Prospective Identification of Ictal Electroencephalogram

Yikai Yang, Nhan Duy Truong, Christina Maher, Armin Nikpour and Omid Kavehei*

***Abstract*—A vast majority of epileptic seizure (ictal) detection on electroencephalogram (EEG) data has been retrospective. Therefore, even though some may include many patients and extensive evaluation benchmarking, they all share a heavy reliance on labelled data. This is perhaps the most significant obstacle against the utility of seizure detection systems in clinical settings. Most retrospectively tested algorithms under-perform when presented with real-world data with different outcomes than those obtained during their training. Another critical challenge is hitting the right balance between sensitivity and false alarms. It is common to hear about tools developed for automated seizure labelling with low or extremely low sensitivities and high or too high false alarm rates. We argue that while achieving the best sensitivity may not be possible alongside the best false alarm rate, hitting the right balance is crucial. In this paper, we present a prospective automatic ictal detection and labelling performed at the level of a human expert (arbiter) and reduces labelling time by more than an order of magnitude. Accurate seizure detection and labelling are still a time-consuming and cumbersome task in epilepsy monitoring units (EMUs) and epilepsy centres, particularly in countries with limited facilities and insufficiently trained human resources. This work implements a convolutional long short-term memory (ConvLSTM) network that is pre-trained and tested on Temple University Hospital (TUH) EEG corpus. It is then deployed prospectively at the Comprehensive Epilepsy Service at the Royal Prince Alfred Hospital (RPAH) in Sydney, Australia, testing nearly $14,590$ hours of EEG data across nine years. Our system prospectively labelled RPAH epilepsy ward data and subsequently reviewed by two neurologists and three certified EEG specialists. Our clinical result shows the proposed method achieves a 92.19% detection rate for an average time of 7.62 mins per 24 hrs of recorded 18-channel EEG. A human expert usually requires about 2 hrs of reviewing and labelling per any 24 hrs of recorded EEG and is often assisted by a wide range of auxiliary data such as patient, carer, or nurse inputs. In this prospective analysis, we consider humans' role as an expert arbiter who confirms to reject each alarm raised by our system. We achieved an average of 56 false alarms per 24 hrs.**

I. Introduction

The lifetime-risk of epilepsy is globally between 3% to 4% with 1% of people globally living with active epilepsy at any time [1]. Globally, nearly 80% of epilepsy patients are living in low and middle-income countries, and 30% of epilepsy diagnosis will not respond to medication [2]. A review on *The Economic Burden of Epilepsy in Australia* by Deloitte in 2020 estimates that currently, $142,740$ people are living with active epilepsy in Australia alone, with an annual incidence rate of $14,603$ new cases across all genders, ages, and locations, which results in a total lifetime cost of $\$22.2$ billion to the economy [3]. The primary symptom of epilepsy is epileptic attacks or seizures that are unprovoked, and consequences may include injuries and even death [4]. The occurrence of seizures, missing episodes, or mischaracterizing them may result in misdiagnosis or delay diagnosis. Electroencephalography (EEG) is the gold standard for the studying, monitoring and diagnosis of epilepsy. It plays an integral role in epilepsy units around the globe. Epilepsy could have a severe impact on patients' life quality due to many factors such as unemployment, social exclusion, memory impairment and association with many psychiatric and psycho-social disorders [5], [6].

In focal epilepsy, seizures arise from a region of the cortex, "the epileptic focus", and spreads in a stereotyped, individualized fashion. Since 1950, the gold standard for localizing the onset of epileptic seizure has been the electroencephalogram (EEG) signals [7]. Localized abnormal discharges or changes in EEG frequency indicate the onset of attacks [8], [9]. While there are many EEG-based seizure detection algorithms in the literature, only a few are used in the clinical setting- where the recommendation is to have less than 1 false alarm (FA) per 24 hours (hrs) and more than 75% sensitivity [10] for non-patient specific seizure detection. [11]. There is a considerable performance drop in many studies based on retrospective data in the real-world [12]. Seizure detection and documentation have several aspects, and that makes it an exciting and popular research challenge. It could have direct benefits for patients in the form of (yet to be realized) automated, accurate and real-time seizure logging system. It also has clinical benefits such as reducing the time and cost overhead of the time-consuming and laborious long EEG review tasks.

A recent study in the US Children's Hospital of Philadelphia and University of Pennsylvania concluded that to achieve 89% identification of electrographic seizures in critically ill children, the decision-maker should be willing to pay more than $\$22,648$ per 48 hrs [13]. Additionally, EEG training to prepare expert labour requires a non-trivial amount of full-time study and dedication over six months to two years. The need for clinical care has spurred the emergence of automated non-patient specific seizure detection algorithms. Among them, deep learning methods provide more accurate and promising ideas for this problem [10], [11], [14], [15]. However, exist-

* Corresponding author.

Y. Yang, N.D. Truong, C. Maher, and O. Kavehei are with the School of Biomedical Engineering, and the Australian Research Council Training Centre for Innovative BioEngineering, Faculty of Engineering, The University of Sydney, NSW 2006, Australia.

N.D. Truong and O. Kavehei are also with The University of Sydney Nano Institute, NSW 2006, Australia.

A. Nikpour is with the Comprehensive Epilepsy Service and Department of Neurology at the Royal Prince Alfred Hospital, NSW 2050. He is also with the Faculty of Medicine and Health, central clinical school, The University of Sydney, NSW 2006, Australia.

E-mail: {yikai.yang, duy.truong, christina.maher, omid.kavehei} @sydney.edu.au, armin@sydneyneurology.com.au.

ing techniques and solutions still cannot meet the minimum requirements for clinical usage. The major bottleneck is high false-alarm if the sensitivity reaches an acceptable level, set by the clinician, who do not want to miss even one seizure or could relax that requirement a bit. Recent analyses have shown that ensuring model generalization across patient populations with different characteristics remains a challenge, necessitating label curation and model retraining to deploy machine learning models to different demographics [16], [17]. Practically, it is unrealistic for patients in the ICU to train a new model for a specific patient and apply it several days. Besides, creating a new full labelled dataset requires physician-months or physician-years of labelling time, making repeated re-labelling campaigns a substantial diversion of resources. Therefore, A generalized pre-trained automated seizure detection model across the different hospital, different gender and age patients with high sensitivity and a reasonably low false-alarm rate is an urgent unmet need for most epilepsy clinics [18]–[20].

The significance of non-patient specific prospective studies is that there will be no pre-information, no bias and no data seen before the actual evaluation and test, which, if successful, holds the potential for making a sound argument on its clinical utility. Unfortunately, implementing automated and real-time seizure logging systems that are non or minimally invasive rely on a markedly reduced spatial resolution of electrodes. There is a long road ahead of research to provide such a solution. As far as we know, there are a few commercialised prospective seizure documentation tools such as Persyst [21] and EpiScan [18], [22]. Our understanding of the two tools, as well as other research reports [15], suggest EpiScan achieves a low sensitivity of 72%, which achieves good control of the false-alarm rate. However, the sensitivity for epilepsy in adults is critical, as they statistically experience lesser seizure frequencies. Hence low sensitivity becomes a major concern. Users and literature also report conflicting reports about Persyst performance. One consistent complaint is about its high number of false alarms and its time-consuming process of automated review. Therefore, it is important to hit the right balance between sensitivity and alarm rate while remaining prospective. Some published results like [23], are also not reported to be reproduced by active researchers.

In this work, we describe an innovative method using a convolutional long short-term memory network (ConvLSTM) [24] to identify seizures on the electroencephalogram prospectively. We then use a lens method to review those alarms validity further (see Fig. 1). We proposed a prospective study framework for the detection of seizure onset and offset. Human expert arbiters review our final alarms and a significant improvement in the time. Hence, the cost of EEG review and labelling is reported without any drop compared to the performance achieved by the conventional and laborious method of EEG seizure annotation and documentation. Extensive experiments are tested on the Temple University Hospital (TUH) Seizure Corpus dataset to encourage research reproducibility and the Royal Prince Alfred Hospital (RPAH) dataset across nine years and 1,142 sessions of multi-days surface EEG recordings. Our system is to achieve high accuracy for the seizure detection task and find the undetected seizures by the clinicians or EEG specialists after their first review. Besides, clinical test results show that the proposed system, on average, provides almost an order of magnitude improvement in the time required for seizure annotation and documentation compared to the commonly in use method, shown in Figs. 1(a) and (b).

TABLE I: Summary of the TUH EEG datasets

Dataset attribute	**Train**	**Dev**
Files	4597	1013
Sessions	1185	238
Patients	592	50
Files with seizures	867	280
Sessions with seizures	343	104
Patients with seizures	202	40
Number of seizures	2370	673
Background duration (hours)	705.6	154.1
Seizure duration (hours)	46.7	16.2
Total duration (hours)	752.3	170.3

II. Dataset

There are two datasets used in this work. The Temple University Hospital (TUH) Seizure Corpus [25] for training the machine learning model and providing an opportunity for the results to be reproduced, and Royal Prince Alfred Hospital (RPAH) datasets of adult epilepsy patients for our prospective clinical test. The final ground-truth for reference to the seizure onset and offset information are acquired based on visual inspection on EEG information by the EEG specialists, as shown in Fig. 1(c). The TUH dataset is the largest publicly available epilepsy database in the US that contains EEG data from $1,185$ sessions with 592 patients (202 patient with seizures) in the training dataset and 238 sessions with 50 patients (40 patient with seizures) in development dataset respectively. Using TUH will provide an avenue for research reproducibility.

To verify the proposed system's clinical utility, we test the trained model with an inference-only mode on the RPAH EEG datasets. There are ethics approved to support our clinical access to this raw data. RPA Hospital is one of Australia's major hospitals, with one of the longest, if not the longest, EEG recordings in Australia in its Comprehensive Epilepsy Services. RPA structurally and reliably maintained data of many adult epilepsy patients across Australia. In this work, we select nine years (2011-2019) data to test with nearly $14,590$ hours of EEG data, from 212 patients out of $1,142$ sessions each session, the average recording length is around 24 hours. The detailed information is showed in Fig. 2 for RPAH dataset. The total length of datasets we used for the test are 922.6 hours of TUH data and the entire $14,591.6$ hours data from RPA.

