




Error-aware CNN improves automatic epileptic seizure detection

Vadim Grubov^{1,a} , Sergei Nazarikov¹, Nikita Utyashev², and Oleg E. Karpov²

¹ Baltic Center for Neurotechnology and Artificial Intelligence, Immanuel Kant Baltic Federal University, 14 A. Nevskogo, Kaliningrad 236016, Russia

² National Medical and Surgical Center named after N.I. Pirogov, Ministry of Healthcare of the Russian Federation, 70 Nizhnaya Pervomayskaya, Moscow 105203, Russia

Received 18 June 2024 / Accepted 1 August 2024

© The Author(s), under exclusive licence to EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract Automated seizure detection is a major challenge in the context of epilepsy diagnostics. There are numerous approaches to this task, but most of them share the same problem—the trade-off between recall and precision, i.e. decent recall is often accompanied by low precision. This ultimately leads to a high number of false positive seizure detections, which in its turn impede automated diagnostics. The purpose of this study is to develop a method to lower the number of false positive predictions in seizure detection task when applied to real EEG recordings. We propose the cascade approach which combines the idea of iterative refinement algorithms and powerful neural networks. The method is tested on unrefined dataset, that includes EEG recordings of epileptic patients from the hospital. Time-frequency analysis based on continuous wavelet transform is used for EEG preprocessing and feature extraction. To provide predictions the approach implements convolutional neural networks. The proposed approach consists of two steps: in the first step a model is trained to provide initial predictions and then in the second step another model is trained with the knowledge of the first model's errors. We evaluate the performance of the approach with the confusion matrix metrics adjusted to the specifics of the epilepsy diagnostics task. We show that the number of false positive predictions decreases by an order of magnitude with the use of the proposed method. We theorize about possible application of this approach within a clinical decision support system.

1 Introduction

Epilepsy is a neurological disease characterized by the recurring seizures that can range from short obscure episodes with no clinical manifestation to prolonged attacks with rigorous shaking and loss of consciousness [1]. Such manifestations negatively affect patient's quality of life, thus antiepileptic treatment becomes crucial [2]. Thanks to the achievements of neuroscience and medicine most patients can achieve remission with medications, neurostimulation or surgery [3–7]. However, any treatment requires epilepsy to be diagnosed in timely fashion, hence this area of medicine is in dire need for practical and precise methods of diagnostics.

Nowadays, the primary diagnostic tool for epilepsy is electroencephalography (EEG). EEG is a non-invasive method of measuring the electrical activity of the brain, which involves placing electrodes on the scalp and recording voltage potentials that result from the current flow from the neurons [8]. Usual diagnostics routine includes prolonged patient monitoring and subsequent deciphering of EEG signals in search of specific epilepsy-related patterns [9]. The prevailing approach to EEG interpretation involves visual inspection, which offers precise diagnosis given the expertise of epileptologist but also presents numerous challenges. Rarity of epileptic seizures combined with variety of their manifestations further complicate manual EEG analysis which is already difficult and time-consuming [10]. Therefore, automated methods for EEG analysis and epilepsy diagnostics are highly desirable [11, 12]. While the idea of fully automated diagnostics is alluring, most of the modern approaches to the task just cannot ensure the same level of precision as an experienced epileptologist. A more realistic solution is Clinical Decision Support System (CDSS) [11], which marks possible seizures, but leaves the final decision to a human. This approach inherits high precision of manual diagnostics, but at the same time reduces the workload of the doctor.

^a e-mail: vvgrubov@gmail.com (corresponding author)

To date, many approaches to the task of automated epilepsy diagnostics were tested, which includes various statistical models [13], expert systems [14], machine learning (ML) [15–17]. The more promising approach, that have seen significant advancements in the recent years, are artificial neural networks (ANNs) [18–20]. This includes a specific variation of ANN—convolutional neural network (CNN), that originally was proposed for image classification, but ultimately earned its place as a powerful instrument for seizure detection [21].

However, most of the approaches for automated seizure detection, including ML- and CNN-based, share the same problem—the trade-off between recall and precision, with decent recall often accompanied by low precision. This ultimately leads to a high number of false positive seizure detections, which in its turn impede automated diagnostics. The main source of this issue is a notorious data imbalance in epileptic datasets, but there are some other reasons related to the way epilepsy diagnostics is handled. In this paper, we proposed a CNN-based algorithm, that eventually generated a large number of false positives. We considered several adjustments to the method aimed at reducing false positives, and ultimately arrived at the idea of cascade algorithm. Multi-stage (cascade) algorithms are widely implemented in ML- and ANN-based tasks [22, 23] and are considered as a standard in industry. The main idea is that cascade algorithms gradually improve the quality of predictions using information from previous stages [22, 24]. Cascade algorithms exist in seizure detection task [13], but an approach that uses multiple CNNs to produce and refine predictions is not well known. In this study, we proposed a two-step algorithm for the seizure detection task. In the first step, we train a CNN model on the original dataset. Then we collect the information about the errors of the first model and use this information to train the more precise second CNN model. We tested this algorithm and compared its effectiveness with the initial CNN approach. General pipeline of the study is shown on Fig. 1.

