Anomaly Detection on Electroencephalography with Self-supervised Learning

Junjie Xu
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, China
xujj28@mail2.sysu.edu.cn

Yaojia Zheng
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, China
zhengyj58@mail2.sysu.edu.cn

Yifan Mao
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, China
maoyf7@mail2.sysu.edu.cn

Ruixuan Wang✉
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, China
wangruix5@mail.sysu.edu.cn

Wei-Shi Zheng
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, China
wszheng@ieee.org

***Abstract*—Epilepsy is one of the most common neurological diseases in humans, and electroencephalography (EEG) is the most widely used method for clinicians to detect epileptic seizures. However, it is error-prone to detect epileptic seizures by manually observing EEG, and labeling epilepsy data is an expensive and time-consuming process. In this study, without requiring any epileptic EEG data and only based on normal EEGs, a new self-supervised learning method is proposed for anomaly detection on EEG signals. In particular, a series of scaling transformations are performed on the original EEG data to generated self-labeled scaled EEG data, where different labels correspond to different scaling transformations. Then using the self-labeled normal EEG dataset, a multi-class classifier can be trained to accurately predict the scaling transformations on new normal EEG data, but not accurately on abnormal (epileptic) EEGs. The inconsistency between the predicted scaling transformations and the ground-truth scaling transformations can then be used to measure the degree of abnormality in a new EEG data. Comprehensive experimental evaluations demonstrate that the proposed self-supervised method outperforms classic anomaly detection methods including one-class support vector machine (SVM) and autoencoders. The robustness of the proposed method also has been empirically proved with different classifier structures and by varying relevant hyper-parameters.**

***Index Terms*—Anomaly detection, Self-supervised learning, Epilepsy detection, Electroencephalography.**

I. Introduction

Epilepsy is a common neurological disorder characterised by recurring epileptic seizures [1], and it is reported that globally 50 million people are suffered from epilepsy [2]. The seizure symptoms include convulsions, loss of consciousness and disturbances in perception, sensation, mood, or other cognitive functions, depending on the regions and the extent of the affected brain. Patients with epilepsy are often more likely to suffer from mental illness, and the risk of premature death in epilepsy patients is three times higher than healthy people. Thus, it is of practical significance to improve the level of epilepsy diagnosis and treatment. As one convenient way to record brain activities, electroencephalography (EEG) has been commonly used to monitor and diagnose epileptic seizures, because epilepsy often cause abnormal brain activities [3]. However, since epilepsy is spasmodic and it is difficult to predict the recurrence of epileptic seizures in advance, clinicians are often required to observe the EEG signals of patients through the whole day, and even the most professional clinicians could make mistakes in the long-term monitoring process. Therefore, it would be much helpful if epileptic seizures can be automatically and accurately detected.

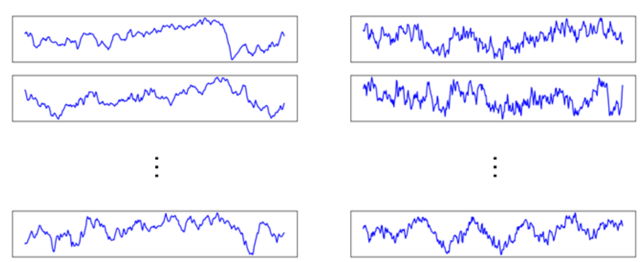

Fig. 1. Two exemplar EGG data. Each EEG consists of multiple sequences. Left: one normal EEG data; Right: one abnormal (epileptic) EEG data.

As for automatic analysis of other medical data, the techniques for automatic detection of epileptic seizure detection with EEG has also been shifted from traditional machine learning approaches to the deep learning approaches. Traditionally, the feature extraction process is manually designed by researchers, and then the hand-crafted features are used to train certain classifiers to determine whether a patient is in the state of epilepsy or not based on EEG signals [4]. Fourier transform and wavelet transform have been widely

✉ Corresponding author.

978-1-7281-6215-7/20/$31.00 @2020 IEEE

used to extract frequency features from EEG [5], and nonlinear dynamics of the EEG signals in the time domain has also been shown helpful in differentiating the epilepsy state from the healthy one [6]. As it is well-known, hand-crafted feature extraction could omit potentially discriminative features for the task of interest. In contrast, deep learning approaches can learn to features directly from the EEG data, without requiring any manual design of feature extraction. It is mainly the feature learning process that enables deep learning models to outperform traditional approaches. As the most well-developed deep learning approach, convolutional neural networks (CNNs) have shown their superior performance in epilepsy detection, either using the general multiple convolutional layers or pyramidal one-dimensional convolutional layers to extract features [7], [8]. Recently, CNN is further combined with recurrent neural networks (RNNs) for epilepsy detection [9], [10], where RNN is used to learn to extract temporal features from the EEG signals. Although deep learning approaches have shown its potential in accurate detection of epilepsy, the time-consuming labeling of EEG signals as the state of epileptic seizures and others have become one bottleneck in successfully applying deep learning approaches. In addition, patients suffering from epilepsy may show individual differences in brain activities and patients suffering from other brain diseases may also show certain abnormal brain activities, in which cases the deep learning model trained with limited amount of epilepsy EEG data may not be well generalized.

In order to alleviate the above issues, we provide a different solution to the automatic detection of epilepsy, without requiring any epileptic EEG data but just based on the healthy or normal EEG data for model training. Here we formulate epilepsy detection as an anomaly detection problem. Anomaly detection is to estimate how abnormal one data is compared to the distribution or content of normal data. In this work, inspired by the self-supervised learning for anomaly detection on image data [11], [12], we propose a new self-supervised learning method specifically for EEG data. The self-supervised deep learning model is trained based only on the normal EEG data and can help detect any potentially abnormal (including epileptic) signals in new EEG data. The main contributions are listed below.

- A new self-supervised learning method based on only normal EEG data is proposed particularly for detection of any abnormal signal in EEG data.
- A simple and effective method is proposed to generate the self-labeled data for self-supervised learning, in which different labels correspond to different scaling transformations on EEG data.
- Comprehensive experimental evaluations show that the proposed self-supervised learning based anomaly detection performs significantly better than existing well-known anomaly detection approaches, and the proposed method is robust to varying model structures and hyper-parameters settings.

II. Methodology

In this paper, we try to solve the problem of anomaly detection on EEGs, with only normal EEGs available during model training. Since only one type of data is used to train the abnormality detector, unlike the general binary classification which aims to classify any data into one of two classes, the anomaly detection often aims to learn a scoring function which can predict the degree of abnormality for any new data. Inspired by the self-supervised learning approach for abnormality detection on image data, we proposed a new self-supervised learning strategy specially for anomaly detection on EEGs. With the observation that abnormal EEG data often include wave signals of higher frequencies, we create multiple pseudo-labels based on each EEG data by resizing the normal sequence data at different scales along the time dimension, and then train a classifier to identify each of the scaling transformations on any EEG data. Since the classifier is trained based on the normal EEGs, it is expected that the classifier would be able to correctly identify the scaling transformation for any new normal data, but may incorrectly identify the scaling of abnormal EEGs due to possible differences in frequency between normal and abnormal EEGs. Therefore, the inconsistency between the predicted scalings by the classifier and the real scalings of the data could be used as the degree of the abnormality for any new EEG data.

A. *Generation of self-labeled EEG data*

In general, each EEG data consists of multiple synchronized sequences of brain activity signals (Figure 2 left), with each sequence collected by one electrode at one unique location on a brain surface over a short period of time (e.g., one second for the data used here). Although brain activity signals are originally continuous, each collected sequence data is from the regular sampling of the original continuous signal, resulting in a limited number of brain activity values.

Based on each EEG data, multiple self-labeled EEG data can be generated (Figure 2), with each pseudo-label corresponding to one specific scaling transformation of the EEG data along the time dimension. Suppose totally K scaling transformations will be applied to each EEG data, and the k-th scaling transformation T_k is associated with a unique scale s_k. For the k-th scaling transformation, given an EEG data represented by a matrix $\mathbf{X}_i$, each sequence of data (a row in $\mathbf{X}_i$) in the EEG is firstly interpolated to generate a generally longer sequence of $s_k \cdot d$ values (Figure 2 middle), where d represents the number of values in the original sequence. Then, the d values around the center (or even from the beginning) of the longer sequence are selected to form a new sequence. For each scale s_k, all the formed new sequences are finally collected to form a new scaled EEG data $T_k(\mathbf{X}_i)$ as the result of the scaling transformation T_k on $\mathbf{X}_i$ (Figure 2 right). Consequently, based on the original EEG dataset $\{\mathbf{X}_i\}_{i=1}^{N}$, a labled dataset $\{(T_k(\mathbf{X}_i), k)\}_{i,k=1}^{N,K}$ is generated based on the multiple scaling transformations, where k represents both the indicator of the k-th scaling transformation and the pseudo-label of the scaled EEG data $T_k(\mathbf{X}_i)$.

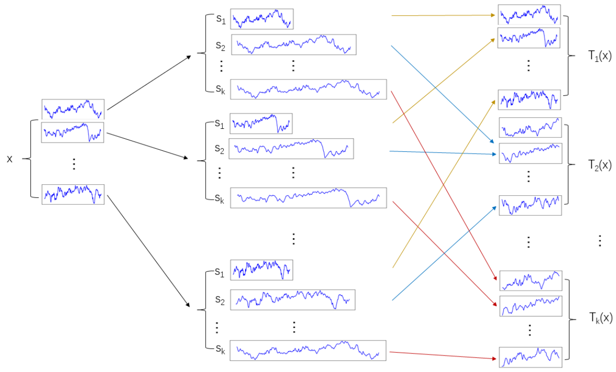

Fig. 2. Generation of self-labeld scaled EEG data. Left: one EEG data consisting of multiple sequences; Middle: each sequence is scaled (elongated) along the time dimension with different scaling transformations; Right: part of each sequence is selected (around the sequence center) and collected to form multiple new self-labeled EEG data, with different labels corresponding to different scaling transformations.

B. CNN classifier for prediction of scaling transformations

Using the self-labeled dataset $\{(T_k(\mathbf{X}_i), k)\}$ which is generated based on only the normal dataset $\{\mathbf{X}_i\}$, a deep convolutional neural network (CNN) $\mathbf{f}_{\boldsymbol{\theta}}(\cdot)$ can be trained to predict the scaling transformation for any scaled EEG data, where $\boldsymbol{\theta}$ represents the CNN parameters to be tuned during model training. Any existing CNN backbone can be used for the classifier, such as the well-known ResNet [13] or VGG-Net [14]. Similar to popular CNNs for image classification, the spatial size of convolutional kernels at each convolutional layer is set 3×3 by default. It is arguable to use kernels of size 1×3 considering there is no explicit spatial neighborhood relationship between neighboring sequences in each EEG data. However, experimental evidence suggests that kernels of size 3×3 could be helpful to capture potential implicit relationship across sequences. The final output of the CNN classifier $\mathbf{f}_{\boldsymbol{\theta}}(T_k(\mathbf{X}_i))$ consists of K values, each representing the probability of one scaling transformation. As usual, the cross-entropy loss is used to train the classifier by comparing the difference between the classifier output and the ground-truth scaling transformation (represented by one-hot vector) over all training data $\{(T_k(\mathbf{X}_i), k)\}$.

C. Anomaly detection

Once the CNN classifier is trained based on the self-labeled normal EEG data, it is expected that the scaling of any new normal EEG would be able to correctly identified by the classifier. In contrast, the higher-frequency signals in an abnormal EEG data would probably mislead the classifier to predict an incorrect scaling transformation for the scaled abnormal EEG data. Therefore, the difference or inconsistency between the predicted scalings by the classifier and the ground-truth scalings of the scaled EEG could be used to represent the degree of abnormality for the new EEG. Formally, for any new EEG data $\mathbf{X}_j$, the K scaling transformations are applied to the data to generate the K scaled EEG data $\{T_k(\mathbf{X}_j)\}_{k=1}^{K}$, and the ground-truth scaling for $T_k(\mathbf{X}_j)$ is represented by the corresponding one-hot vector $\mathbf{y}_k$ whose k-th element is 1.0 and all others are 0's. Then the inconsistency between the classifier output $\mathbf{f}_{\boldsymbol{\theta}}(T_k(\mathbf{X}_j))$ and the ground-truth scaling $\mathbf{y}_k$ for the scaled EEG $T_k(\mathbf{X}_j)$ can be calculated by certain measurement $g(\mathbf{f}_{\boldsymbol{\theta}}(T_k(\mathbf{X}_j)), \mathbf{y}_k)$, where the measurement g could be cross-entropy, L_1 distance, or L_2 distance between $\mathbf{f}_{\boldsymbol{\theta}}(T_k(\mathbf{X}_j))$ and $\mathbf{y}_k$. Over all the scaling transformations, the degree of the abnormality for the new EEG data $\mathbf{X}_j$ can be calculated by (also see Figure 3)

$$a(\mathbf{X}_j) = \frac{1}{K} \sum_{k=1}^{K} g(\mathbf{f}_{\boldsymbol{\theta}}(T_k(\mathbf{X}_j)), \mathbf{y}_k) \tag{1}$$

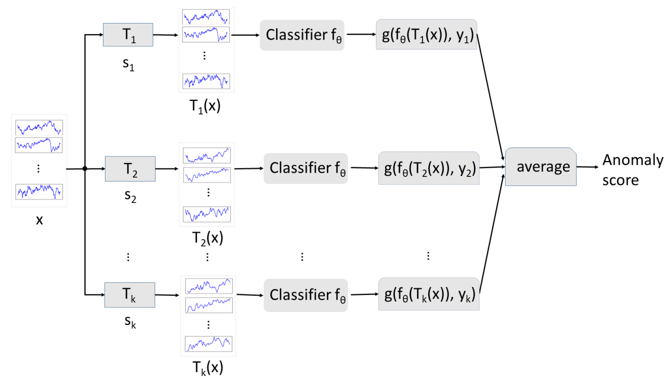

Fig. 3. Anomaly detection framework.

III. Experimental Results

A. Experimental settings

Our method is evaluated based on the UPenn and Mayo Clinic's Seizure Detection Challenge dataset [15]. The original dataset contains EEG data of 4 dogs and 8 human patients. For each subject, multiple (segments of) EEG data were recorded, with some in epileptic state and others in normal state. Since the EEG recording equipment is identical and well registered across all the 4 dogs, but varies in equipment manufacturers and the number of recorded sequences for the 8 human patients, only the EEG data of 4 dogs were used in the experimental evaluation. The EEG statistics of the 4 dogs was summarized in Table I. Each (segment of) EEG data of the dogs contains 16 sequences, and each sequence is from the regular sampling of 400 values over a period of one second by a unique electrode.

TABLE I
The number of different types of EGG data for each dog.

	Subject1	Subject2	Subject3	Subject4	Total
Normal	418	1148	4760	2790	9116
Abnormal (Epilepsy)	178	172	480	257	1087

Unless mentioned otherwise, the normal (non-epileptic) EEG data were randomly split into the training set and test set by a 4:1 ratio, and all the epileptic EEG data were used as another part of the test set. Each value in EEGs was normalized into the range $[-1, 1]$ based on the minimum and maxmum values in the training EEG dataset. By default, three scaling transformations were adopted for the self-supervised model training, with the scales $s_1 = 1.0$, $s_2 = 2.0$, and $s_3 = 3.0$. The Resnet34 was used as the default CNN backbone. Each CNN model was trained by the Adam optimizer, with the learning rate 0.00001 and the batch size 32. In all experiments, CNN training was observed convergent within 200 epochs. During testing, the cross-entropy loss was used as the measurement function g to calculate the difference between predicted scaling transformation and the ground-truth scaling. L_1 and L_2 distance led to similar results and therefore not reported. The Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC) were reported for performance evaluation.

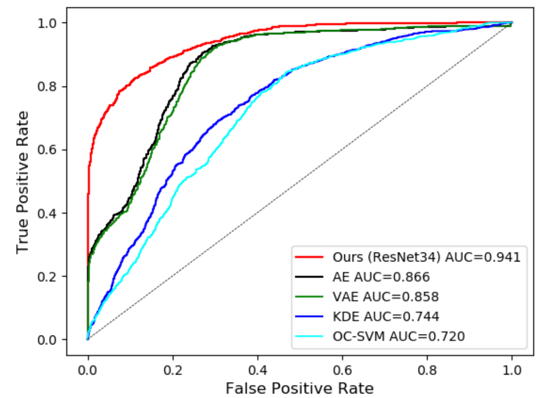

Fig. 4. Comparison of different approaches for anomaly detection.

B. Comparisons with baselines

The proposed self-supervised anomaly detection was compared with several well-known anomaly detection methods, including the one-class support vector machine (OC-SVM) [16], the statistical kernel density estimation (KDE) method, the autoencoder (AE) [17] and the variational autoencoder (VAE) [18]. OC-SVM learns a decision boundary between distribution of normal data and the origin in a higher-dimensional space, and then uses this boundary to determine how far a new data is from the normal training data. KDE naturally construct the statistical distribution of the normal data, which can then be used to estimate the probability of any new data belonging to normal data distribution. Both AE and VAE use the reconstruction error between the original input and the reconstruction to measure the degree of abnormality for the new input data. For all the baseline methods, similar efforts were taken to tune relevant hyper-parameters. In particular, for OC-SVM and KDE, the dimension of each sequence of data was reduced from 400 to 64 by principal component analysis (PCA) based on all the training set, and then the 16 dimension-reduced sequences from each EEG data were reorganized to get a vector for the model input. For AE and VAE, three convolutional layers followed by one fully connected layer were chosen for the encoder, and symmetrically one fully connected layer followed by three deconvolutional layers were chosen for the decoder. Figure 4 shows that our method clearly outperform all the baseline methods. Compared to the performance of the second best method (AE, AUC=0.866), our method improved the anomaly detection by a large margin (AUC=0.941). To more realistically simulate the epilepsy detection scenario, we also evaluate our method at the subject level, i.e., using three subjects' normal EEG data for model training and using the remaining subject's normal EEG data and all subjects epilepsy data for testing. As expected, our method still works best compared to the baselines (Figure 5). Furthermore, our method works more stably than the others, with much smaller variance in detection performance for the 4-fold cross-validation test. All the results support that the scaling transformation is effective to help train a self-supervised CNN model for anomaly detection, via which the characteristics of normal EEG data can be well implicitly learned and represented in the CNN model.

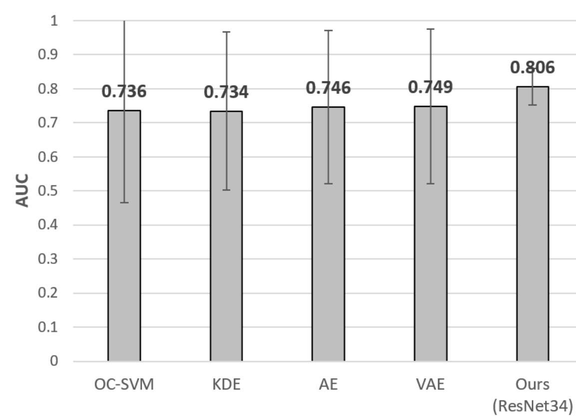

Fig. 5. Comparison of different approaches for anomaly detection at the subject level. Each vertical line represents the standard deviation of the AUCs over the 4-fold cross validations by the corresponding method.

C. Robustness of the self-supervised anomaly detection

To show the robustness of the proposed method, we tested the effect of the CNN structures, number of scaling transformations and scaling ranges, and the selected positions from scaled (elongated) EEG sequences on the performance of anomaly detection.

Classifier structures. To check whether the proposed method is robust to CNN structures, different CNN backbones were adopted for the classification of scaling transformations, including the well-known VGG19, ResNet18, ResNet34, ResNet50, and DenseNet121. As Figure 6 shows, while the performance varies a bit across these backbones, the differences in performance are relatively small and all the performances are clearly higher than that (dashed line in figure) of the strongest baseline method. This suggests that the proposed self-supervised anomaly detection works stably with varying structures of classifiers.

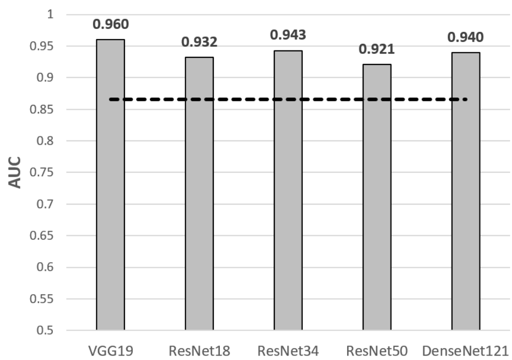

Fig. 6. Robustness to model structures. Dashed line represents the performance of the strongest baseline (AE).

Scaling numbers. With the same scaling range $[1, 3]$, different numbers of scaling transformations were respectively adopted for the self-supervised anomaly detection, i.e., by varying K from 2 to 5. In this case, the scales $\{s_k\}$ are respectively $\{1, 3\}$, $\{1, 2, 3\}$, $\{1, 1.67, 2.33, 3\}$, $\{1, 1.5, 2, 2.5, 3\}$. Figure 7 (Left) shows that the performance of the self-supervised anomaly detection method varies relatively small when changing the number of scaling transformations and all outperform the strong baseline (blue dash line), supporting that the proposed self-supervised method is robust to the number of scaling transformations.

Scaling range. With fixed number of 3 scales, five scaling ranges were tried from $[1, 2]$ to $[1, 4]$. In this case, the scales $\{s_k\}$ are respectively $\{1, 1.5, 2\}$, $\{1, 1.75, 2.5\}$, $\{1, 2, 3\}$, $\{1, 2.25, 3.5\}$, and $\{1, 2, 4\}$. Figure 7 (Right) shows that the performance of the self-supervised anomaly detection method changes little with varying ranges of scaling transformations, suggesting that the proposed method works stably within different ranges of scaling transformations.

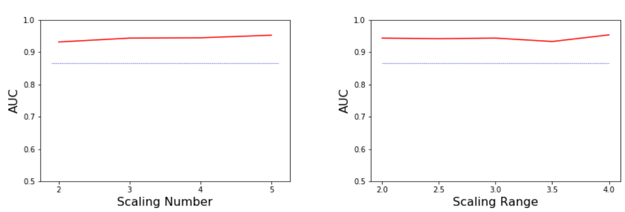

Fig. 7. Performance of the self-supervised anomaly detection with different numbers of scaling transformations (Left) and varying ranges of scales (Right). Blue dashed line represents the performance of the strongest baseline (AE).

Sampling position. When selecting part of the scaled (elongated) sequences to form new scaled EEG $T_k(\mathbf{X})$ for each scale s_k, the selection could start from the beginning position, the one-third position, or around the center of the scaled sequences etc. With such different sampling positions, the proposed self-supervised method work relatively stable (Figure 8), suggesting that different local windows in the sequences may be equivalently important for the prediction of scaling transformation and subsequent anomaly detection. The relatively small variance in performance again proves the robustness of the anomaly detection method.

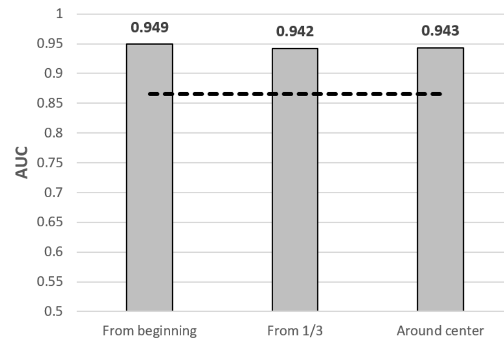

Fig. 8. Performance of the self-supervised anomaly detection with different sampling positions from scaled EEG sequences. Dashed line represents the performance of the strongest baseline (AE).

D. Ablation study

Here we evaluate the effect of kernel shape on anomaly detection. Table II shows that when replacing the shape of kernels by 1×3, the performance of the anomaly detector clearly decreased compared to the setting of 3×3 kernels. Kernels of size 3×3 may help capture potential relationship between different sequences, considering brain activities captured by these multiple sequences show strong correlations between different brain regions. This suggests that, although the two dimensions (brain location vs. time) in the EEG signals are quite different, the implicit relationships in recorded signals across the dimensions may still be well captured by the two-dimensional convolutional kernels.

TABLE II
EFFECT OF KERNEL SHAPE. 1×3 AND 3×3 REPRESENT DIFFERENT KERNEL SHAPES.

Backbone	1×3	3×3 (ours)
ResNet34	0.934	0.943
VGG19	0.947	0.960

IV. CONCLUSION

In this study, we provide a new self-supervised learning based anomaly detection specifically for epilepsy detection on EEG data. Considering the temporal property of the EEG data, scaling transformations along the time dimension are applied to EEG sequences to create self-labeled data. With the self-labeled normal data, a CNN classifier can be trained to accurately predict the scaling transformations on any new normal EEG data, but not accurately on abnormal (epileptic) data. With such inconsistency between predicted scaling transformation and the created ground-truth scaling transformation, epilepsy can be well detected. Comparisons with strong anomaly detection baselines supports the effectiveness of the proposed method. In addition, the robustness of the proposed method has also been empirically proved from various aspects.

Future work includes the evaluation of the method on human patient data and its application to the real clinical support.

ACKNOWLEDGMENT

This work is supported in part by the National Key Research and Development Program (grant No. 2018YFC1315402), the Guangdong Key Research and Development Program (grant No. 2019B020228001), the National Natural Science Foundation of China (grant No. U1811461), and the Guangzhou Science and Technology Program (grant No. 201904010260).

REFERENCES

[1] V. D. Nunes, L. Sawyer, J. Neilson, G. Sarri, and J. H. Cross, "Diagnosis and management of the epilepsies in adults and children: summary of updated nice guidance," *British Medical Journal*, vol. 344, pp. 281–287, January 2012.

[2] I. Megiddo, A. Colson, D. Chisholm, T. Dua, A. Nandi, and R. Laxminarayan, "Health and economic benefits of public financing of epilepsy treatment in india: An agent-based simulation model," *Epilepsia*, vol. 57, pp. 464–474, March 2016.

[3] S. Noachtar and J. Rémi, "The role of eeg in epilepsy: A critical review," *Epilepsy & Behavior*, vol. 15, pp. 22–33, May 2009.

[4] F. Abd El-Samie, T. Alotaiby, S. Alshebeili, T. Alshawi, and I. Ahmad, "EEG seizure detection and prediction algorithms: A survey," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, pp. 183–203, December 2014.

[5] G. Fiscon, E. Weitschek, A. Cialini, G. Felici, P. Bertolazzi, S. De Salvo, A. Bramanti, P. Bramanti, and M. C. De Cola, "Combining eeg signal processing with supervised methods for alzheimer's patients classification," *BMC Medical Informatics and Decision Making*, pp. 35–44, May 2018.

[6] Y. Zhou, R. Huang, Z. Chen, X. Chang, J. Chen, and L. Xie, "Application of approximate entropy on dynamic characteristics of epileptic absence seizure," *Neural Regeneration Research*, vol. 7, pp. 572–577, March 2012.

[7] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Computers in Biology & Medicine*, vol. 100, pp. 270–278, September 2018.

[8] I. Ullah, M. Hussain, E.-u.-H. Qazi, and H. Aboalsamh, "An automated system for epilepsy detection using EEG brain signals based on deep learning approach," *Expert Systems with Applications*, vol. 107, pp. 61–71, October 2018.

[9] M. C. Tjepkema-Cloostermans, R. C. de Carvalho, and M. J. van Putten, "Deep learning for detection of epileptiform discharges from scalp EEG recordings," *Clinical Neurophysiology*, vol. 129, pp. 2191–2196, May 2018.

[10] C. Huang, W. Chen, and G. Cao, "Automatic epileptic seizure detection via attention-based cnn-birnn," in *Proceedings of the 13th IEEE International Conference on Bioinformatics and Biomedicine*, 2019, pp. 660–663.

[11] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 9781–9791.

[12] L. Bergman and Y. Hoshen, "Classification-based anomaly detection for general data," *arXiv:2005.02359*, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[15] A. Temko, A. Sarkar, and G. Lightbody, "Detection of seizures in intracranial EEG: Upenn and mayo clinic's seizure detection challenge," in *Proceedings of IEEE International Engineering in Medicine and Biology Conference*, 2015, pp. 6582–6585.

[16] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Proceedings of the 12th International Conference on Neural Information Processing*, 2000, pp. 582–588.

[17] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, pp. 233–243, February 1991.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2013.