31). However, the CNN model showed a higher F_1-score (0.79 vs 0.69) and sensitivity (0.77 vs 0.59) compared with the baseline model without age, with the paired bootstrap analysis indicating evidence of a difference for the F_1-score (baseline without age–STFT-CNN without age –0.10, 95% CI –0.20 to 0.00; P=.05) and sensitivity (baseline without age–STFT-CNN without age: –0.17, 95% CI –0.30 to –0.05; P=.01). A similar trend was observed for the ROC-AUC, where the DeLong test indicated an evidence of a difference between the 2 models (P=.05). Overall, these results suggest that, in the absence of age information, the STFT-CNN model retains its discriminative ability, whereas the baseline model appears to rely more strongly on age-related information for optimal performance. The ROC curves computed from the predicted probabilities further support this observation, with the STFT-CNN model achieving a slightly larger area under the curve compared with the baseline model without age (Figure 8A). This indicates improved discrimination between abnormal and normal EEG recordings across different classification thresholds.

Models' Performances on Different Age Groups

In Figure 8B, a graphical representation of the accuracies, sensitivities, and specificities is provided for both STFT-CNN and baseline models trained with and without age and measured across all age groups. All other performance metrics can be inspected in Table S3 in Multimedia Appendix 1. Additionally, the number of normal and abnormal EEG recordings in the test dataset across all age groups is provided in Table S4 in Multimedia Appendix 1.

The STFT-CNN model trained with and without age outperforms the baseline models in all age groups except for g_0. The STFT-CNN model trained with the patient's ages shows higher or equal performances when compared to the same model trained without this feature, except in the group $g_{6,12}$. The highest performances are measured on the age group $g_{3,5}$, where this model reaches accuracy, sensitivity, and specificity of 92%, 89%, and 94%, respectively. On g_0, the baseline model trained with the age feature outperforms the STFT-CNN models, reaching a test accuracy of around 90%. A decrease in accuracy and sensitivity of around 14% and 26% is measured on the performance of the baseline model trained without age, calculated on the subset g_0.

In general, in all age groups (with the exception of g_0), the STFT-CNN models tend to show higher sensitivities than the baseline models, whereas the latter models show similar or noticeably higher specificities than the one measured with the STFT-CNN models, thereby reflecting the general performances measured across the whole test data comprising all ages.

Discussion

In the present work, an EEG classification pipeline was designed. The pipeline was trained and tested on labeled normal and abnormal pediatric EEG signals acquired in the EEG department and the PICU of a tertiary children's hospital. A model designed by Albaqami et al [11] was reproduced and modified to accommodate the purpose of the present work. This baseline model was compared to a CNN-based model that classified images generated from the extracted WPD features prior to the aggregation process and from STFT spectrograms of the individual EEG signals. The training of the CNN fed with WPD images was discontinued after the grid search tuning, as 176 of the 192 (92%) grid search runs performed worse than the CNNs initialized with the same hyperparameters but fed with STFT images.

The effect of age on the model's performance was explored by removing this input feature from both approaches. The overall best performances are provided by the STFT-CNN model trained with the age feature. This trend was observed across almost all performance metrics (except for specificity) and was reflected also in the ROC curves. A plausible explanation is that the time-frequency representation of the STFT spectrograms conserves the temporal information stored in the input signals better than the aggregation of the WPD features. The STFT-CNN model may therefore be able to learn time-dependent transient patterns and rhythmic abnormalities directly from local spectral changes. In contrast, the feature aggregation step condenses those time-dependent patterns into single features, with a potential loss of temporal information. The consistently lower performance observed in the CNN model trained using WPD images is compatible with this interpretation. In a PICU screening context, sensitivity may be considered particularly relevant because failure to detect an abnormal EEG could delay further clinical assessment or monitoring. The higher sensitivity observed for the STFT-CNN model compared with the baseline model may therefore be clinically relevant from a screening perspective, enabling a window of opportunity for neuroprotective interventions with the potential to improve patient outcomes downstream.