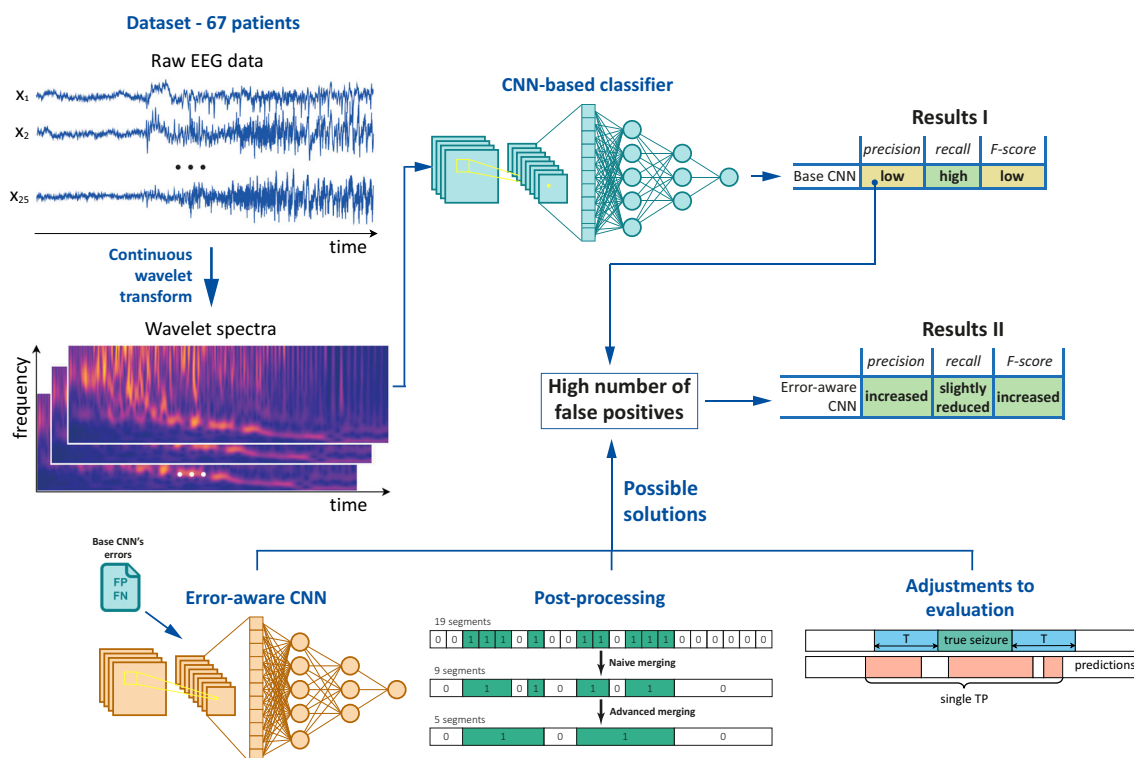


Fig. 1 General pipeline of the study. EEG data from an epilepsy clinical dataset were used to propose a CNN-based classifier for seizure detection. The main issue of the proposed algorithm was high number of false positives. We considered several possible approaches to reduce the number of false positives: (i) post-processing of CNN's predictions, (ii) adjustments to evaluation procedure and (iii) two-step approach with error-aware CNN, which uses information about base CNN's errors during training. We performed evaluation for these approaches and reported a slight reduction in recall but a significant improvement in precision

2 Materials and methods

2.1 Dataset

The dataset analyzed in this work was provided by the National Medical and Surgical Center named after N.I. Pirogov of the Ministry of Health of the Russian Federation (Moscow, Russia). The set includes EEG recordings of patients of the Center who were diagnosed with focal epilepsy in 2017–2019. EEG signals were recorded as a part of long-term patient monitoring aimed at verification of epileptogenic zones for further clinical treatment. The monitoring was conducted during normal daily routine, including sleep and wakefulness. The patient were also exposed to the sessions of photic stimulation and hyperventilation in order to provoke epileptiform activity [25, 26], however, none of the recorded seizures was caused by these sessions. All medical procedures were carried out in accordance with the medical rules of the Center and the Helsinki Declaration, and were approved by the ethics committee. All patients provided written informed consent before participation. The personal data were not included, thus the dataset was anonymized.

Length of monitoring depended on the patient's condition, thus the duration of individual recording varies from 8 to 84 h and includes from 1 to 5 epileptic seizures. Collected EEG data were deciphered manually by an expert, who marked all epileptic episodes and confirmed the diagnosis “focal epilepsy”. Epileptic foci were located in the frontal, temporal, or parietal regions of the left, right, or both hemispheres. There were no uniformity in the manifestation of epilepsy, so the analyzed dataset is a good representation of unstructured data that are dealt with in hospitals. It is important to test diagnostics methods on the data collected through routine clinical practice, since it helps to align the methods with the medical problem they are dealing with [17].

“Micromed” encephalograph (Micromed S.p.A., Italy) was used to record EEG data from $N = 25$ channels at 128 Hz sampling rate. Recording electrodes were placed according to “10–20” scheme uniformly covering the whole cortex [27]. The initial dataset includes recordings of 83 patients, however our previous research on the same set revealed that the data of 16 patients have poor “signal-to-noise” ratio [28–31]. Such issues with the quality of the data make them ill-suited for automated and even manual diagnostics, since expert-made analysis on these data is also dubious. Therefore, we excluded those 16 patients and analyzed the data of the rest 67 patients.

