Healthier Lives, Digitally Enabled
M. Merolli et al. (Eds.)

doi:10.3233/SHTI210012

Automated Inter-Ictal Epileptiform Discharge Detection from Routine EEG

Duong NHU [a,1], Mubeen JANMOHAMED [b,d], Lubna SHAKHATREH [b,d], Ofer GONEN [b,d], Patrick KWAN [b,d], Amanda GILLIGAN [c], Chang WEI TAN [a] and Levin KUHLMANN [a]

[a] *Faculty of Information Technology, Monash University, Clayton VIC, Australia*
[b] *Epilepsy Clinic, Alfred Health Hospital, Melbourne VIC, Australia*
[c] *Neurosciences Clinical Institute, Epworth Healthcare Hospital, Melbourne VIC, Australia*
[d] *Department of Neurology, Central Clinical School. Monash University, Melbourne VIC Australia*

Abstract. Epilepsy is the most common neurological disorder. The diagnosis commonly requires manual visual electroencephalogram (EEG) analysis which is time-consuming. Deep learning has shown promising performance in detecting interictal epileptiform discharges (IED) and may improve the quality of epilepsy monitoring. However, most of the datasets in the literature are small ($n \leq 100$) and collected from single clinical centre, limiting the generalization across different devices and settings. To better automate IED detection, we cross-evaluated a Resnet architecture on 2 sets of routine EEG recordings from patients with idiopathic generalized epilepsy collected at the Alfred Health Hospital and Royal Melbourne Hospital (RMH). We split these EEG recordings into 2s windows with or without IED and evaluated different model variants in terms of how well they classified these windows. The results from our experiment showed that the architecture generalized well across different datasets with an AUC score of 0.894 (95% CI, 0.881-0.907) when trained on Alfred's dataset and tested on RMH's dataset, and 0.857 (95% CI, 0.847-0.867) vice versa. In addition, we compared our best model variant with Persyst and observed that the model was comparable.

Keywords. Resnet, deep learning, automation, epileptiform discharges, epilepsy

1. Introduction

Epilepsy is a neurological disorder in which a patient has an enduring tendency for recurring seizures. In Australia, 3-3.5% of the population is affected by epilepsy at some time during their lives [1]. Electroencephalography (EEG) is an important tool in the diagnosis of epilepsy. Routine EEG records the voltage fluctuations resulting from neuronal post-synaptic potentials within the brain, using surface scalp electrodes. Interictal epileptiform discharges (IED) are abnormal EEG waveforms that are often sharp, standing out from the background rhythm, and are seen in patients with epilepsy. Neurologists use epileptiform transients on EEG to support the diagnosis of epilepsy. Automated IED detection algorithms have received a lot of research interest. A recent

[1] Corresponding Author, Duong Nhu, Faculty of Information Technology, Monash University, Clayton VIC, Australia; E-mail: duong.njy1@monash.edu.

review of an extensive number of machine learning methods for automated IED detection (SVM, KNN, etc.) reported sensitivity from 30% to 99% [2]. Among all the existing methods, Persyst [3], the industry-standard IED detection software developed by Persyst Corporation, is the only software with FDA approval and has been shown to have similar performance to skilled neurologists [4]. In recent years, deep learning methods have emerged as powerful computational methods, superior to human experts in various tasks [5,6]. Researchers have demonstrated that the convolutional neural network (CNN) has had promising performance in IED detection [7,8]. However, most of these works only studied a small number of patients ($n \leq 100$). The research with the largest datasets studied 1,051 IED and 8,520 non-IED EEG recordings collected at the Massachusetts General Hospital between 2012 and 2016 [9]. Furthermore, datasets in the literature were collected from single hospitals which might limit the generalizability across different devices and settings.

To address the above limitations, we performed a study of deep learning methods in automated IED detection on a large set of routine EEG recordings of patients with idiopathic generalized epilepsy (IGE), collected from 2 hospitals. As routine EEG is a clinical standard step in epilepsy diagnosis, we implemented a general architecture which was invariant to the diversity of patients. To evaluate the generalizability of the proposed architecture, we trained different model variants on a dataset from one hospital and tested it on the other hospital. We also compared the performance of our architecture with Persyst 14 on a small independent set of routine EEG recordings.