In terms of accuracy, the baseline model trained with the age feature showed comparable performance as the STFT-CNN models. The removal of the patients' ages from the baseline model resulted in a drop in all measured metrics. Contrary to expectations, the CNN model without age performed equally

XSL•FO
RenderX

well as the CNN model trained with the age feature, with no statistically significant difference measured across the performance metrics. Most probably, the concatenation of this single feature to a layer (Fc2) with a rather large number of input connections (n=512) is not ideal. Therefore, the information provided by age is not treated with enough weight to influence the performance of the model. It is also possible that some age-dependent patterns (like the different rhythm velocities in background activity [32]) undergo some kind of intrinsic normalization process due to the various scaling and feature extraction steps applied before the classification. Despite the observed differences in performance, most comparisons between the models did not reach statistical significance. Therefore, the results do not provide sufficient evidence to conclude that the STFT-CNN approach is superior to the baseline approach. Rather, they indicate a consistent performance trend that is compatible with the hypothesis that preserving temporal information in the time-frequency representation may be beneficial. Increasing the number of EEG recordings in the test set would improve the statistical power of the analyses and allow smaller differences in model performance to be detected with greater confidence.

The performance metrics were also computed for different subsets of the test dataset. Those 5 separate age groups follow the scheme proposed by Eisermann et al [30]. This analysis gave further insight into some of the general findings observed on the whole test data. In almost all groups, the CNN models showed slightly higher accuracies and sensitivities than the baseline models, whereas those latter ones tended to show higher specificities. The big drop in performance of the baseline model when the age feature was removed was especially visible in the group g_0. This latter group was the only subset where the baseline model trained with the age feature outperformed the STFT-CNN models.

Several limitations of the dataset should be considered when interpreting these results. First, the number of labeled abnormal EEGs was relatively limited compared with the normal dataset, and the abnormal recordings were not evenly distributed across age groups. In particular, 129 of the abnormal EEGs were acquired from patients aged 0 years, resulting in comparatively limited numbers of recordings in the older age groups. Therefore, the performance observed for these age groups, as well as the conclusions regarding the influence of age, should be interpreted with caution. A larger and more age-balanced dataset containing abnormal EEGs across the entire pediatric age range would be required to assess the generalizability of the models more reliably.

An additional limitation arises from differences in the populations from which the normal and abnormal EEGs were obtained. The normal EEGs were recorded in healthy children without neuroactive medication or known cerebral disease, whereas the abnormal EEGs were obtained from critically ill PICU patients with manifest or suspected neurological disease. The models may therefore have learned not only features related to EEG abnormality, but potentially also characteristics associated with the different patient populations and recording settings. The present findings should thus be considered an initial demonstration of the models' ability to distinguish normal from abnormal EEGs within this dataset. Further validation in an independent and representative PICU cohort is needed before the models can be considered applicable for general bedside abnormality detection.

A further limitation concerns the patient-level distribution of the recordings. While each patient contributed only 1 EEG measurement to the normal dataset, repeated measurements were available for some patients in the abnormal dataset. As the dataset was split at the EEG measurement level, a retrospective analysis identified 24 patients with recordings represented in both the training and test datasets. In critically ill patients, EEG activity may change considerably over time depending on the neurological and clinical state, such that repeated recordings from the same patient may capture distinct electrophysiological states rather than represent redundant observations. Nevertheless, these recordings share patient-specific characteristics and cannot be considered fully independent. The resulting patient-level overlap may therefore introduce some dependence between the training and test datasets and may lead to optimistic estimates of model performance. Therefore, the reported test performance should not be interpreted as a fully independent assessment of generalization to previously unseen patients. A patient-level resplitting and subsequent re-evaluation of the present dataset was not feasible. In addition, the EEG labels were based on expert interpretation by a certified pediatric neurologist without independent review or formal consensus adjudication; thus, interrater reliability could not be assessed. Future studies should evaluate the models in temporally separated and clinically well-defined cohorts with patient-level separation and, ideally, independent expert assessment of EEG labels to further establish model generalizability and the robustness of the reference standard.

A further limitation concerns the frequency range represented in the STFT-CNN input. Although the EEG signals were band-pass filtered between 0.3 and 70 Hz during preprocessing, only the first 96 STFT frequency bins were retained, corresponding to a frequency range of 0-9.5 Hz. Consequently, spectral information above 9.5 Hz was not available to the STFT-CNN. The performance of this model should therefore be interpreted as reflecting classification based predominantly on low-frequency EEG information rather than the full preprocessed frequency range. Future studies should investigate whether including a broader frequency range provides additional discriminative information and further improves classification performance.