Raw EEG signals went through basic preprocessing, since prolonged recordings are commonly contaminated with artifacts [32]. We applied filtration: band-pass 1–60 Hz filter was used to mitigate some low and high frequency parasitic components in EEG and 50 Hz notch filter suppressed interference from the power grid. Additional preprocessing included procedure based on independent component analysis [33], which aimed to reduce some interfering artifacts such as eye blinks. All EEG data preprocessing was performed with the help of Fieldtrip toolbox for MATLAB [34].

2.2 Data processing

We analyzed EEG signals in time-frequency domain and constructed feature space using continuous wavelet transform (CWT) [35, 36]. CWT suggests convolution of EEG signal $x_n(t)$ with the set of basis functions $\psi_{s,\tau}$:

$$W_n(s, \tau) = \int_{-\infty}^{\infty} x_n(t) \psi_{s,\tau}^*(t) dt, \quad (1)$$

where $n = 1, 2, \dots, N$ is the number of the channel ($N = 25$), $W_n(s, \tau)$ are the coefficients of CWT, and $*$ corresponds to complex conjugation. Each basis function $\psi_{s,\tau}(t)$ is derived from the mother wavelet ψ_0 through some transformation:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi_0\left(\frac{t-\tau}{s}\right), \quad (2)$$

where s is the time scale defining expansion and compression of the original function and τ is the time shift of the original function. In this study, we used complex Morlet wavelet:

$$\psi_0(\eta) = \frac{1}{\sqrt[4]{\pi}} e^{j\omega_0\eta} e^{-\frac{\eta^2}{2}}, \quad \eta = \frac{t-\tau}{s}, \quad (3)$$

where ω_0 is the central frequency of the wavelet. Here we use a special case $\omega_0 = 2\pi$, which provides simple relation between time scales of CWT and frequencies: $f \approx 1/s$.

The most common CWT-based characteristic is wavelet power (WP), as it reflects the time-frequency structure of EEG signals [35, 37]. Here, we considered WP as:

$$w_n(f, \tau) = |W_n(f, \tau)|^2, \quad (4)$$

We considered WPs for each EEG channel in the frequency range of 1–40 Hz as a feature space for the models. We used Python package MNE to perform CWT-based analysis [38].

In the temporal domain we segmented WP into 10-s intervals, thus the task of epileptic seizure detection was reduced to the task of classification of 10-s intervals. Length of 10 s was chosen based on the typical time-frequency structure of epileptic seizure. The average duration of seizure ranges from 60 s to 120 s, and its frequency range is 1–5 Hz [39]. In this context 10-s interval includes 10–50 periods of epileptic EEG, which provides a sufficient basis for accurate prediction. One may think, that shorter intervals would result in more precise prediction. While this is technically true, we should remember that CNN treats several consecutive intervals as independent instances of data, and this way some temporal dynamics is lost. Epileptic seizures share certain frequency features with other EEG patterns, for instance sleep spindles, thus the nuanced temporal dynamics is the only feature that separates seizures from such patterns. As a result, shorter intervals of non-epileptic activity can be classified as seizures and vice versa.

Since we use CNNs in this study, it should be noted that CNNs tend to provide more rapid and stable convergence when the input data is normally distributed with close-to-zero mean and constrained variance [40]. In our case, many values of WP are close to zero, which results in close to exponential asymmetrical distribution. To address this issue, we consider the normalized logarithm of the WP as the input for the models:

$$w_n^{\log}(f, \tau) = \ln(w_n(f, \tau)), \quad (5)$$

$$w_n^{\text{norm}}(f, \tau) = \frac{w_n^{\log}(f, \tau) - \mu(w_n^{\log})}{\sigma(w_n^{\log})}, \quad (6)$$

where $\mu(\cdot)$ —mean value, $\sigma(\cdot)$ —standard deviation.

2.3 CNN-based classifier

As we mentioned earlier, the detection of epileptic seizures on the EEG recording was reduced to the classification of 10-s intervals. We proposed the initial solution for this task in a form of CNN-based classifier. In our research, we have chosen ResNet-18 architecture for CNN [41], as it is a common choice for the image classification task. Indeed, EEG wavelet spectrum shares structural features with images, so seizure detection task can be partially substituted with image classification task. This transition, however, cannot be direct, and requires some modifications to the algorithm to be made. In this context the choice of ResNet-18 is obvious: it is not the current state-of-the-art model, however, it is well-tested and dependable, which is crucial in designing custom architecture. In our case, modifications to the original architecture included accommodation of WP spectrum (25 channels \times 10 s “image”) as an input and binary prediction as an output. To achieve this, the first convolutional layer was adjusted to have 25 input channels, while the final fully connected layer responsible for prediction was modified to have a single neuron to generate the output. The resulting architecture included 18 layers and ~ 11.3 million parameters.

During CNN training, 100 examples were chosen for training each epoch, and $\sim 50\%$ of these examples contained epileptic activity. The size of the training set with 100 examples was chosen to provide reasonable time costs of training.

CNN was trained with following hyperparameters:

- number of epochs: 10,
- learning rate: 0.001,
- batch size: 4,
- optimizer: Adam.

The Adam optimizer was chosen since it is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters [42].

CNN loss function was in the form of Binary Cross Entropy (BCE):

$$\text{BCE} = -\frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} (t_i \log(p_i) + (1 - t_i) \log(1 - p_i)) \quad (7)$$

where N_{data} is the number of training samples, p_i is the model prediction, t_i is the true label.