2. Methods

2.1. Objective

The objective of this study was to automate the routine EEG review specifically for generalized IED detection from routine and outpatient EEG recordings whose durations vary from 30 minutes to 1 hour. For proof of concept, we focused on patients with idiopathic generalized epilepsy. As routine EEG is an initial standard step of epilepsy diagnosis procedure, we aimed to develop general models that would be invariant to the demographics of patients, cover a variety of artefacts and waveforms, and would be most suitable for deployment. Cross-evaluation between 2 hospitals will be carried out to confirm the generalizability of the models. In addition, the architecture will be compared with Persyst 14 [3] on an independent set of EEG recordings in IED detection and abnormal and normal EEG classification. We considered an EEG recording to be abnormal if it had at least one unequivocal IED generalized discharge or fragment.

2.2. Datasets and Labelling

We collected routine EEG recordings, between 2008 and 2019, from patients with idiopathic generalized epilepsy (IGE) seen at the Alfred Health Hospital ($n = 94$) and Royal Melbourne Hospital (RMH; $n = 110$) hospitals in Melbourne, Australia. These consist of 956 and 1,518 IED, respectively. In addition, normal control recordings were obtained from these sites ($n = 98$ and 120, respectively). The demographics of patients are summarized in Figure 1. All EEG recordings were recorded with the 10-20 system and annotated by 3 board certified neurologists with accredited training in EEG reporting. We then trained the architecture on one set and tested on the other set.

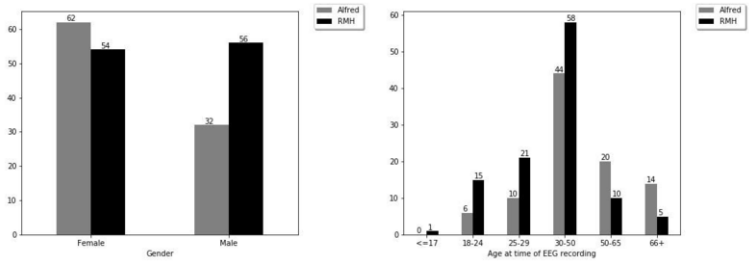

Figure 1. Demographics of patients with IGE.

To compare with Persyst 14, an independent experiment to review clinical utility was conducted with a neurologist at Alfred Health hospital. 8 EEG recordings with generalized IED and 11 normal EEG recordings were randomly selected from 2 hospitals.

2.3. Methodology

2.3.1. Preprocessing

A band-pass Butterworth filter of 0.5 - 50 Hz was used to remove muscle artefacts. We split the EEG into 2s windows of IED and normal with 50% overlap. We used 19 channels and all windows were resampled to 256 Hz. To avoid using any IED, which are missed by the neurologists, as normal windows, only normal windows from normal EEG recordings were used. All windows were z-score normalized.

2.3.2. Architecture Design

Residual network (Resnet) introduces residual connections to overcome the vanishing gradient problem when the deep learning network gets deeper [10]. Resnet has been demonstrated to be effective in image classification [10, 11], and recently in time series classification [12]. In our experiment, we considered each 2s window from an EEG recording as a multivariate time series with 512 timesteps and 19 features, and implemented the Resnet architecture from [12] (Resnet-TSC). Resnet-TSC consists of 3 residual blocks with 3 different number of filters: 64, 128, and 256.

2.3.3. Data Augmentation

As labelling IED is a resource intensive task, data augmentation is a solution to create a variance in the dataset. We implemented the data augmentation method from [7]. In each batch, a reference channel is randomly chosen. The rest of channels are ranked according to the Pearson correlation with the reference channel. This aims to make the model invariant with respect to the location of the channel and keep the local similarities in which IEDs are visible in spatially adjacent channels.

2.3.4. Tackling Imbalanced Dataset

In the collected datasets, the number of windows without IDE is significantly higher than that of windows with IDE. The ratio of IED windows to normal windows from Alfred and RMH are 1:100 and 1:30, respectively. In order to address this, we studied 3

strategies: oversampling, focal loss, and focal loss with oversampling. In terms of oversampling, the windows with IED were oversampled so that the numbers of samples in the two categories were equal. Focal loss [13] was introduced by a research team at Facebook AI Research (FAIR) and shown to be effective in objects detection where background classes significantly outnumbered foreground classes. Focal loss modifies the binary cross-entropy by adding a tuneable parameter γ and a balanced parameter α. The focal loss is defined as $FL(p)=-\alpha(1-p)^{\gamma}log(p)$. We used the same values as in the original paper for these parameters.

3. Results

3.1. Cross-evaluation Results

We trained the architecture with a batch size of 64. In addition, the cyclical learning rate in [14] was used for faster convergence. The stochastic gradient descent was used with the maximum and minimum learning rate of 0.001 and 0.0001, respectively. The step size was set to 8. Table 1 and Table 2 show the 3-folds results in which sessions were divided into 3 different groups and the results from cross-evaluation on the 2 datasets. The results from our experiment indicated that the architecture generalized well across different datasets. The focal loss strategy had the highest AUC score, 0.894 (95% CI, 0.881-0.907) when it was trained on Alfred's dataset and tested on RMH's dataset. Conversely, the focal loss with oversampling strategy had the highest AUC score, 0.857 (95% CI, 0.847-0.867) when it was trained on RMH's dataset and tested on Alfred's dataset.

To verify if the observed differences among these AUC scores are random, we applied the method of comparison of AUC by Hanley and McNeil [15] with the cut-off of 1.96 ($\alpha = 0.05$). In addition, we applied the Benjamini-Hochberg procedure [16] with the control level of 0.05 to control the false discovery rate. The results indicated that the above 2 AUC scores were the highest and significantly different from that of other model variants within each dataset ($p \leq 0.04$).

Table 1. Results of Resnet-TSC trained on Alfred's dataset.

	Trained on Alfred's dataset	**Tested on RMH's dataset**
	Mean AUC of 3 folds	AUC
Oversampling	0.936	0.884
Focal loss	0.923	0.894
Focal loss with oversampling	0.940	0.877

Table 2. Results of Resnet-TSC trained on RMH's dataset.

	Trained on RMH's dataset	**Tested on Alfred's dataset**
	Mean AUC of 3 folds	AUC
Oversampling	0.921	0.815
Focal loss	0.925	0.842
Focal loss with oversampling	0.924	0.857

3.2. Comparing with Persyst

In this experiment, we tested all model variants on the second dataset. We observed that in terms of classifying a whole EEG as normal or abnormal, the Resnet-TSC with oversampling trained on Alfred's dataset resulted in the highest sensitivity and specificity, compared to other variants, 100% and 36%, respectively. The sensitivity was 84.5% in terms of detecting 2s windows overlap with annotated IED.

Sensitivity and specificity of Persyst 14 (at moderate spike detection sensitivity setting) in EEG classification were 100% and 58%, respectively. The sensitivity of Persyst in individual IED detection was 82.7%. Overall, the results are comparable to the industry standard. Results are shown in Table 3. Moreover, we also explored the false positive samples detected by our model and observed that most of them were ocular artefacts.

Table 3. Resnet-TSC vs Persyst.

EEG classification			
Resnet-TSC		Persyst	
Sensitivity	Specificity	Sensitivity	Specificity
100%	36%	100%	58%
IED Detection			
Resnet-TSC		Persyst	
Sensitivity	Precision	Sensitivity	Precision
84.5%	27%	82.7%	37%

4. Discussions

Despite the fact that our 2 datasets are not as large as in [9], we demonstrated the Resnet-TSC with the 3 strategies of tackling imbalanced dataset generalized well across two different hospitals. In the second experiment, we collected a small newly recorded set of routine EEG data and showed that Resnet-TSC with oversampling trained on the Alfred hospital was comparable to Persyst 14. A larger sample size is needed to confirm this. In addition, over-classifying ocular artefacts as IED was found to be a limitation of the model in this experiment. This indicates an additional ocular artefact removal is needed. We will integrate this into our future work.

5. Conclusions

In this paper, we studied a Resnet architecture in automated IED detection on EEG recordings datasets from 2 hospitals. We also evaluated 3 different strategies to tackle the imbalanced dataset problem, oversampling IED samples, focal loss, focal loss with oversampling. Our models generalized well across the 2 datasets. We also compared the models with Persyst, industry-standard software for IED detection, on a separate test dataset. The model with an oversampling strategy trained on Alfred's dataset had the best performance and was comparable to Persyst. We also found that an additional ocular artefacts removal step was needed. Our future work includes improving the models and collecting another dataset from additional hospitals with the aim of providing an interictal epileptiform detector that will be reliable across multiple settings and usable in the early stages of epilepsy diagnosis involving both routine and sleep-deprived EEG.

References

[1] Epilepsy Action Australia. Help Us Fight the Stigma. Epilepsy Action Australia;. Available from: https://www.epilepsy.org.au/about-us/for-the-media/.
[2] Abd El-Samie FE, Alotaiby TN, Khalid MI, Alshebeili SA, Aldosari SA. A Review of EEG and MEG Epileptic Spike Detection Algorithms. IEEE Access. 2018;6:60673–60688.
[3] Corporate P. Persyst; Available from: https://www.persyst.com/.
[4] Scheuer M. Spike detection: Inter-reader agreement and a statistical Turing test on a large data set. Clinical Neurophysiology. 2017;128(1):243–250. Available from: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3bscp=85007206710origin=inward.
[5] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. arXiv:151200567 [cs]. 2015 Dec. ArXiv: 1512.00567. Available from: http://arxiv.org/abs/1512.00567.
[6] Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. WaveNet: A Generative Model for Raw Audio. arXiv:160903499 [cs]. 2016 Sep. ArXiv: 1609.03499. Available from: http://arxiv.org/abs/1609.03499.
[7] Hao Y, Khoo HM, von Ellenrieder N, Zazubovits N, Gotman J. DeepIED: An epileptic discharge detector for EEG-fMRI based on deep learning. NeuroImage: Clinical. 2018;17(June 2017):962–975. Available from: https://doi.org/10.1016/j.nicl.2017.12.005.
[8] Clarke S. Computer-assisted EEG diagnostic review for idiopathic generalized epilepsy. Epilepsy and Behavior. 2019. Available from: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3bscp=85074486818origin=inward.
[9] Jing J, Sun H, Kim JA, Herlopian A, Karakis I, Ng M, et al. Development of Expert-Level Automated Detection of Epileptiform Discharges During Electroencephalogram Interpretation. JAMA Neurology. 2019 Oct. Available from: https://jamanetwork.com/journals/jamaneurology/fullarticle/2752666.
[10] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv:151203385 [cs]. 2015 Dec. ArXiv: 1512.03385. Available from: http://arxiv.org/abs/1512.03385.
[11] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:160207261 [cs]. 2016 Feb. ArXiv: 1602.07261. Available from: http://arxiv.org/abs/1602.07261.
[12] Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Mining and Knowledge Discovery. 2019 Jul;33(4):917–963. ArXiv: 1809.04356. Available from: http://arxiv.org/abs/1809.04356.
[13] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. arXiv:170802002′ [cs]. 2018 Feb. ArXiv: 1708.02002. Available from: http://arxiv.org/abs/1708.02002.
[14] Smith LN. Cyclical Learning Rates for Training Neural Networks. arXiv:150601186 [cs]. 2017 Apr. ArXiv: 1506.01186. Available from: http://arxiv.org/abs/1506.01186.
[15] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983 Sep;148(3):839–843.
[16] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57(1):289–

300. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x.