The CNN training procedure included several measures to reduce the risk of overfitting. Hyperparameters including dropout, weight decay, batch size, and learning rate were investigated using a grid search based on validation accuracy. The final models were trained using 3-fold cross-validation, and the learning rate was progressively reduced during training. Nevertheless, overfitting remains a relevant concern, given the limited number of independent patients and the relatively small number of abnormal EEG recordings. Future studies should therefore evaluate the models on larger datasets and, where possible, use independent external test cohorts.

XSL•FO
RenderX

Future work should also investigate potential design modifications to the STFT-CNN architecture in order to enhance the role of age as an input feature. In addition, more detailed annotations of abnormal EEGs covering a broader spectrum of possible abnormalities and identifying their location within the signals could enable more targeted classification approaches. In such a manner, more powerful approaches could be used to classify the data. The generation of images could be performed by means of a continuous wavelet transform [33]. Due to the confined abnormality regions, such images would be of manageable size and would reflect better the different duration of the various frequency components. In addition, the CNN architecture could be explored with the use of GradCAM methods [34] in order to compare the results to the annotated signals and acquire more explainability.

With the implemented CNN approach, the age of the patients seems to play a marginal role when classifying the measurements. Therefore, it could be advantageous to also use normal and abnormal EEGs from adult patients (like the TUH Abnormal EEG Corpus version 2.0.0 [9]) to enhance the quantity and variability of the data and consequently improve the performance of the models. It would be also interesting to assess the performance of more recent EEG interpretation pipelines (eg, Standardized Computer-based Organized Reporting of EEG-Artificial Intelligence [13]) and compare their performance to the ones achieved by our model. It is important to note that most of the studies conducted on this topic usually exclude abnormal readings from intensive care unit or pediatric patients [9,13]. For this reason, lower performance should be expected when feeding those pipelines with the data collected for this study.

In conclusion, the present study demonstrated the potential for the AI-supported automated interpretation of pediatric EEGs, with a focus on EEGs acquired in the PICU and comparison with EEGs from healthy children outside the PICU. Future enhancement in data diversity and improvements in the AI approaches are needed to validate our findings.

Acknowledgments

The authors acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to reasons of data protection and privacy but are available from Data-Governance@uk-essen.de on reasonable request.

Funding

This research received no specific funding from any funding agency in the public, commercial, or not-for-profit sectors.

Generative AI Use

ChatGPT and Grammarly supported manuscript drafting and improved clarity and readability. All AI-assisted content and code were reviewed, verified, and revised by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

Authors' Contributions

NB, SG, L Paul, CD-S, ADM, and UF-M provided the labeled electroencephalography recordings. L Poggi, SSB, NF, and RH designed and implemented the classification pipeline. L Poggi, RH, and NB cowrote the manuscript. CJS, FN, and JK critically revised the manuscript. All authors approved the final content of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary information on model hyperparameter optimization and age-specific performance evaluation, including the grid search parameters used to train the baseline and convolutional neural network (CNN) models, model performance metrics across individual age groups, and the distribution of normal and abnormal electroencephalography (EEG) recordings across age groups in the test dataset.
[DOCX File , 1755 KB-Multimedia Appendix 1]

References

1. Tatum WO. EEG essentials. Contin Minneap Minn. 2022;28(2):261-305. [doi: 10.1212/con.0000000000001129]
2. Ahrens S, Twanow JD, Vidaurre J, Gedela S, Moore-Clingenpeel M, Ostendorf AP. Electroencephalography technologist inter-rater agreement and interpretation of pediatric critical care electroencephalography. Pediatr Neurol. 2021;115:66-71. [FREE Full text] [doi: 10.1016/j.pediatrneurol.2020.10.016] [Medline: 33333462]