2.4 Reducing false positives

2.4.1 Dealing with the data imbalance

Most tools for seizure detection deal with an abundance of false positive detections, which seriously hinder automated diagnostics. The major reason for that is the imbalance between seizures and non-seizures in the data. Epileptic seizures are rare events typically occurring once in a few hours or even days. This type of temporal dynamics naturally leads to the high imbalance in epileptic EEG. For example, the total duration of the dataset in the present study is over 900 h, of which only $\sim 0.5\%$ is epileptic activity. The analysis of such datasets presents a serious challenge to various classification methods including CNNs [43].

There are certain commonly accepted techniques to overcome imbalance in data, at least to some degree. First, we implemented basic but effective approach, that includes oversampling the minority class (epileptic EEG) and undersampling the majority class (normal EEG) [44]. It is achieved by manipulating the likelihood of the data segments to be selected for the training. In a signal of L segments, the probability of each segment to be selected for training is the same— F . However, when the aforementioned approach is used for an imbalanced dataset of L_n segments of normal activity, the probability of the normal segment to be selected becomes F_n , and the probability of the epileptic segment to be selected becomes F_e :

$$F = \frac{1}{L}, \quad F_n = \frac{1}{2L_n}, \quad F_e = \frac{1}{2(L - L_n)}. \quad (8)$$

Second, to further increase the robustness of the model, we used augmentation. With this approach, minor modifications are made to the dataset to artificially increase its size. Here, we utilize two augmentation techniques:

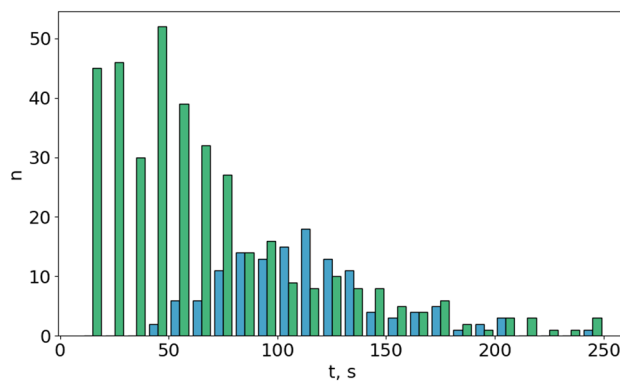
- random mirroring of data segments in temporal dimension;
- SpecAugment [45] that modifies WP spectrum by zeroing random range in temporal and/or frequency dimension.

2.4.2 Post-processing

It is worth noting that a fraction of false positives comes from the representation of CNN's predictions. Normally a binary classifier, like CNN-based one in the current study, provides prediction individually for each segment of data—10-s interval in our case. However, this is a formalistic approach that does not take into account the specifics of epilepsy diagnostics task. Even the shortest seizures are much longer than individual 10-s interval. Analysis of classification results shows that predictions tend to merge into longer sequences of the same class—"seizure" or "non-seizure". Indeed, such behavior is well aligned with observations made on epileptic EEG. CDSS by design aims to mark all seizures rather than all 10-s intervals containing epileptic activity. The latter task is far more challenging even for experienced epileptologist as the onset and offset of a seizure are often debatable. Ultimately, this kind of precision is excessive in CDSS, where the final decision is made by a human.

Additionally, pre-analysis of the results obtained with CNN-based algorithm revealed a large number of predictions with short duration. On Fig. 2, one can see the comparison of length distribution between true seizures (blue histogram) and predictions by CNN (green histogram). Blue histogram correctly reflects known features of epileptic seizures: average duration is between 60 and 120 s, while minimum duration exceeds 40 s. In contrast,

Fig. 2 Length distributions for true epileptic seizures (blue) and prediction made by CNN (green)



green histogram have a peak in the area of 10–50 s, and such short predictions do not match the definition of seizure, so we treated them as false positives.

In the context of everything discussed above, we applied two types of post-processing to the results of CNN's prediction. First, we used a median filter with a kernel size of $K = 7$ to smooth the output of the CNN. This filtering technique aims to reduce the stochasticity in predictions, leading to the elimination of sporadic short predictions.

Second, we implemented a merging algorithm that transforms sequences of 10-s predictions into “events”. Each event corresponds to a continuous interval of epileptic or non-epileptic activity with arbitrary length. The algorithm consists of two steps:

- Naive merging: neighboring segments of the same predicted class are merged together into a single longer segment of this class;
- Advanced merging: positive prediction segments that are separated by a single negative prediction segment are merged into a longer positive prediction segment.

The second step may seem far-fetched, but it originates from the features of data: epileptic seizure is a single prolonged EEG pattern, so the situation described in the second step should be treated as a classifier's inaccuracy.

2.4.3 Adjustments to evaluation procedure

Commonly evaluation of the model is done with the standard metrics based on the 1st and 2nd kind errors from statistical hypothesis testing, namely precision (P), recall (R) and F_1 -score (F_1):

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F_1 = \frac{2PR}{P + R}, \quad (9)$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative predictions, respectively.

Similar to post-processing the evaluation should consider peculiarities of the seizure detection task, which results in modifications made to definition of TP, FP, and FN predictions. The modified guidelines describing how these metrics are calculated go as follows:

- If one or more predicted events lie in a T -second vicinity of true seizure, we treat them collectively as a single TP prediction;
- Predicted events, that do not lie in a T -second vicinity of any true seizure, are treated as a FP predictions;
- If no predicted events lie in a T -second vicinity of true seizure, we treat this case as a FN prediction.

Parameter T was introduced to support the specifics of epilepsy diagnostics and CDSS performance: there is no need in precise definition of the seizure's onset and offset, as we mentioned earlier. Additionally, this more loose definition of a seizure acknowledges the fact, that expert's marking of seizures could be imprecise. This may come simply from the human factor, but there is also evidence of some underlying EEG activity prior to the seizure [46]. Such activity commonly possesses very high frequencies, which may go unnoticed for a human observer. However, CNN is a powerful technique, that can possibly consider some features of the data skipped by an expert. In our case, $T = 60$ s, which was chosen following the same reasoning about average length of epileptic seizure as in Sect. 2.4.2.

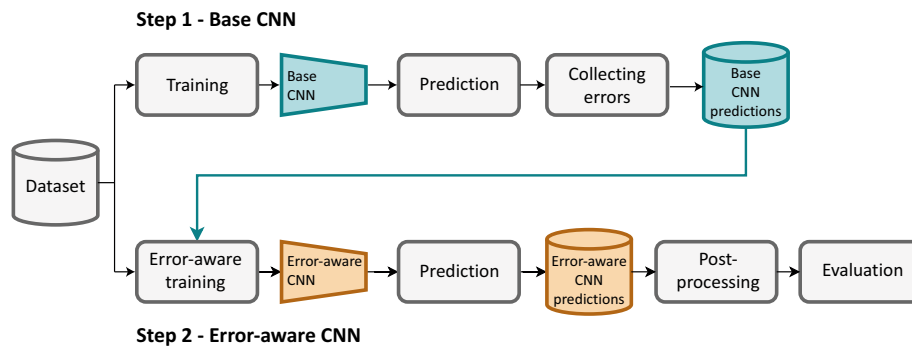
Before P , R , and F_1 metrics could be computed, raw predictions should be transformed into binary ones. It was achieved by thresholding, and the threshold was chosen to maximize precision while maintaining decent (> 0.8) recall on a validation set. This approach was imposed by the goals of CDSS, which aims to detect the most possible amount of seizures. In fact, this philosophy is the main reason for high numbers of false positive predictions.

Another thing worth noting is that traditional metrics such as P , R , and F_1 may be ill-suited to represent the efficiency of CDSS. Seizures are very rare and not evenly distributed through EEG data, and the number of false positives is high, thus the shorter recording with the same amount of seizures as some longer one would demonstrate misleading values of precision and recall. For example, in the current dataset, the recording length can differ between two patients by an order of magnitude (8 h vs 84 h). Thus, we considered the hourly equivalents of the traditional metrics— TP_h , FP_h , and FN_h :

$$TP_h = \frac{TP}{H}, FP_h = \frac{FP}{H}, FN_h = \frac{FN}{H}, \quad (10)$$

where H is the total duration of the recording in hours.

Fig. 3 The flowchart of the error-aware algorithm: the trapezoid represents CNN model, the rectangle stands for the data processing steps, the cylinder block describes the data used for processing steps, arrows illustrate data flow



It should be noted, that the adjustments described in the last two sections, do not directly reduce the number of false positive predictions. However, they are aimed to modify the definition of false positive itself with respect to the needs of medical task. This provides a fresh point of view on what should or should not be treated as a false positive, and a better understanding of what makes an efficient CDSS.

2.4.4 Error-aware algorithm

The final approach to reduce the number of false positive predictions lies in modification of CNN training procedure. We introduced the error-aware approach, which was inspired by cascade algorithms and consists of two steps. In the first step, we train the CNN model according to the procedure described earlier. The purpose of this step is to find out which of the examples are the most difficult for the CNN to classify, i.e. the examples on which the model made a false prediction. With this information, we can proceed to the second step of our method. In the second step, we train another model of the same architecture from scratch, which pays special attention to those examples where the base model made a mistake. This approach allows the model to learn on the most complex and most informative cases, which positively affects the training and overall performance. We call the model from the second step an error-aware CNN.

From the technical point of view, the primary distinction between the error-aware and initial models lies in the methodology employed to construct the training dataset and select samples for it. For the error-aware CNN model, the half of the samples were selected in the same way as for the initial model, and the other half were selected from examples where the initial model made a mistake. One may think, that usage of the error samples only could make the error-aware CNN more effective. However, we rejected this approach to avoid overfitting. Thus, for error-aware CNN model, Eq. (8) take the following form:

$$F_n = \frac{1}{4L_n}, F_a = \frac{1}{4(L - L_n)}, F_e = \frac{1}{2(L_{FP} + L_{FN})}, \tag{11}$$

where F_e —probability of error segment to be selected, L_{FP} —number of segments with false positive prediction by the initial CNN, L_{FN} —number of segments with false negative prediction.

The flowchart of the proposed error-aware algorithm is shown on Fig. 3.

3 Results and discussion

We evaluated the proposed model for the epileptic seizure detection task, based on CNN of ResNet-18 architecture that utilizes the information about the errors made by the base model of the same architecture. The evaluation results are presented in Table 1, which provides a comprehensive analysis of two approaches: Base CNN and Error-aware CNN. Additionally, the Table includes information on the effect provided by each technique employed to enhance the performance of the model. Namely, we analyzed models at four different stages of the prediction pipeline:

1. Raw predictions (raw)—Results of binary classification on 10-s segments without any post-processing;
2. Filtered predictions (filtered)—Results of binary classification on 10-s segments after applying a median filter ($K = 7$) to the raw predictions;
3. Naive merged (naive)—Results on segments of arbitrary length after naive merging of 10-s segments;
4. Advanced merged (advanced)—Results on segments of arbitrary length after advanced merging of 10-s segments.

Table 1 Results of comparative study on various techniques for reducing the number of false positives

Model	Add-on technique	Precision	Recall	<i>F</i> -score	FN	FP	TP	FN _h	FP _h	TP _h
Base	Raw	0.0638	0.7169	0.1171	167	6212	423	0.4122	15.3318	1.0440
	Filtered	0.1462	0.7339	0.2438	157	2529	433	0.3875	6.2418	1.0687
	Naive	0.1134	0.9608	0.2029	2	383	49	0.0049	0.9452	0.1209
	Advanced	0.1273	0.9608	0.2248	2	336	49	0.0049	0.8292	0.1209
Error-aware	Raw	0.2763	0.5559	0.3692	262	859	328	0.6466	2.1201	0.8095
	Filtered	0.4567	0.5271	0.4894	279	370	311	0.6886	0.9132	0.7676
	Naive	0.5000	0.8627	0.6331	7	44	44	0.0173	0.1086	0.1086
	Advanced	0.5366	0.8627	0.6617	7	38	44	0.0173	0.0938	0.1086

As one can see, here, we are only concerned of the post-processing techniques described in Sect. 2.4.2. We did not include the methods from Sect. 2.4.1 since they are essential for treating the class imbalance in the data, and it would be nearly impossible to properly train a CNN-based model without them. We also did not test the effect of evaluation adjustments from Sect. 2.4.3, so these techniques were applied to all variants of the model. Changes to evaluation procedure virtually decrease the number of false positives, but this is achieved by softening criteria for true positive detection. While this approach is justified in CDSS, it cannot be treated as a universal method for direct improvement of classifier's performance.

There are several observations that can be made from the results in Table 1. First, one can see that applying median filtration to the raw CNN predictions significantly reduces the number of FP 10-s segments. Specifically, for the Base CNN model, filtering significantly improves precision (from 0.0638 to 0.1462) and *F*-score (from 0.1171 to 0.2438), while maintaining high recall (0.7169 vs 0.7339). This can be explained by significantly reduced number of FP (from 6212 to 2529). Filtration for the Error-aware CNN results in similar behavior: it dramatically enhances precision (from 0.2763 to 0.4567) and *F*-score (from 0.3692 to 0.4894), while modestly decreasing recall (from 0.5559 to 0.5271). This result may indicate the potential loss of correctly predicted seizure onsets/offsets since the overall performance is higher in the Error-aware model than in the Base model. The number of FP predictions drops from 859 to 370.

Second, one can also clearly see the effect of the merging. Combining 10-s intervals into events of arbitrary length significantly reduces the total number of segments that need to be checked by a doctor (TP + FP). In this context, raw numbers of TP or FP are unrepresentative, but precision and recall become more relevant. Namely, for the Base CNN recall increases significantly from 0.7339 to 0.9608. The same can be said for recall in the Error-aware CNN (from 0.5271 to 0.8627). The advanced merging keeps the same high recall (0.9608 for the Base CNN and 0.8627 for the Error-aware CNN), but slightly increases precision (from 0.1134 to 0.1273 for the Base CNN and from 0.5000 to 0.5366 for the Error-aware CNN) due to further reduction in the number of FP.

Third, by comparing results for the Base CNN and Error-aware CNN, one can see that the Error-aware CNN consistently outperforms the Base CNN in terms of precision, *F*-score, and, specifically, the number of FP predictions. The only metric in which the Error-aware CNN falls behind is recall (0.9608 vs 0.8627 for the Base CNN and the Error-aware CNN with all additional techniques correspondingly). The reduced recall can be explained by the cascade nature of the Error-aware CNN, that inevitably loses some TP predictions with each iteration. However, massively increased precision for the Error-aware CNN allows it to still beat the Base CNN in terms of *F*-score (0.6617 vs 0.2248 correspondingly).

From the point of view of CDSS, we can say that the Base CNN produces a huge amount of FP predictions, which is not acceptable in the considered case. Number of segments that should be checked by an expert in CDSS is represented as TP + FP, so large number of FP predictions leads to an increased workload on human, which defies the initial goal of CDSS. On the other hand, the number of errors produced by the Error-aware CNN model is less by an order of magnitude and, in fact, is comparable to the number of true seizures. This combination of TP + FP will require much less time from a doctor to analyze, and therefore, such model is a suitable candidate for CDSS. This point is even more evident in terms of hourly metrics. For example, in the best version of the Error-aware model FP_h is even lower than TP_h (0.0938 vs 0.1086), which means that we will encounter FP predictions less frequently than TP. Given that a common EEG recording contains from 1 to 5 seizures with total duration of 5–10 min, the recording for human analysis would be 10–20 min long, which is a great improvement from 8–80 h long initial recording. In contrast, with the Base CNN, we would likely to have about 8–9 FP per single TP prediction, which is still manageable for individual cases, but becomes a problem as the number of patients in the dataset increases.

4 Conclusion

In this study, we proposed a CNN-based classifier for detecting epileptic seizures, that eventually generated a large number of false positives. We considered several adjustments to the method aimed at reducing false positives, which includes techniques for dealing with data imbalance, post-processing of the CNN's predictions and advanced approach to evaluation of classifier's performance. Ultimately, we arrived at the idea of cascade algorithm for the seizure detection task, that includes two steps. In the first step, we train a CNN model on the original dataset. Then, we collect the information about the errors of the first model and use this information to train the more precise second CNN model. We tested this algorithm and compared its effectiveness with the initial CNN approach, as well as tested the effect of all other techniques aimed at reducing the number of false positive predictions.

In conclusion, we can say the following:

- Median filtering significantly reduces FP in both Base and Error-aware CNN models;
- Merging 10-s segments into longer events simplifies the evaluation process and improves the real-world applicability of the model by reducing the number of segments to be reviewed by a doctor;
- The Error-aware CNN approach consistently outperforms the Base CNN, highlighting the benefits of error-aware modeling.

It should be noted, that this work has certain limitations and room for further research. First, the CNN architecture is not optimized. We used conventionally strong architecture ResNet-18, but we did not perform much optimization, which can become a goal for the future research. Second, the calculation time is high. This is mostly due to the use of CWT, which is computationally intensive. Thus, the future studies may consider some other similar but less demanding techniques. Third, the proposed approach lacks interpretability. While this is a common issue for CNNs, the real-world implementation of the approach requires certain transparency. This means, that in the future research we should pay extra attention to the important features used by CNN in classification.

Acknowledgements This research was funded by Academic Leadership Program PRIORITY'2030 of Immanuel Kant Baltic Federal University of the Ministry of Science and Education of Russian Federation.

Data availability The data that support the findings of this study are available from the National Medical and Surgical Center named after N. I. Pirogov of Russian Healthcare Ministry but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of the National Medical and Surgical Center named after N. I. Pirogov of Russian Healthcare Ministry.

References

1. E. Beghi, The epidemiology of epilepsy. *Neuroepidemiology* **54**(2), 185–191 (2020)
2. R.D. Thijs, R. Surges, T.J. O'Brien, J.W. Sander, Epilepsy in adults. *Lancet* **393**(10172), 689–701 (2019)
3. E. Perucca, T. Tomson, The pharmacological treatment of epilepsy in adults. *Lancet Neurol.* **10**(5), 446–456 (2011)
4. G. Luijtelaar, A. Lüttjohann, V.V. Makarov, V.A. Maksimenko, A.A. Koronovskii, A.E. Hramov, Methods of automated absence seizure detection, interference by stimulation, and possibilities for prediction in genetic absence models. *J. Neurosci. Methods* **260**, 144–158 (2016)
5. V.A. Maksimenko, S. Van Heukelum, V.V. Makarov, J. Kelderhuis, A. Lüttjohann, A.A. Koronovskii, A.E. Hramov, G. Van Luijtelaar, Absence seizure control by a brain computer interface. *Sci. Rep.* **7**(1), 2487 (2017)
6. V. Maksimenko, A. Lüttjohann, S. Heukelum, J. Kelderhuis, V. Makarov, A. Hramov, A. Koronovskii, G. Luijtelaar, Brain-computer interface for the epileptic seizures prediction and prevention, in *2020 8th International Winter Conference on Brain-Computer Interface (BCI)* (IEEE, 2020). p. 1–5
7. J.W. Miller, S. Hakimian, Surgical treatment of epilepsy. *Contin. Lifelong Learn. Neurol.* **19**(3), 730–742 (2013)
8. R. Cooper, J.W. Osselton, J.C. Shaw, *EEG Technology* (Butterworth-Heinemann, Oxford, 2014)
9. W.O. Tatum IV., *Handbook of EEG Interpretation* (Springer, Berlin, 2021)
10. C.E. Stafstrom, L. Carmant, Seizures and epilepsy: an overview for neuroscientists. *Cold Spring Harbor Perspect. Med.* **5**(6), 022426 (2015)
11. S. Beniczky, S. Wiebe, J. Jeppesen, W.O. Tatum, M. Brazdil, Y. Wang, S.T. Herman, P. Ryvlin, Automated seizure detection using wearable devices: a clinical practice guideline of the international league against epilepsy and the international federation of clinical neurophysiology. *Clin. Neurophysiol.* **132**(5), 1173–1184 (2021)
12. O.E. Karpov, E.N. Pitsik, S.A. Kurkin, V.A. Maksimenko, A.V. Gusev, N.N. Shusharina, A.E. Hramov, Analysis of publication activity and research trends in the field of ai medical applications: network approach. *Int. J. Environ. Res. Public Health* **20**(7), 5335 (2023)