III. Method

We trained a machine learning interface using the TUH dataset and used the trained model to assess patients' susceptibility to seizure in the RPAH dataset. We use a deterministic method to refine the AI (Artificial Intelligence) results of high susceptibility. EEG specialists and clinicians then review the output through our graphical user interface (GUI). The

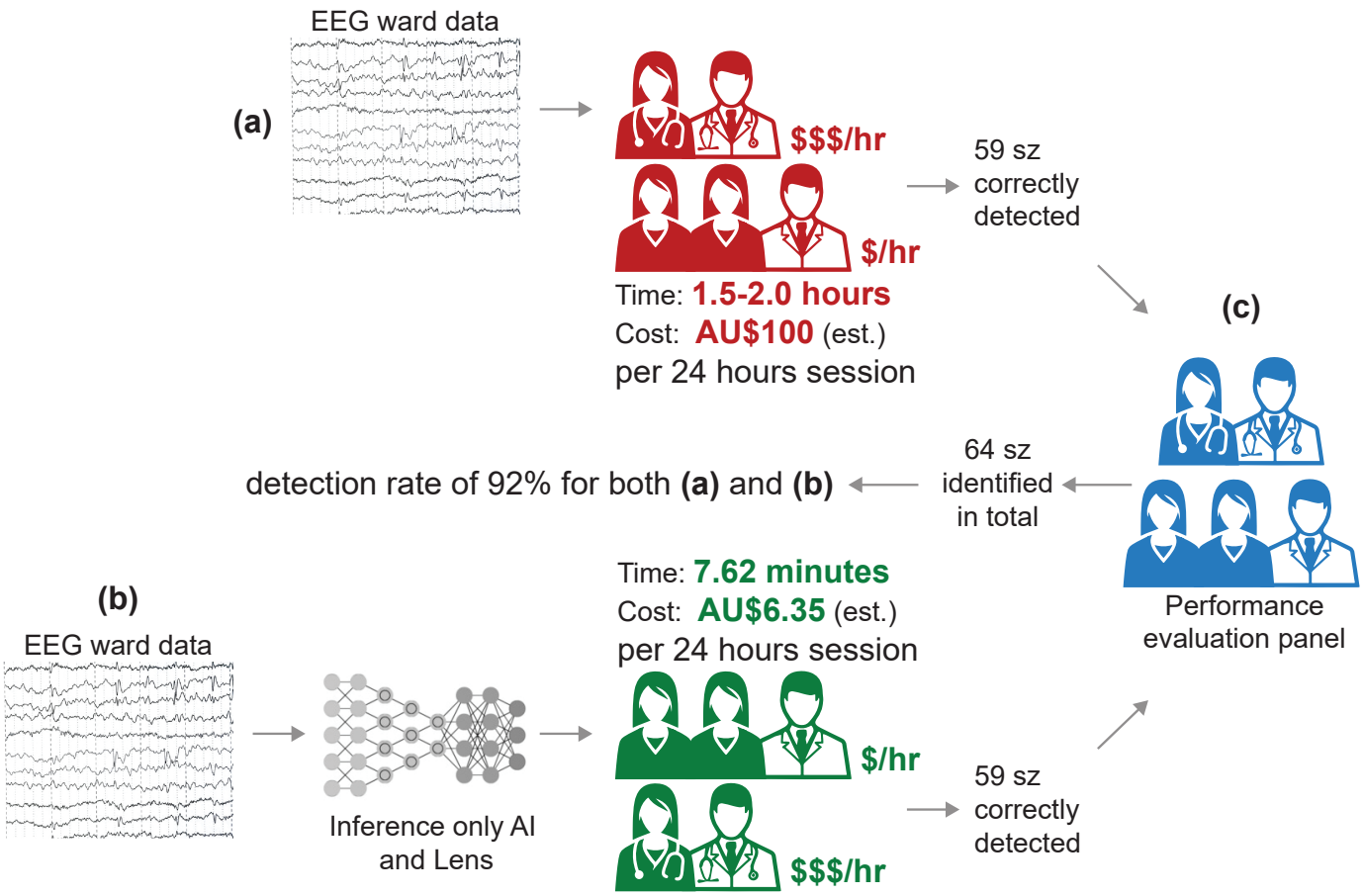

Fig. 1: Limiting the role of human experts to arbiters (seizure or no-seizure) with a prospectively and clinically evaluated system could significantly reduce the time and hence the cost associated with the time-consuming process of seizure (Sz) detection and labelling. (a) Represents a commonly used approach in which trained nurses ($/hour), EEG technicians ($/hour), and neurologists ($$$/hour) review and label EEG data. In this method, on average, each 24 hours recording session requires about 1.5 to 2.0 hours for detailed review. This time is a conservative measure, as some cases may require more or multiple reviews. (b) In contrast, we suggest a model that is accurate enough, with high sensitivity and reasonably low false alarm, to provide a series of alarms with potential onsets and offsets of seizure events to arbiters (human experts). We verified our method using multiple arbiters in the series. Clinicians (human arbiters) only review the highlighted EEG periods by our graphical user interface (GUI). (c) To validate the entire system performance against a ground-truth, a team of experts provided their full independent review of data (the ground-truth) with 64 identified seizure events. Conservatively, we assume an average of 50/hr between all expert staff.

system's outcome is a combined AI and human ability to detect the seizure, and the whole procedure is shown in Fig. 1. We did not use the RPAH dataset during training.

A. Pre-processing

Although raw EEG data information can be directly fed into a neural network, the lack of frequency information mixed with artefacts will make it harder for the network to extract essential features. To address this, we used two signal processing techniques, independent component analysis (ICA) [26] and short-time Fourier Transform (STFT) [27].

First, we split EEG signals into 12-s segments and applied the ICA algorithm to decompose the signal into several statistically independent components and removed the near-eye montages components. Then, We perform a window length of 250 (or 1s) and 50% overlapping when doing the STFT and remove the DC component so that the data shape will become $(n \times 23 \times 125)$.

B. Machine learning interface

We train our machine learning interface only use the TUH dataset, using a deep learning network that consisted of three ConvLSTM blocks [24] combined with three fully connected layers, and the detailed structure can be found in the Supplementary Methods.

Our model is implemented in Python 3.6 with the use of Keras 2.0 in the Tensorflow 1.4.0 backend. In order to avoid overfitting issue, we not only use the dropout [28] layer with 0.5 probability applied into all dense layers but also apply early-stopping technique considering the both training loss and valuation loss, which stop training when the combined loss have not increased for 20 epochs. In addition, model training

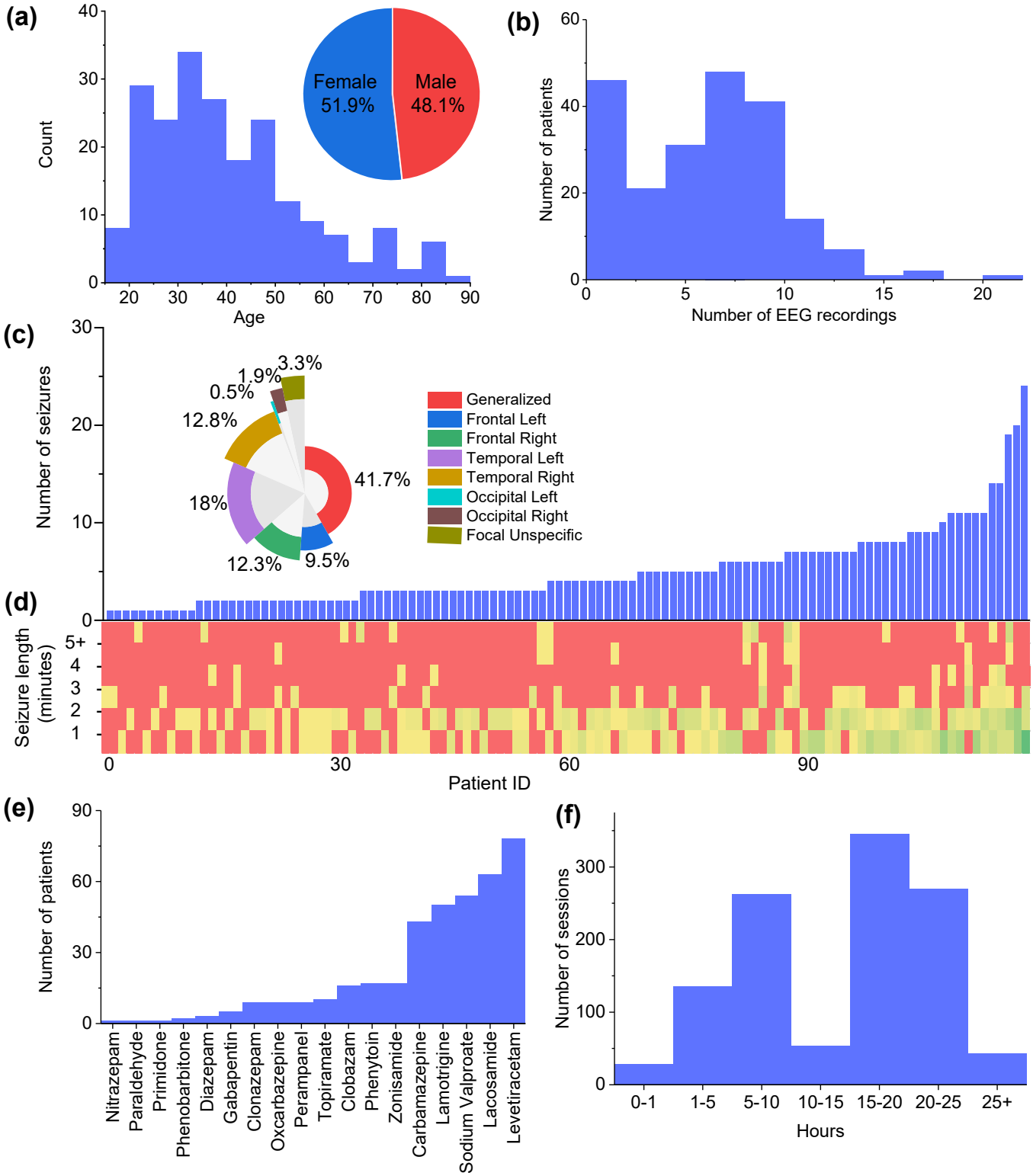

Fig. 2: Summary of the Royal Prince Alfred Hospital (RPAH) dataset. (a) Patients' age distribution and gender (inset). (b) The number of EEG recording files per patient. (c) Distribution of the number of seizures per patient (only those with detected and documented seizures are plotted, based on the final ground-truth), and their seizure types (inset). (d) Heat-map of seizure lengths for each patients with detected seizures; Changing color from green to red represents an increase in the number of seizures in that band. (e) Histogram of anti-epileptic drugs (AED) administered for patients (AED types may overlap in a given patient). (f) Monitoring sessions lengths.