XSL•FO
RenderX

3. Alkhachroum A, Appavu B, Egawa S, Foreman B, Gaspard N, Gilmore EJ, et al. Electroencephalogram in the intensive care unit: a focused look at acute brain injury. Intensive Care Med. 2022;48(10):1443-1462. [FREE Full text] [doi: 10.1007/s00134-022-06854-3] [Medline: 35997792]
4. Löffelhardt VT, Della Marina A, Greve S, Müller H, Felderhoff-Müser U, Dohna-Schwake C, et al. Characterization of aEEG during sleep and wakefulness in healthy children. Front Pediatr. 2021;9:773188. [FREE Full text] [doi: 10.3389/fped.2021.773188] [Medline: 35127587]
5. Paul L, Greve S, Hegemann J, Gienger S, Löffelhardt VT, Della Marina A, et al. Association of bilaterally suppressed EEG amplitudes and outcomes in critically ill children. Front Neurosci. 2024;18:1411151. [FREE Full text] [doi: 10.3389/fnins.2024.1411151] [Medline: 38903601]
6. Greve S, Löffelhardt VT, Della Marina A, Felderhoff-Müser U, Dohna-Schwake C, Bruns N. The impact of age and electrode position on amplitude-integrated EEGs in children from 1 month to 17 years of age. Front Neurol. 2022;13:952193. [FREE Full text] [doi: 10.3389/fneur.2022.952193] [Medline: 36090865]
7. Benedetti GM, Silverstein FS, Rau SM, Lester SG, Benedetti MH, Shellhaas RA. Sedation and analgesia influence electroencephalography monitoring in pediatric neurocritical care. Pediatr Neurol. 2018;87:57-64. [doi: 10.1016/j.pediatrneurol.2018.05.001] [Medline: 30049426]
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
9. López S, Suarez G, Jungreis D, Obeid I, Picone J. Automated identification of abnormal adult EEGs. IEEE Signal Process Med Biol Symp. 2015;2015. [FREE Full text] [doi: 10.1109/SPMB.2015.7405423] [Medline: 27195311]
10. Gemein LAW, Schirrmeister RT, Chrabąszcz P, Wilson D, Boedecker J, Schulze-Bonhage A, et al. Machine-learning-based diagnostics of EEG pathology. Neuroimage. 2020;220:117021. [FREE Full text] [doi: 10.1016/j.neuroimage.2020.117021] [Medline: 32534126]
11. Albaqami H, Hassan GM, Subasi A, Datta A. Automatic detection of abnormal EEG signals using wavelet feature extraction and gradient boosting decision tree. Biomed Signal Process Control. 2021;70:102957. [doi: 10.1016/j.bspc.2021.102957]
12. Darvishi-Bayazi M-J, Ghaemi MS, Lesort T, Arefin MR, Faubert J, Rish I. Amplifying pathological detection in EEG signaling pathways through cross-dataset transfer learning. Comput Biol Med. 2024;169:107893. [FREE Full text] [doi: 10.1016/j.compbiomed.2023.107893] [Medline: 38183700]
13. Tveit J, Aurlien H, Plis S, Calhoun VD, Tatum WO, Schomer DL, et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. JAMA Neurol. 2023;80(8):805-812. [FREE Full text] [doi: 10.1001/jamaneurol.2023.1645] [Medline: 37338864]
14. Zou Z, Chen B, Xiao D, Tang F, Li X. Accuracy of machine learning in detecting pediatric epileptic seizures: systematic review and meta-analysis. J Med Internet Res. 2024;26:e55986. [FREE Full text] [doi: 10.2196/55986] [Medline: 39661965]
15. Gleason A, Richter F, Beller N, Arivazhagan N, Feng R, Holmes E, et al. Accurate prediction of neurologic changes in critically ill infants using pose AI. medRxiv. Preprint posted online on June 10, 2024. [FREE Full text] [doi: 10.1101/2024.04.17.24305953] [Medline: 38699362]
16. Azriel R, Hahn CD, De Cooman T, Van Huffel S, Payne ET, McBain KL, et al. Machine learning to support triage of children at risk for epileptic seizures in the pediatric intensive care unit. Physiol Meas. 2022;43(9). [doi: 10.1088/1361-6579/ac8ccd] [Medline: 36007520]
17. Si Y, Gotman J, Pasupathy A, Flanagan D, Rosenblatt B, Gottesman R. An expert system for EEG monitoring in the pediatric intensive care unit. Electroencephalogr Clin Neurophysiol. 1998;106(6):488-500. [doi: 10.1016/s0013-4694(97)00154-5] [Medline: 9741748]
18. Montazeri S, Pinchefsky E, Tse I, Marchi V, Kohonen J, Kauppila M, et al. Building an open source classifier for the neonatal EEG background: a systematic feature-based approach from expert scoring to clinical visualization. Front Hum Neurosci. 2021;15:675154. [FREE Full text] [doi: 10.3389/fnhum.2021.675154] [Medline: 34135744]
19. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016. Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016:770-778; Las Vegas, NV, United States. [doi: 10.1109/CVPR.2016.90]
20. Britto JW, Frey LC, Hopp JL. Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants. Chicago, IL. American Epilepsy Society; 2016.
21. Larson E, Gramfort A, Engemann DA, Leppakangas J, Brodbeck C, Jas M, et al. MNE-Python. URL: https://zenodo.org/records/14519545 [accessed 2026-10-15]
22. Poggi L. Diagram of the implemented EEG classification pipeline. BioRender. 2026. URL: https://app.biorender.com/citation/6ac5542e3eced8c749a9006d [accessed 2026-10-23]
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn. 2011;12:2825-2830. [FREE Full text]
24. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016. Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016:785-794; San Francisco, CA, United States. [doi: 10.1145/2939672.2939785]