13. Z. Chen, G. Lu, Z. Xie, W. Shang, A unified framework and method for eeg-based early epileptic seizure detection and epilepsy diagnosis. *IEEE Access* **8**, 20080–20092 (2020)
14. A. Aarabi, R. Fazel-Rezai, Y. Aghakhani, A fuzzy rule-based system for epileptic seizure detection in intracranial eeg. *Clin. Neurophysiol.* **120**(9), 1648–1657 (2009)
15. P. Vanabelle, P. De Handschutter, R. El Tahry, M. Benjelloun, M. Boukhebouze, Epileptic seizure detection using eeg signals and extreme gradient boosting. *J. Biomed. Res.* **34**(3), 228 (2020)
16. J. Yuan, X. Ran, K. Liu, C. Yao, Y. Yao, H. Wu, Q. Liu, Machine learning applications on neuroimaging for diagnosis and prognosis of epilepsy: a review. *J. Neurosci. Methods* **368**, 109441 (2022)
17. A. Miltiadous, K.D. Tzamourta, N. Giannakeas, M.G. Tsipouras, E. Glavas, K. Kalafataki, A.T. Tzallas, Machine learning algorithms for epilepsy detection based on published eeg databases: a systematic review. *IEEE Access* **11**, 564–594 (2022)
18. Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* **16**(5), 051001 (2019)
19. W. Zhao, W. Zhao, W. Wang, X. Jiang, X. Zhang, Y. Peng, B. Zhang, G. Zhang et al., A novel deep neural network for robust detection of seizures using eeg signals. *Comput. Math. Methods Med.* **2020**, 9689821 (2020)
20. U. Asif, S. Roy, J. Tang, S. Harrer, SeizureNet: multi-spectral deep feature learning for seizure type classification, in *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology: Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3* (Springer, 2020). p. 77–87
21. M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, B.C. Van Esesn, A.A.S. Awwal, V.K. Asari, The history began from alexnet: a comprehensive survey on deep learning approaches. arXiv preprint [arXiv:1803.01164](https://arxiv.org/abs/1803.01164) (2018)
22. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 52 (2017)
23. L. Jing, Y. Chen, Y. Tian, Coarse-to-fine semantic segmentation from image-level labels. *IEEE Trans. Image Process.* **29**, 225–236 (2019)
24. J.H. Friedman, Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001)
25. D.K.-N. Trenite, Photosensitivity and epilepsy. *Clin. Electroencephalogr.* **29**, 487–495 (2019)
26. M.D. Holmes, A.S. Dewaraja, S. Vanhatalo, Does hyperventilation elicit epileptic seizures? *Epilepsia* **45**(6), 618–620 (2004)
27. R.W. Homan, The 10–20 electrode system and cerebral location. *Am. J. EEG Technol.* **28**(4), 269–279 (1988)
28. O.E. Karpov, V.V. Grubov, V.A. Maksimenko, N. Utashev, V.E. Semerikov, D.A. Andrikov, A.E. Hramov, Noise amplification precedes extreme epileptic events on human eeg. *Phys. Rev. E* **103**(2), 022310 (2021)
29. O.E. Karpov, V.V. Grubov, V.A. Maksimenko, S.A. Kurkin, N.M. Smirnov, N.P. Utyashev, D.A. Andrikov, N.N. Shusharina, A.E. Hramov, Extreme value theory inspires explainable machine learning approach for seizure detection. *Sci. Rep.* **12**(1), 11474 (2022)
30. O.E. Karpov, M.S. Khoymov, V.A. Maksimenko, V.V. Grubov, N. Utyashev, D.A. Andrikov, S.A. Kurkin, A.E. Hramov, Evaluation of unsupervised anomaly detection techniques in labelling epileptic seizures on human eeg. *Appl. Sci.* **13**(9), 5655 (2023)
31. O.E. Karpov, S. Afinogenov, V.V. Grubov, V. Maksimenko, S. Korchagin, N. Utyashev, A.E. Hramov, Detecting epileptic seizures using machine learning and interpretable features of human eeg. *Eur. Phys. J. Spec. Top.* **232**(5), 673–682 (2023)
32. D.M. White, C.A. Van Cott, Eeg artifacts in the intensive care unit setting. *Am. J. Electroneurodiagn. Technol.* **50**(1), 8–25 (2010)
33. J. Iriarte, E. Urrestarazu, M. Valencia, M. Alegre, A. Malanda, C. Viteri, J. Artieda, Independent component analysis as a tool to eliminate artifacts in eeg: a quantitative study. *J. Clin. Neurophysiol.* **20**(4), 249–257 (2003)
34. R. Oostenveld, P. Fries, E. Maris, J.-M. Schoffelen, Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 1–9 (2011)
35. A.E. Hramov, A.A. Koronovskii, V.A. Makarov, V.A. Maksimenko, A.N. Pavlov, E. Sitnikova, *Wavelets in Neuroscience* (Springer, Berlin, 2021)
36. A. Aldroubi, M. Unser, *Wavelets in Medicine and Biology* (Routledge, London, 2017)
37. V. Grubov, E. Sitnikova, A. Pavlov, A. Koronovskii, A. Hramov, Recognizing of stereotypic patterns in epileptic eeg using empirical modes and wavelets. *Phys. A Stat. Mech. Appl.* **486**, 206–217 (2017)
38. A. Gramfort, M. Luessi, E. Larson, D.A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, Meg and eeg data analysis with mne-python. *Front. Neurosci.* **7**, 70133 (2013)
39. E. Trinka, J. Höfler, A. Zerbs, Causes of status epilepticus. *Epilepsia* **53**, 127–138 (2012)
40. L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, L. Shao, Normalization techniques in training dnns: methodology, analysis and application. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(8), 10173–10196 (2023)
41. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA* (2016), pp. 770–778
42. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
43. S. Henning, W. Beluch, A. Fraser, A. Friedrich, A survey of methods for addressing class imbalance in deep-learning based natural language processing. arXiv preprint [arXiv:2210.04675](https://arxiv.org/abs/2210.04675) (2022)

44. J. Hernandez, J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad, An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I 18* (Springer, 2013), p. 262–269
45. D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint [arXiv:1904.08779](https://arxiv.org/abs/1904.08779) (2019)
46. H. Daoud, M.A. Bayoumi, Efficient epileptic seizure prediction based on deep learning. *IEEE Trans. Biomed. Circuits Syst.* **13**(5), 804–813 (2019)

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.