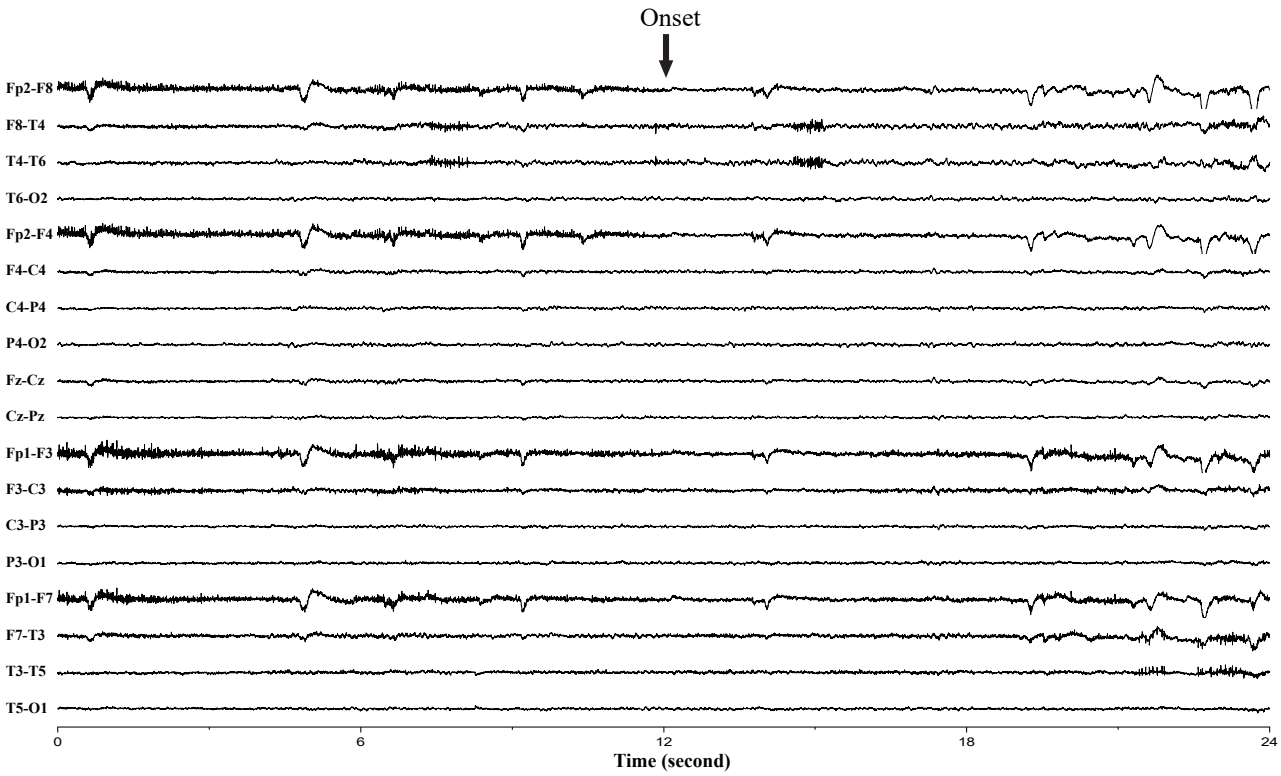

Fig. 3: Clinical test extra seizure example. This human arbiter missed this seizure at first but it was detected by AI. A review by five clinicians agreed that it was a valid seizure detection. We can see the frequency evolution are at the frontal (F8-T4, T4-T6, F3-C3) and the slowing at the temporal leads (Fp1-F3, Fp1-F7, FP2-F8, Fp2-F4) in this 24-second EEG window.

is accomplished using Adam optimizer, and the learning rate is set as 5×10^{-4}. Training the model in the TUH train set of the EEG signals took approximately one day.

C. Post Processing

After the machine learning model is trained with the TUH dataset, we used it to run the inference model directly on the RPAH without any further training and knowledge of RPAH annotations. In this process, automatically, the high risk (probability $>= 10\%$) seizure areas were selected by the model to be sent for a lens focus algorithm in charge of reviewing the results and generating alerts. The lens is a deterministic signal processing method called periodic waveform analysis (PWA), initially presented by the EpiScan tool team [29]. Periodic energy index (PEI) and PWI values for alpha, beta, theta, delta, gamma rhythms were calculated and, through an automated and adaptive process of threshold setting for the power in each band (rhythm), the lens was able to identify most likely ictal events. Details of this method are described in the Supplementary. Using this technique, we achieved a significantly higher sensitivity than EpiScan techniques. We maintained an acceptable level of false alarms based on the results collected at RPAH during our system's test and deployment.

D. Clinical test with human arbiters

After the post-processing, the left potential seizure areas are highlighted in the interface. Two board-certified practising epilepsy neurologists and three board-certified practising EEG specialists consist of our human arbiters committees. All members in our committees only visualized the high-risk area and made their decisions based on their previous clinical experience. The final results are decided by the committees' majority votes and then compared with the reference results.

E. Performance metrics

To assess how well the proposed method performs for the seizure detection task, we compute several metrics, including the Area Under the Receiver Operating Characteristic curve (AUC), sensitivity or true positive rate, false-positive rate (FPR), seizure detection rate (SDR) and false alarm per 24 hours (FA/24 hrs).

The SDR rate is calculated by the number of seizures detected over the total number of seizures. Moreover, the AUROC score is used to measure the model ability to classify the seizure and non-seizure clips regardless of the threshold selection. The value of recall verse FPR derives from the ROC curve. In the clinical test, the accuracy of seizure time and incorrect alarms are two important indicators to influence the patient and clinicians. Therefore, we use SDR and FA/24 hrs to measure how many seizures are detected correctly and how many incorrect predictions over 24 hours.

IV. Results

We test the deep neural network (DNN) on the TUH Development dataset and do a prospective study on the RPAH

dataset. For the TUH dataset test, we compare our performance with the Khaled *et al.* [15], where we only use 12-s input but still improve 6% average AUC. We also test AUC on the RPAH dataset, which reaches 0.82. For the 66 sessions clinical test, our model missed 8 seizures (after reviewing by neurologists, confirmed only 5 of them as seizures) compared with purely human identification. Still, with the combination of the proposed model and human arbiter, we find 5 more extra seizures. We compare the traditional human method with our proposed method into three aspects: accuracy, time cost, and money cost, which show in Fig. 5. We tie on the accuracy, but for each 24 hours recording, the proposed method takes more than ten times less time (7.62 mins versus 1.5 hrs).

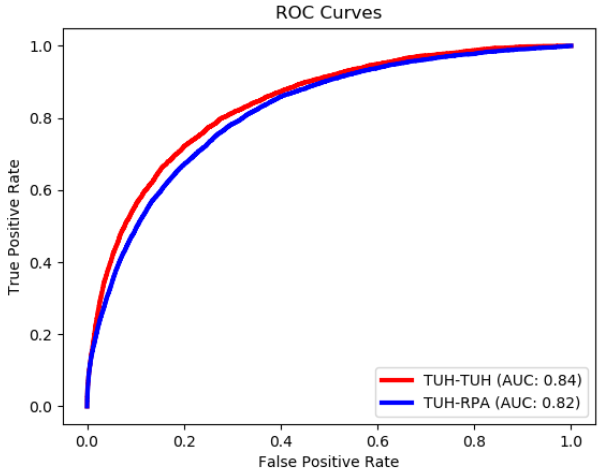

Fig. 4: Receiver operating characteristic (ROC) curves. Two curves are using the area under the ROC curve (AUC) metric. TUH-TUH represents the model is trained on the TUH training dataset and tested on the TUH development dataset. TUH-RPAH represents the model is trained on the TUH training dataset and tested on the RPAH dataset.

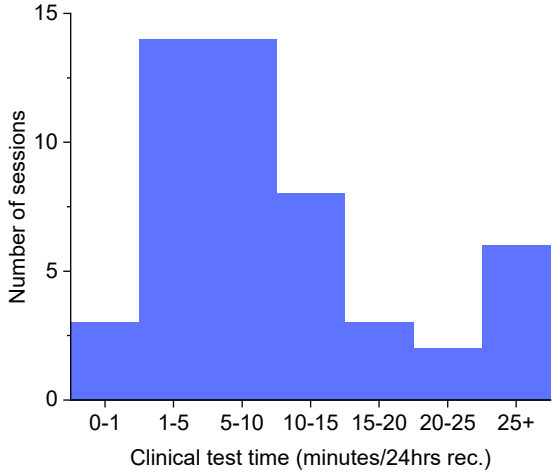

Fig. 5: Time consumption of the 66 sessions during clinical test. This is a histogram showing the actual human time spent on each clinical session. The average time cost is 7.62 mins./24 hrs recording.

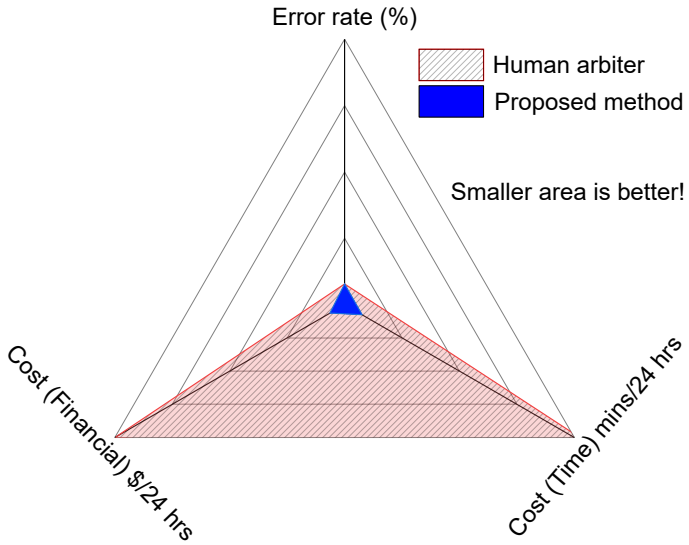

Fig. 6: Performance comparison between human arbiter and the proposed method. This is a radar curve with the money cost, time cost and error rate on three axes. The smaller triangle area means the better overall performance.

V. Discussion

We trained our deep learning model with two types of information: 1) background information - anytime except seizure happens, and 2) seizure information - seizure onset to seizure offset period. We were interested in how the deep network performs with continuous EEG recording and the practical usage in real clinical settings. Thus we did a prospective study on the RPAH dataset, trained our model on the TUH training dataset, and ran the inference on the RPAH dataset. First, we test our model on the TUH dataset and get a 0.84 AUROC score (show on Table II) using 12-s window, which is better than the performance achieved by Khaled *et al.* [15]. Although the AUC score only improve by 0.06, we use a 12-s window, which is 5 times shorter than the Khaled *et al.*. We decided this as there were a large number of seizures less than 60-s and evidence proved in Khaled *et al.* where the F1 score for using 12-s window are 0.1 and 0.27 smaller than the 60-s window on the pediatric and adult LPCH (Luckile Packard Children's Hospital) dataset respectively. To identify our model's generalization, we tested the pre-trained network directly on the RPAH dataset, which across nine years and include 1142 sessions, the AUROC score reaches 0.82, which is slightly lower than the TUH dataset. The detail curve comparison is shown in Fig. 4.

Furthermore, we did a prospective clinical test on the 66 EEG recording sessions. We asked two epilepsy neurologists and three experienced EEG specialists to label the seizure with and without the AI interface's help. The results are shown in the Fig. 1. We can see that human arbiter can find 59/64 seizures by reviewing all the EEG recordings. In comparison,

TABLE II: Results comparison

Dataset	Prospective	Method	Detect seizure length	AUC	Evaluation method	Sensitivity	FA/24 hrs
NCR+MUV+KEMP	Y	EpiScan [18]	Onset only	–	SDR*	72.00%	7.05
TUH EEG Corpus v1.1.0	N	Golmohammadi *et al.* [30]	Y	–	OVLP	39.15%	22.83
TUH EEG Corpus v1.1.0	N	Meysam *et al.* [11]	Y	–	OVLP	32.97%	73.52
TUH EEG Corpus v1.1.0	N	Meysam *et al.* [14]	Y	–	OVLP	35.35%	77.39
TUH EEG Corpus v1.4.1	N	Meysam *et al.* [10]	Y	–	OVLP	30.83%	6.75
TUH EEG Corpus v1.4.1	N	Khaled *et al.* [15]	Y	0.78	–	–	–
Stanford Hospital	Y	Khaled *et al.* [15]	Y	0.70	–	–	–
TUH EEG Corpus v1.5.1	N	Proposed AI	Y	0.84	–	–	–
RPAH	Y	Proposed AI	Y		SDR*	78.54%	56.52
RPAH	Y	Proposed AI	Y	0.82	SDR	76.68%	56.55
RPAH (66 sessions clinical test)	Y	EpiScan [18]	Onset only	–	SDR	62.50%	7.02
RPAH (66 sessions clinical test)	Y	Human arbiter	Y	–	SDR	92.19%	0
RPAH (66 sessions clinical test)	Y	Proposed AI+Human arbiter	Y	–	SDR	92.19%	0

Note: We explain the metrics AUC, SDR, FA/24 hrs in detail in Section III-E. OVLP is referred to "Any Overlap Metric" [31]. The evaluation method OVLP considers the result is correct if the detection is within the reference event or multiple shorter events detected within the long reference event. The sensitivity, FA/24 hrs, refers to the sensitivity and the number of false alarms per 24 hours and is calculated by the corresponding evaluation method. In real clinical settings, neurologists are more concerned about when the seizure happened and the frequency of seizures; thus, SDR is more suitable for the real application. The SDR* method combines the false alarms within 30 seconds into one and considers all seizure lengths are 3 minutes.
NCR: Neurological Center Rosenhuegel in Vienna
MUV: Medical University of Vienna
KEMP: Epilepsy Center Kempenhaeghe in Heeze, the Netherlands
TUH: Temple University Hospital
RPAH: Royal Prince Alfred Hospital

with AI's help, human arbiter only needs to check the area with high susceptibility, saving lots of time and move cost as in a clinical setting but still tied to the performance. General speaking, averagely, it takes $1-1.5$ hours and $1.5-2$ hours for one neurologist and one EEG specialist to review a 24 hours surface EEG recording, respectively. Interestingly, with AI's help, one human arbiter found 15 potential extra seizures that were not labelled and missed 8 seizures which the AI does not highlight. The five human arbiters confirmed 5 out of 15 are valid seizures ignored by the first time of review (without AI). One of the extra seizures is showed in Fig. 3, and neurologists found that these seizures have the common characteristic that has short subtle frequency evolution. Thus it is hard to identify when human arbiter first time reviewing the whole recordings. For the 8 seizures that AI missed, three of them were confirmed on video and not but EEG information, and video input to confirm the seizure. In another five missed seizures, neurologists found the majority of them were very brief clinical seizures. Understandably, short seizures are quite hard to detect as the EEG biomarker or patterns could be ambiguous. Another weakness for the AI is the clinical seizure, in which the patient usually reports and event. Still, there is no change in the surface EEG signals – not all seizures are associated with EEG surface EEG change. Overall, with the help of AI, the human arbiter find $59/64$ seizures, and the time and the cost comparison is shown in Fig. 6. From the figure, we can see that the proposed method save around 10 times money and time cost, while the error rate is tied. When considering the overall performance, the proposed method should be quite useful in the clinical setting, but the drawback still evident for patients and neurologists who do not want to miss even one seizure. The model needs to be further improved to increase accuracy, but the proposed framework will be useful in future clinical usage.

VI. Conclusion

Seizure prediction and detection capability has been studied and improved over the last four decades. A board-certified EEG specialist is required by law to diagnose the epilepsy, however, it takes several years to train a clinician, and the ability to generate data far exceeds the human ability to translate the data. Therefore, a reliable detection system will relieve the clinicians work as well as be helpful for patient to have a more manageable life. In the meantime, the false alarm is particular important to application as it impacts the workload for the clinicians and patients. Our proposed method show the advantage of largely reducing the time and money cost while maintain a high accuracy level, and can apply directly into clinical without acquiring the training data.

Acknowledgements

The authors would like to thank neurologists Armin Nikpour, Kaitlyn Sharp and EEG specialists Sumika Ochida, Maricar Senturias, Toh Hock Wong, IT technologist Satendra Pratap at the Comprehensive Epilepsy Service at the Royal Prince Alfred Hospital (RPAH), Sydney, Australia, for their dedication and contribution. The authors would also like to thank Sheng-Ku Lin for his logistical help during the time-consuming process of converting the EEG data. Yikai Yang would like to acknowledge the Research Training Program (RTP) support provided by the Australia Government. Omid Kavehei acknowledges the support provided by The University of Sydney through a SOAR Fellowship and Microsoft's support through a Microsoft AI for Accessibility grant.

Ethics declarations

Ethics X19-0323-2019/STE16040: Validating epileptic seizure detection, prediction and classification algorithms approved on 19 September 2019 by NSW Local Health District (LHD), for implementation in the Comprehensive Epilepsy

Services at the Department of Neurology at the Royal Prince Alfred Hospital.

COMPETING INTEREST

The authors declare no competing interests.

CODE AVAILABILITY

The code used to generate all results in this manuscript can be made available upon request.

REFERENCES

[1] P. N. Banerjee, D. Filippi, and W. A. Hauser, "The descriptive epidemiology of epilepsy—a review," *Epilepsy research*, vol. 85, no. 1, pp. 31–45, 2009.

[2] P. Kwan, S. C. Schachter, and M. J. Brodie, "Drug-resistant epilepsy," *New England Journal of Medicine*, vol. 365, no. 10, pp. 919–926, 2011.

[3] D. A. Economics, "The economic burden of epilepsy in australia, 2019–2020," *Epilepsy Australia: Sydney*, pp. 1–53, 2020.

[4] L. Ridsdale, J. Charlton, M. Ashworth, M. P. Richardson, and M. C. Gulliford, "Epilepsy mortality and risk factors for death in epilepsy: a population-based study," *Br J Gen Pract*, vol. 61, no. 586, pp. e271–e278, 2011.

[5] R. Nickel, C. E. Silvado, F. M. B. Germiniani, L. d. Paola, N. L. d. Silveira, J. R. B. d. Souza, C. Robert, A. P. Lima, and L. M. Pinto, "Quality of life issues and occupational performance of persons with epilepsy," *Arquivos de neuro-psiquiatria*, vol. 70, no. 2, pp. 140–144, 2012.

[6] R. S. Fisher, B. G. Vickrey, P. Gibson, B. Hermann, P. Penovich, A. Scherer, and S. Walker, "The impact of epilepsy from the patient's perspective i. descriptions and subjective perceptions," *Epilepsy research*, vol. 41, no. 1, pp. 39–51, 2000.

[7] B. Litt, R. Esteller, J. Echauz, M. D'Alessandro, R. Shor, T. Henry, P. Pennell, C. Epstein, R. Bakay, M. Dichter *et al.*, "Epileptic seizures may begin hours in advance of clinical onset: a report of five patients," *Neuron*, vol. 30, no. 1, pp. 51–64, 2001.

[8] M. Dichter and W. A. Spencer, "Penicillin-induced interictal discharges from the cat hippocampus. i. characteristics and topographical features." *Journal of neurophysiology*, vol. 32, no. 5, pp. 649–662, 1969.

[9] ——, "Penicillin-induced interictal discharges from the cat hippocampus. ii. mechanisms underlying origin and restriction." *Journal of Neurophysiology*, vol. 32, no. 5, pp. 663–687, 1969.

[10] J. P. Iyad Obeid, Ivan Selesnick, *Signal processing in medicine and biology*. Springer, 2020.

[11] M. Golmohammadi, S. Ziyabari, V. Shah, I. Obeid, and J. Picone, "Deep architectures for spatio-temporal modeling: Automated seizure detection in scalp eegs," *Proc. IEEE Machine Learning and Applications (ICMLA)*, pp. 745–750, 2018.

[12] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019.

[13] N. S. Abend, A. A. Topjian, and S. Williams, "How much does it cost to identify a critically ill child experiencing electrographic seizures?" *Journal of clinical neurophysiology*, vol. 32, no. 3, p. 257, 2015.

[14] M. Golmohammadi, A. H. Harati Nejad Torbati, S. Lopez de Diego, I. Obeid, and J. Picone, "Automatic analysis of eegs using big data and hybrid deep learning architectures," *Frontiers in human neuroscience*, vol. 13, p. 76, 2019.

[15] K. Saab, J. Dunnmon, C. Ré, D. Rubin, and C. Lee-Messer, "Weak supervision as an efficient approach for automated seizure detection in electroencephalography," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–12, 2020.

[16] P. Thodoroff, J. Pineau, and A. Lim, "Learning robust features using deep learning for automatic seizure detection," *Proc. Machine learning for healthcare*, pp. 178–190, 2016.

[17] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

[18] F. Fürbass, P. Ossenblok, M. Hartmann, H. Perko, A. Skupch, G. Lindinger, L. Elezi, E. Pataraia, A. Colon, C. Baumgartner *et al.*, "Prospective multi-center study of an automatic online seizure detection system for epilepsy monitoring units," *Clinical Neurophysiology*, vol. 126, no. 6, pp. 1124–1131, 2015.

[19] C. Baumgartner, J. P. Koren, and M. Rothmayer, "Automatic computer-based detection of epileptic seizures," *Frontiers in neurology*, vol. 9, p. 639, 2018.

[20] R. Hopfengärtner, B. S. Kasper, W. Graf, S. Gollwitzer, G. Kreiselmeyer, H. Stefan, and H. Hamer, "Automatic seizure detection in long-term scalp eeg using an adaptive thresholding technique: a validation study for clinical routine," *Clinical Neurophysiology*, vol. 125, no. 7, pp. 1346–1352, 2014.

[21] M. L. Scheuer, S. B. Wilson, A. Antony, G. Ghearing, A. Urban, and A. I. Bagić, "Seizure detection: Interreader agreement and detection algorithm assessments using a large dataset." *Journal of Clinical Neurophysiology*, 2020.

[22] F. Fürbass, "Eeg monitoring based on automatic detection of seizures and repetitive discharges," Ph.D. dissertation, Wien, 2018.

[23] A. Gabor, "Seizure detection using a self-organizing neural network: validation and comparison with other detection strategies," *Electroencephalography and clinical Neurophysiology*, vol. 107, no. 1, pp. 27–32, 1998.

[24] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, pp. 802–810, 2015.

[25] V. Shah, E. Von Weltin, S. Lopez, J. R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, "The temple university hospital seizure detection corpus," *Frontiers in neuroinformatics*, vol. 12, p. 83, 2018.

[26] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[27] P. Li, X. Wang, F. Li, R. Zhang, T. Ma, Y. Peng, X. Lei, Y. Tian, D. Guo, T. Liu *et al.*, "Autoregressive model in the lp norm space for eeg analysis," *Journal of neuroscience methods*, vol. 240, pp. 170–178, 2015.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] M. M. Hartmann, F. Fürbaß, H. Perko, A. Skupch, K. Lackmayer, C. Baumgartner, and T. Kluge, "Episcan: online seizure detection for epilepsy monitoring units," *Proc. IEEE Engineering in Medicine and Biology Society*, pp. 6096–6099, 2011.

[30] V. Shah, M. Golmohammadi, S. Ziyabari, E. Von Weltin, I. Obeid, and J. Picone, "Optimizing channel selection for seizure detection," *Proc. IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5, 2017.

[31] S. Ziyabari, V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, "Objective evaluation metrics for automatic classification of EEG events," *arXiv preprint arXiv:1712.10107*, 2017.

[32] I. N. Sneddon, *Fourier transforms*. Courier Corporation, 1995.

[33] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on signal processing*, vol. 45, no. 2, pp. 434–444, 1997.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Advances in neural information processing systems*, pp. 1097–1105, 2012.

[35] T. Kluge, M. Hartmann, C. Baumgartner, and H. Perko, "Automatic detection of epileptic seizures in scalp eeg-recordings based on subspace projections: 1.047," *Epilepsia*, vol. 50, pp. 26–27, 2009.

VII. Supplementary Information

A. Method

1) Pre-processing: In this work, we used the short-time Fourier transform (STFT) to translate raw EEG signals into a spectrogram with three different window lengths of 12 seconds. The Eq. 1 shows the Fourier transform that calculates the information for the frequency domain, where m represents the window length and n for the n^{th} sample. The magnitude is calculated by the square of the absolute value after the STFT [32].

$$X(m,\omega) = \sum_{n=0}^{end} x[n]\omega[n-m]e^{-j\omega n} \tag{1}$$

$$M(m,\omega) = |X(m,\omega)|^2 \tag{2}$$

The EEG data consists of recordings of different locations electrical potentials on the scalp which are presumably generated by mixed components of brain activities. The ICA [26] algorithm is applied to decompose the signal into several statistically independent topographic maps using the Blind Source Separation (BSS) approaches [33]. The Eq. 3 shows the principle of the BSS, ICA algorithms assume the matrix A contains statistically independent topographic maps and M represents for the time courses.

$$T \approx MA^T \tag{3}$$

Then we removed the components related to the EOG detected on the channel 'FP1' and 'FP2' to get the signal after removing the eye artefact.

2) Machine Learning Interface: CNN and LSTM have been two widely used methods for computer vision and natural language processing [24], [34]. In this work, we use three Conv-LSTM blocks [24] combined with three fully connected layers. The code is implemented with Keras Conv-LSTM module, and the first Conv-LSTM layer has 16 $n\times2\times3$ kernels using 1×2 stride where *n* represent the channel numbers. The next two Conv-LSTM blocks both use 1×2 stride and 1×3 kernel sizes, whereas the Conv-LSTM block 2 use 32 filters and Conv-LSTM block 3 use 64 filters. Following the three Conv-LSTM blocks are two fully connected layers with sigmoid activation and output sizes of 256 and 2, respectively. The detailed structure is showed on the Fig 7. For the clinical result, we built the interface that show the possibility that AI suggest which enable the clinician to make their decision based on the EEG they saw and record the review time automatically.

3) Post-Processing Lens: The AI output of potential seizure areas are further analyzed by the periodic waveform analysis (PWA) method [29] that is used here as a lens and on its own achieved a low score on sensitivity (see Table II). This methods is used to calculate the rhythmic EEG patterns [35], which is the most frequent patterns for the temporal lobe ellipses. First the total harmonic energy (E_τ) is calculated within a certain period of EEG signal (x_t). The PEI value is defined as the maximum value in that periodical area, which is show on the equation below:

$$E_\tau = \sum_{m>0} |\frac{1}{\sqrt{\tau}} \int_{-\infty}^{\infty} x_t \psi^*_{\frac{t}{\tau}} e^{-j2\pi\frac{mt}{\tau}} dt|^2 \tag{4}$$

$$PEI = \max_{t_{min}\leq\tau\leq t_{max}} E_\tau \tag{5}$$

Then the signal energy value is calculated in that period.

$$N_\tau = \frac{1}{\sqrt{\tau}} \int_{-\infty}^{\infty} |x_t \psi^*_{\frac{t}{\tau}}|^2 dt \tag{6}$$

Finally, the PWI value is defined as:

$$PWI = \frac{E_\tau}{N_\tau} \tag{7}$$

Then the PWI_α,PWI_β,PWI_θ,PWI_δ,PWI_γ value are split based on the different brain signal frequency range. Only the value of PWI and PEI for all brain signals are higher than the adaptive threhold (certain percenatge of 2 hours within that area PWI and PEI value), the risk seizure area will reported to the interface.

B. Results

The human arbiter (neurologists and EEG specialists) labelled $1,142$ sessions across 9 years. In contrast, 136 sessions have seizures that are hard to make decisions without the help of the video information or miss the sEEG whole electrode information, thus to better compare with the AI performance, we test $1,006$ sessions use our prospective study. The detail results are showed on the Table IV, while human check and label all $1,142$ sessions. After that, neurologists select 66 sessions use the proposed methods to do the clinical test, and the detail information are shown on the Table III.

During the 66 sessions clinical test, as mentioned the before, neurologists found 5 extra seizure with the help of the interface, and compared with the seizure that AI miss, three of them are confirmed can not directly find with only EEG information.

1) Extra seizures detected by AI: The examples are shown in Fig 9, 10, 11, 12

2) AI miss seizures: Three examples are shown in Fig 13, 14, and 15, which are confirmed only on video and not but EEG information.

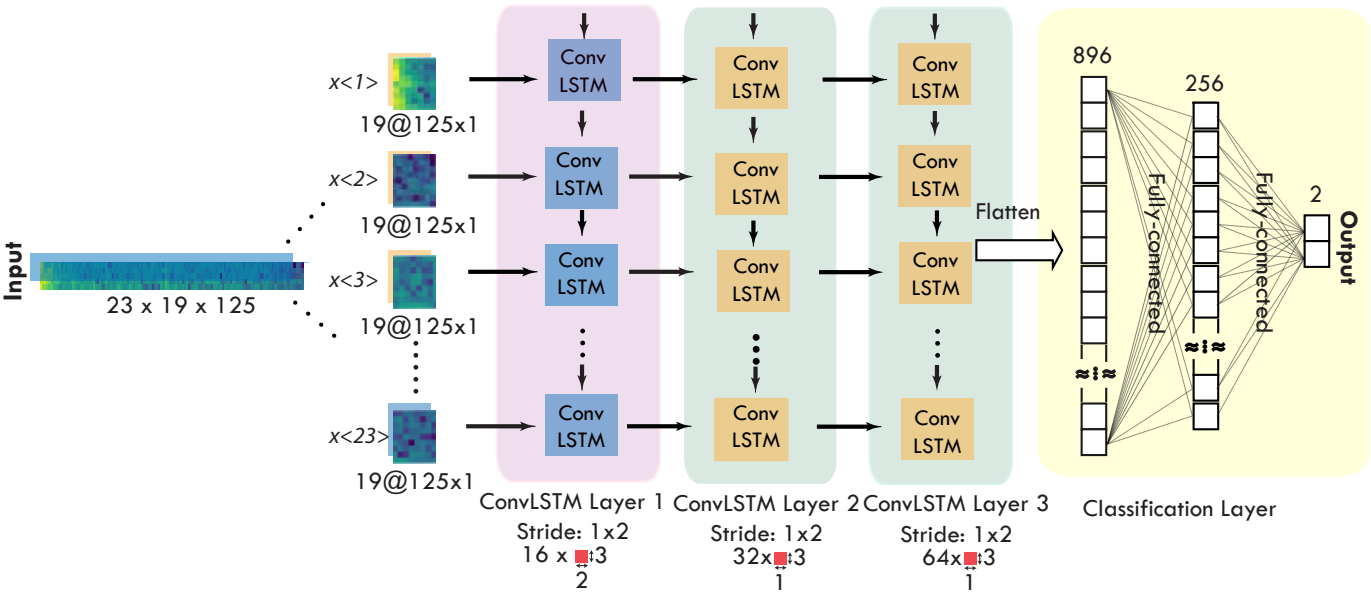

Fig. 7: Deep Learning Model

Fig. 8: Interface

TABLE III: Clinical test detail information

Session Number	No. seizures (Human arbiter)	No. seizures (AI+Human arbiter)	Clinical test time	Recording length
1	0	0	0 : 07 : 03	18 : 23 : 53
2	3	3	0 : 05 : 28	23 : 44 : 16
3	8	9	0 : 11 : 07	17 : 27 : 31
4	3	4	0 : 01 : 02	06 : 10 : 57
5	0	0	0 : 00 : 11	16 : 27 : 02
6	0	0	0 : 05 : 54	18 : 37 : 44
7	0	0	0 : 01 : 06	18 : 15 : 15
8	0	0	0 : 02 : 24	17 : 11 : 10
9	0	0	0 : 01 : 00	0 : 51 : 48
10	0	0	0 : 03 : 12	14 : 04 : 13
11	0	0	0 : 00 : 23	10 : 18 : 34
12	0	0	0 : 02 : 28	19 : 51 : 37
13	0	0	0 : 01 : 10	06 : 40 : 56
14	0	0	0 : 00 : 28	06 : 21 : 43
15	0	0	0 : 05 : 13	06 : 50 : 31
16	0	0	0 : 03 : 34	25 : 23 : 21
17	1	1	0 : 03 : 21	02 : 46 : 42
18	1	1	0 : 07 : 31	23 : 41 : 59
19	0	0	0 : 09 : 25	17 : 00 : 39
20	2	2	0 : 08 : 37	24 : 01 : 36
21	1	1	0 : 02 : 54	24 : 42 : 02
22	0	0	0 : 09 : 54	20 : 41 : 51
23	1	1	0 : 00 : 33	01 : 25 : 52
24	1	1	0 : 11 : 09	30 : 16 : 09
25	0	0	0 : 00 : 25	02 : 24 : 13
26	0	0	0 : 03 : 35	20 : 31 : 00
27	0	0	0 : 04 : 30	16 : 19 : 59
28	0	0	0 : 00 : 29	01 : 24 : 35
29	0	0	0 : 00 : 03	25 : 49 : 53
30	0	0	0 : 00 : 31	01 : 19 : 17
31	0	0	0 : 00 : 53	03 : 10 : 56
32	0	0	0 : 07 : 41	24 : 12 : 53
33	1	1	0 : 10 : 58	04 : 55 : 00
34	0	0	0 : 12 : 20	23 : 47 : 51
35	2	3	0 : 02 : 37	21 : 26 : 48
36	0	0	0 : 09 : 08	21 : 39 : 24
37	0	0	0 : 02 : 24	20 : 18 : 14
38	0	0	0 : 04 : 44	09 : 00 : 22
39	0	0	0 : 04 : 41	23 : 52 : 12
40	0	0	0 : 04 : 47	14 : 53 : 24
41	0	0	0 : 00 : 10	00 : 25 : 32
42	0	0	0 : 05 : 23	06 : 26 : 45
43	3	3	0 : 05 : 24	17 : 00 : 31
44	0	0	0 : 00 : 46	02 : 49 : 18
45	0	0	0 : 00 : 09	01 : 19 : 10
46	0	0	0 : 02 : 18	01 : 33 : 55
47	1	1	0 : 13 : 09	23 : 51 : 00
48	1	1	0 : 11 : 15	18 : 44 : 52
49	0	0	0 : 06 : 33	23 : 01 : 36
50	0	0	0 : 02 : 02	06 : 23 : 43
51	0	0	0 : 00 : 46	00 : 29 : 46
52	1	1	0 : 02 : 13	00 : 56 : 00
53	2	2	0 : 08 : 08	08 : 04 : 06
54	17	15	0 : 11 : 17	24 : 16 : 04
55	3	2	0 : 11 : 37	15 : 10 : 56
56	0	0	0 : 02 : 15	06 : 03 : 09
57	0	0	0 : 03 : 47	16 : 13 : 56
58	0	0	0 : 01 : 05	05 : 42 : 41
59	2	2	0 : 04 : 29	14 : 37 : 38
60	4	4	0 : 07 : 54	15 : 27 : 53
61	1	1	0 : 00 : 46	08 : 27 : 35
62	0	0	0 : 02 : 46	16 : 13 : 24
63	0	0	0 : 00 : 54	16 : 13 : 55
64	0	0	0 : 00 : 53	07 : 40 : 35
65	0	0	0 : 00 : 18	06 : 58 : 55
66	0	0	0 : 01 : 04	18 : 38 : 22
Overall	59	59	4 : 42 : 14	889 : 14 : 39

TABLE IV: Detailed Results of RPAH Comparison

Year	AUC	SDR	FA/24 hrs	SDR*	FA/24 hrs*	Total duration (hours)	Number of sessions
2011	0.8993	82.45%	58.00	85.96%	57.96	1114.69	75
2012	0.9107	83.52%	56.77	83.52%	56.76	1752.62	117
2013	0.896	83.33%	47.11	86.67%	47.09	2090.99	118
2014	0.7382	73.02%	74.86	74.60%	74.84	1792.13	111
2015	0.8215	78.69%	64.78	78.69%	64.74	2075.35	139
2016	0.8547	80.25%	44.75	80.25%	44.72	1506.03	101
2017	0.6827	67.65%	53.41	70.58%	53.36	2171.21	174
2018	0.7286	58.49%	50.17	66.04%	50.07	1181.05	100
2019	0.7877	75.00%	56.54	75.00%	56.48	907.60	71
Overall	0.8172	76.68%	56.55	78.54%	56.52	14591.6	1006

Note: The metric AUC, SDR, FA/24 hrs are explained in detail in the Section III-E.
SDR* and FA/24 hrs* method combines all false alarms within 30 seconds as one and considers all seizure lengths are 3 minutes.

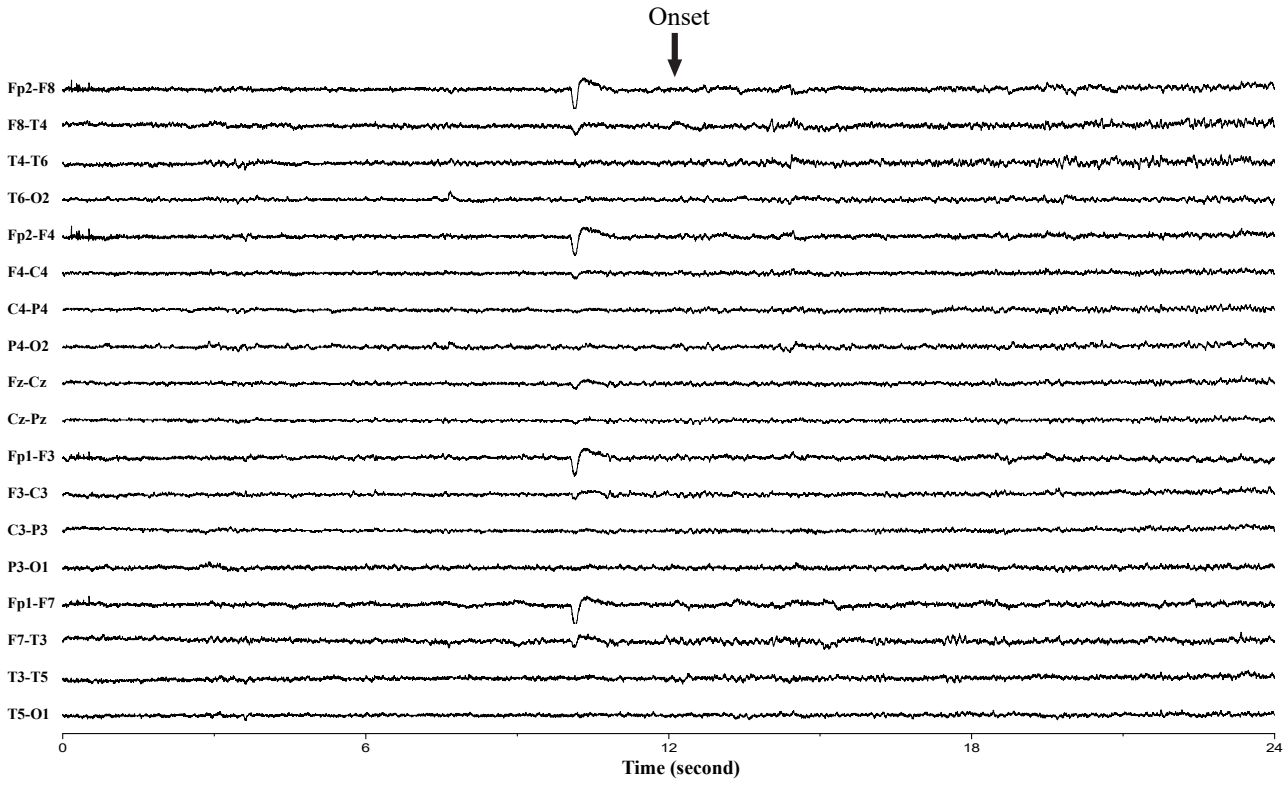

Fig. 9: Extra seizure detected by AI (verified by the neurologist).

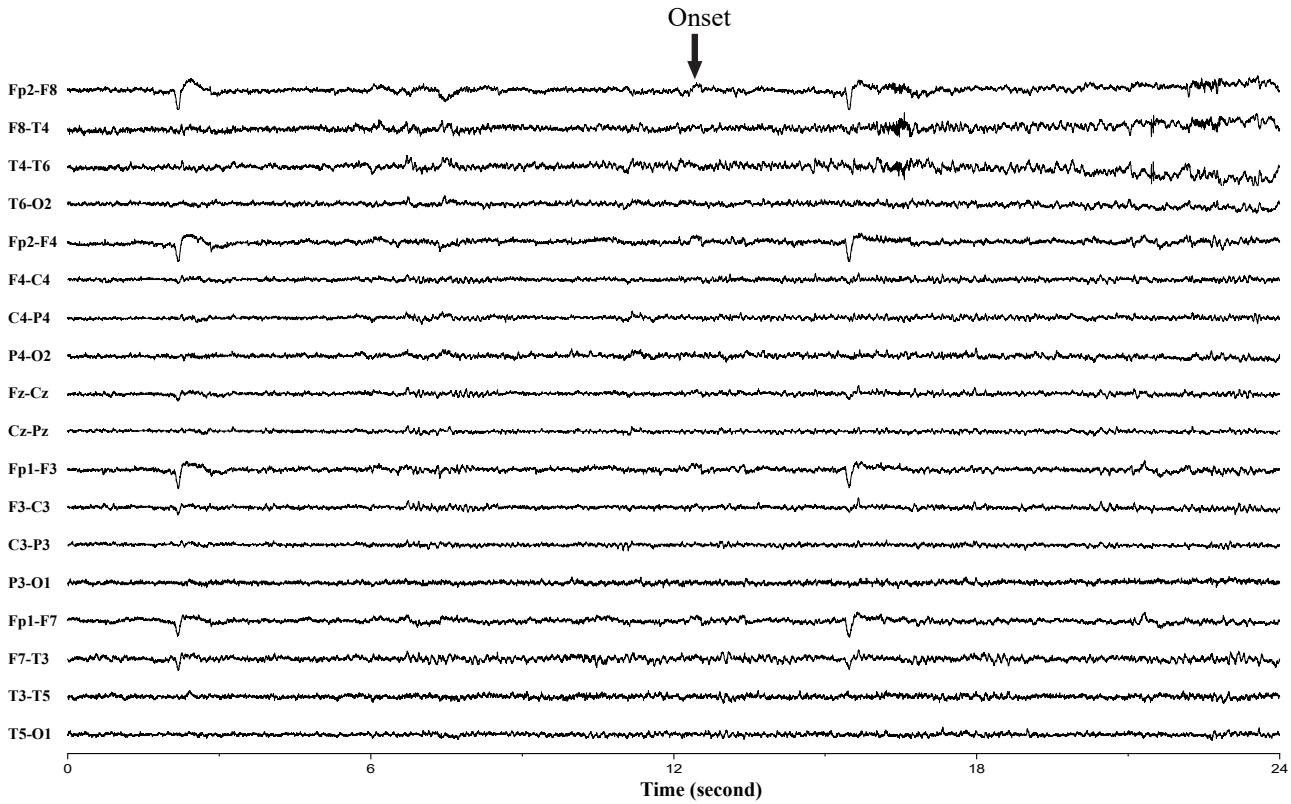

Fig. 10: Extra seizure detected by AI (verified by the neurologist).

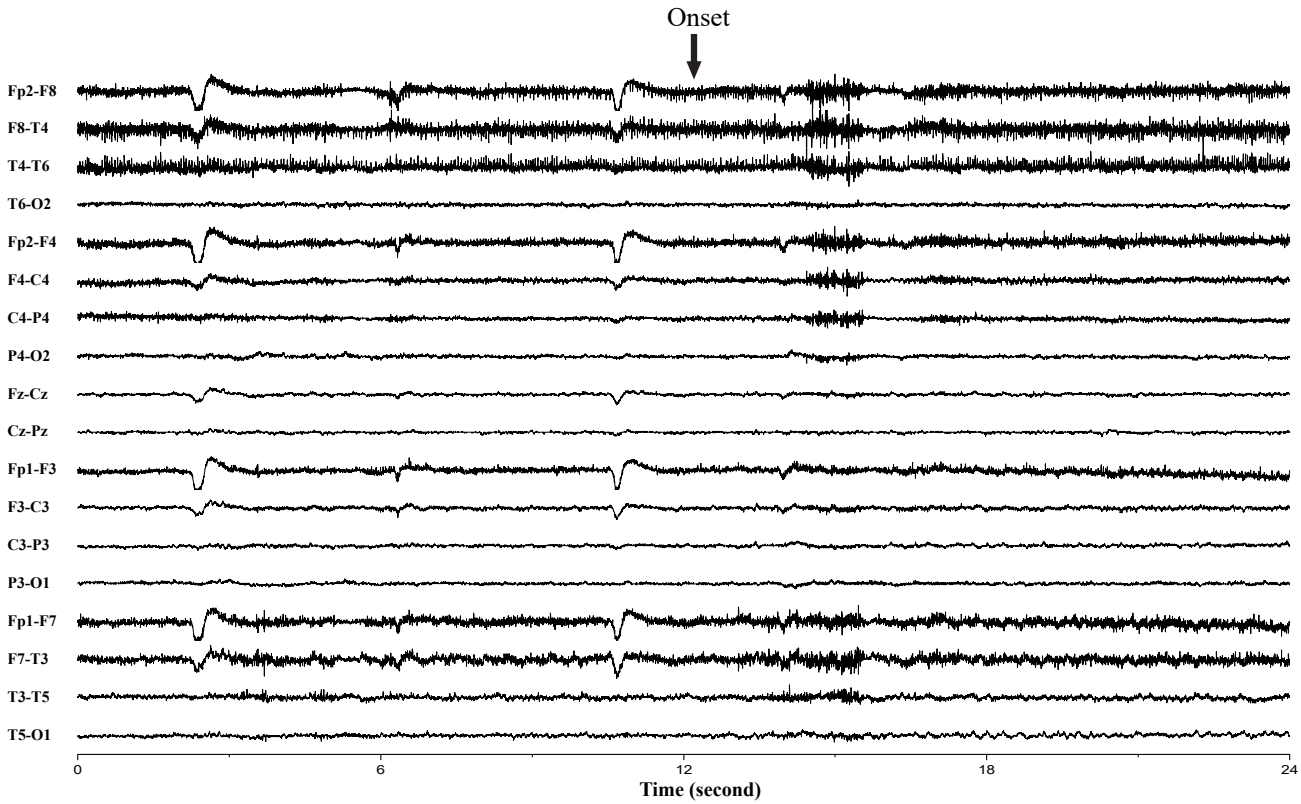

Fig. 11: Extra seizure detected by AI (verified by the neurologist).

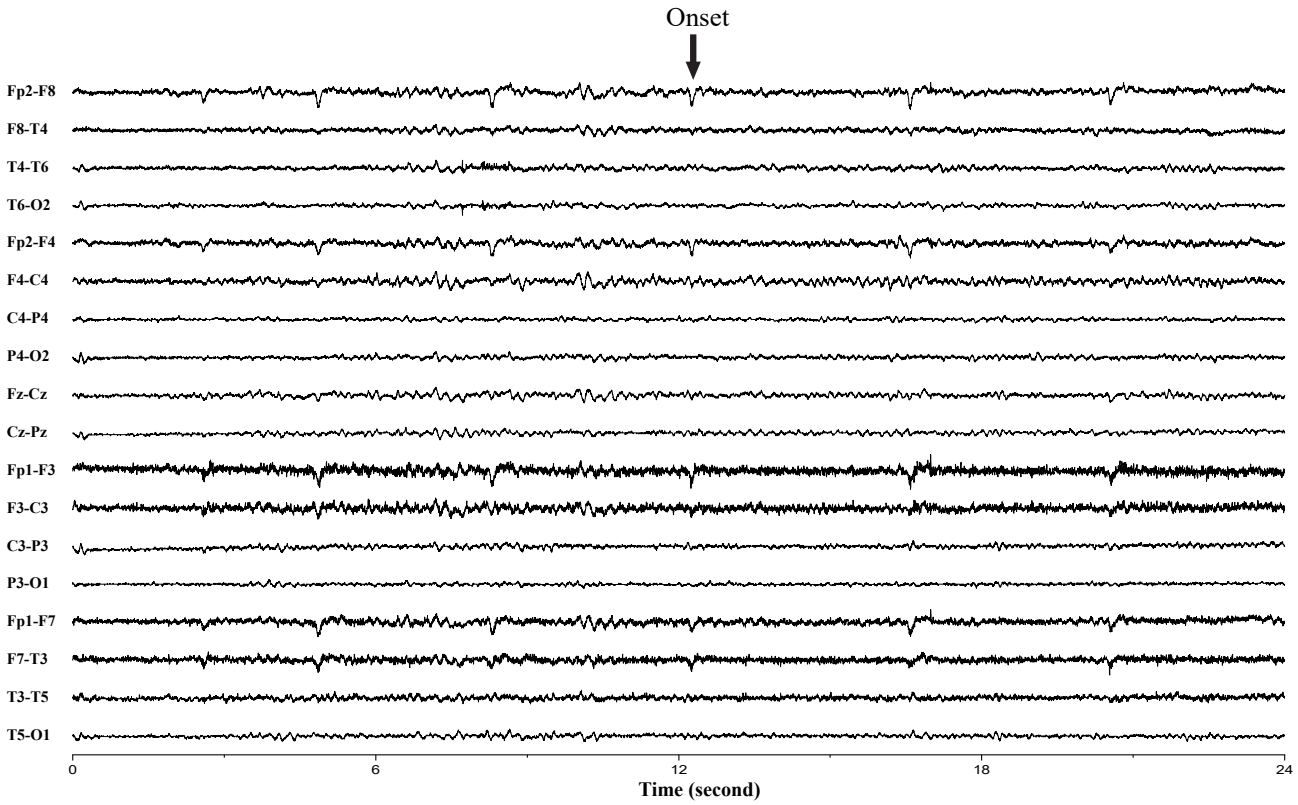

Fig. 12: Extra seizure detected by AI (verified by the neurologist).

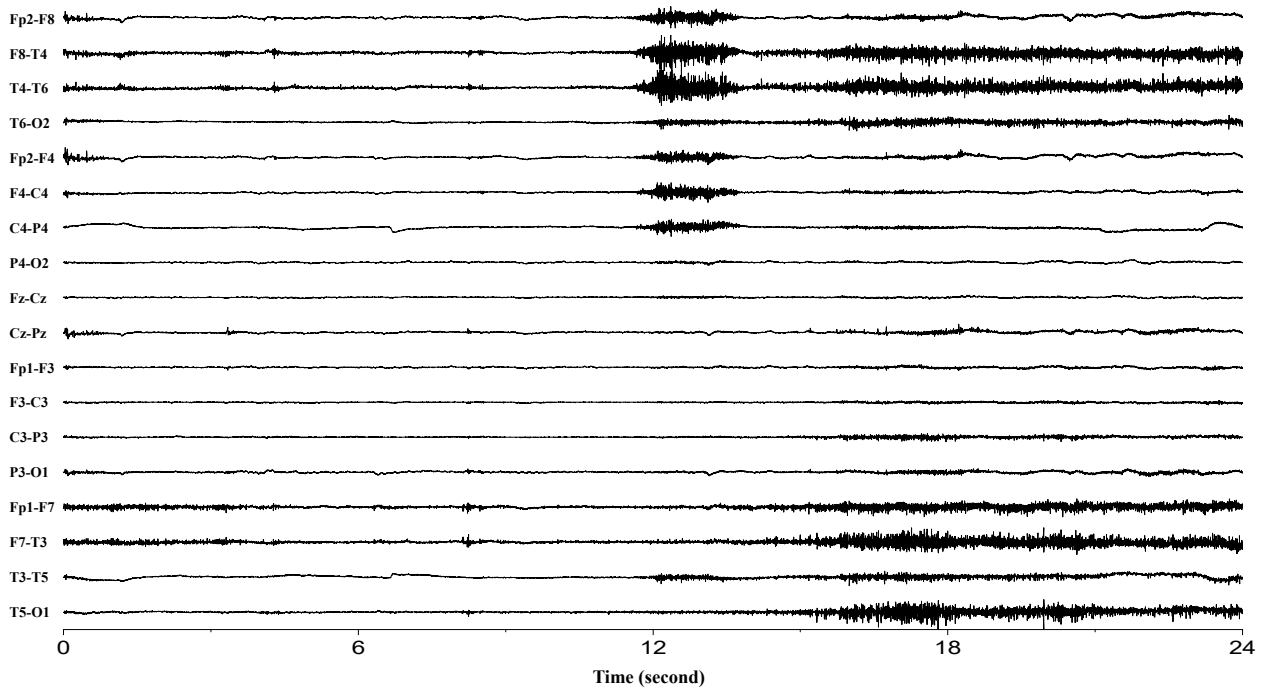

Fig. 13: Seizure undetected by AI. There is no clear frequency evolution and there are lots of muscle and eye artifacts which can only be confirmed by the video (verified by the neurologist).

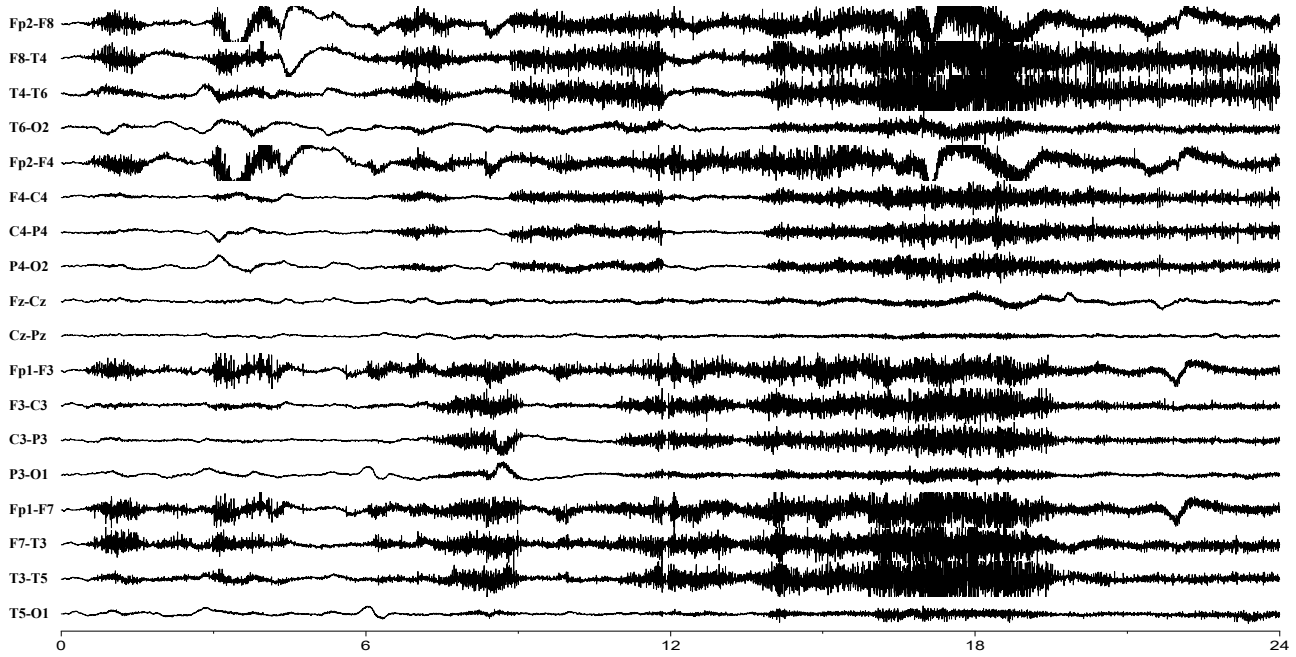

Fig. 14: Seizure undetected by AI. There is no clear frequency evolution and there are lots of muscle artifacts which can only be confirmed by the video (verified by the neurologist).

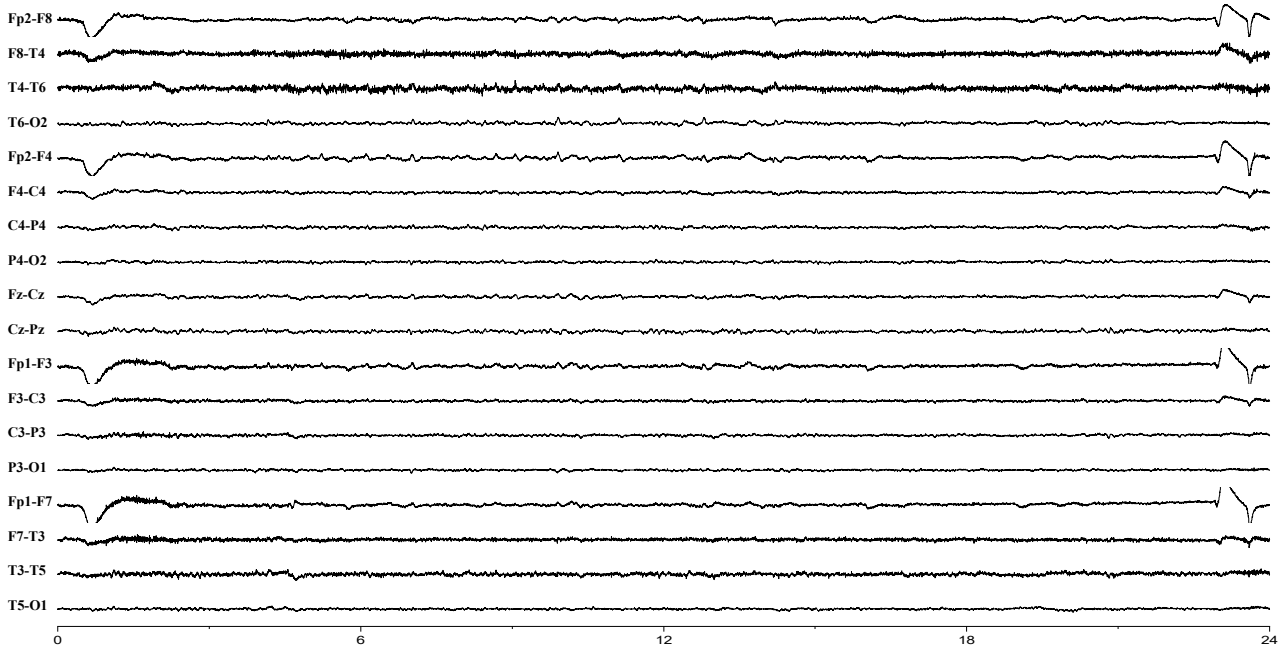

Fig. 15: Seizure undetected by AI. There is no clear frequency evolution and there are lots of muscle artifacts which can only be confirmed by the video (verified by the neurologist).