XSL•FO
RenderX

25. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. 2018. Presented at: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems; December 3-8, 2018:6639; Montréal, Canada. [doi: 10.5555/3327757.3327770]
26. Poggi L. CNN-based architecture is used to classify the images. BioRender. 2026. URL: https://app.biorender.com/citation/6ac554731e43b34dca9396fb [accessed 2026-10-23]
27. Durak L, Arikan O. Short-time fourier transform: two fundamental properties and an optimal implementation. IEEE Trans Signal Process. 2003;51(5):1231-1242. [doi: 10.1109/tsp.2003.810293]
28. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab. 2012;10(2):486-489. [FREE Full text] [doi: 10.5812/ijem.3505] [Medline: 23843808]
29. Hesterberg T. Bootstrap. WIREs Comput Stat. 2011;3:497-526. [doi: 10.1002/wics.182]
30. Eisermann M, Kaminska A, Moutard M-L, Soufflet C, Plouin P. Normal EEG in childhood: from neonates to adolescents. Neurophysiol Clin. 2013;43(1):35-65. [doi: 10.1016/j.neucli.2012.09.091] [Medline: 23290174]
31. Burkholder DB, Britton JW, Rajasekaran V, Fabris RR, Cherian PJ, Kelly-Williams KM, et al. Routine vs extended outpatient EEG for the detection of interictal epileptiform discharges. Neurology. 2016;86(16):1524-1530. [FREE Full text] [doi: 10.1212/WNL.0000000000002592] [Medline: 26984946]
32. Nayak CS, Anilkumar AC. EEG Normal Waveforms. Treasure Island, FL. StatPearls Publishing; 2025.
33. Wang T, Lu C, Sun Y, Yang M, Liu C, Ou C. Automatic ECG classification using continuous wavelet transform and convolutional neural network. Entropy (Basel). 2021;23(1):119. [FREE Full text] [doi: 10.3390/e23010119] [Medline: 33477566]
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017. Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); October 22-29, 2017; Venice, Italy. URL: https://ieeexplore.ieee.org/document/8237336 [doi: 10.1109/iccv.2017.74]

Abbreviations

AUC: area under the curve
CatBoost: categorical boosting
CNN: convolutional neural network
DL: deep learning
EEG: electroencephalography
ML: machine learning
PICU: pediatric intensive care unit
ROC: receiver operating characteristic
STFT: short-time Fourier transform
WPD: wavelet packet decomposition
XGBoost: extreme gradient boosting

Edited by A Benis; submitted 30.Mar.2026; peer-reviewed by M Al-Agil, X Liang; comments to author 14.Jun.2026; revised version received 07.Oct.2026; accepted 07.Oct.2026; published 13.Nov.2026

Please cite as:
Poggi L, Flaschel N, Becker SS, Nensa F, Kohnke J, Greve S, Paul L, Dohna-Schwake C, Felderhoff-Müser U, Della Marina A, Stein CJ, Hosch R, Bruns N
Automated Detection of Abnormal Pediatric Electroencephalograms Using Machine Learning: Retrospective Study
JMIR Med Inform 2026;14:e96601
URL: https://medinform.jmir.org/2026/1/e96601
doi: 10.2196/96601
PMID:

©Leonardo Poggi, Nils Flaschel, Sven Steffen Becker, Felix Nensa, Judith Kohnke, Sandra Greve, Luisa Paul, Christian Dohna-Schwake, Ursula Felderhoff-Müser, Adela Della Marina, Christopher Johannes Stein, René Hosch, Nora Bruns. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX