Seizure Type Classification Using EEG Signals and Machine Learning: Setting a Benchmark

S. Roy[1], U. Asif[2], J. Tang[2] and S. Harrer[2]

1. IBM Research Australia, now with Google Health, London, UK
2. IBM Research Australia, Melbourne, VIC, AU
roy.subhrajit20@gmail.com, {umarasif, jbtang, sharrer}@au1.ibm.com

Abstract— Accurate classification of seizure types plays a crucial role in the treatment and disease management of epileptic patients. Epileptic seizure types not only impact the choice of drugs but also the range of activities a patient can safely engage in. With recent advances being made towards artificial intelligence enabled automatic seizure detection, the next frontier is the automatic classification of seizure types. On that note, in this paper, we explore the application of machine learning algorithms for multi-class seizure type classification. We used the recently released TUH EEG seizure corpus (v1.4.0 and v1.5.2) and conducted a thorough search space exploration to evaluate the performance of a combination of various pre-processing techniques, machine learning algorithms, and corresponding hyperparameters on this task. We show that our algorithms can reach a weighted $F1$ score of up to 0.901 for seizure-wise cross validation and 0.561 for patient-wise cross validation thereby setting a benchmark for scalp EEG based multi-class seizure type classification.

keywords: Seizure type classification, Machine learning, Electroencephalography

I. Introduction

Despite many new advances in drug therapy and disease understanding, our capabilities in treating and managing epilepsy are extremely limited. Roughly 1% of the world's population, 65 million people, suffer from epilepsy [1]. For one third of these patients, no medical treatment options exist. These patients need to find ways to live with their condition and manage their daily lives around it. For the remaining two thirds of the patient population, medical treatment options are available but have vastly differing and constantly changing results and quality of treatment. These shortcomings in diagnosis and treatment options are caused by the fact that epilepsy is a highly individualized condition, i.e. it does not look the same in all patients and even for an individual patient disease expression changes over time. As a result, until recently, the lack of data and measurements made the correct matching of patients and drugs into an unnecessary, long process of trial and error. Manual diaries are the basic data source, but these have been proven to be only 50% accurate[2].

With the advent of mobile devices that allow collection of patient information in real-time, continuously and at the point of sensing, and leveraging miniaturization and IoT data collection platforms, new efforts are being directed towards building individualized patient management systems. Data that is more accurate and more extensive can be used to gain a patient specific understanding of the disease and provide support for decision-making in managing it.

Machine Learning has been successfully used to address a large variety of problems in the biomedical field, ranging from image classification in cancer diagnosis to the automatic interpretation of electronic health records[3–7]. Recently, we reported results demonstrating feasibility of using specialized neural networks to classify EEG data into normal/abnormal EEG [8] and to automatically detect and predict seizures [9]. A current review of epileptic seizure detection techniques using machine learning classifiers is provided in [10].

Recently, the International League Against Epilepsy (ILAE) released an updated seizure type classification framework [11]. In our paper, we expand on this work and discuss the feasibility of using machine learning algorithms for automatically distinguishing between different types of seizures as they are detected. This technology could support automatic, patient-specific seizure type logging in digital seizure diaries. Such seizure diaries could then be used to improve the performance of clinical trials through more efficient and reliable patient monitoring for endpoint detection, adherence control and patient retention [12].

II. Datasets

We used the TUH EEG Seizure Corpus (TUSZ) [13], which is the largest open source corpus of its type. This dataset includes the time of occurrence and type of each seizure.

The dataset covers a total of 8 different types of seizures: Focal Non-Specific Seizure (FNSZ): Focal seizures not further specified by type; Generalized Non-Specific Seizure (GNSZ): Generalized seizures not further classified into one of the groups below; Simple Partial Seizure (SPSZ): Partial seizures during consciousness; Type specified by clinical signs only; Complex Partial Seizure (CPSZ): Partial Seizures during

unconsciousness; Type specified by clinical signs only; Absence Seizure (ABSZ): Absence Discharges observed on EEG; patient loses consciousness for few seconds (Petit Mal); Tonic Seizure (TNSZ): Stiffening of body during seizure (EEG effects disappear); Tonic Clonic Seizure (TCSZ): At first stiffening and then jerking of body (Grand Mal) and Myoclonic Seizure (MYSZ): Myoclonus jerks of limbs.

v1.4.0 of the dataset released in Oct 2018 contains 2012 seizures as shown in Table 1. v1.5.2 of the dataset released in May 2020 contains 3050 seizures as shown in Table 2. Since the number of MYSZ samples was too low for statistically meaningful analysis, we did not include MYSZ seizures in our study hence making it a 7-class classification problem.

III. Methods

In this section, we briefly discuss the data preparation strategies, pre-processing techniques, machine learning algorithms and hyperparameter tuning methodologies we have explored.

For pre-processing the dataset, we re-sampled all channels' data to 250Hz, then Temporal Central Parasagittal (TCP) montage [14] is applied to create differentiated signal. We used two-popular methods which have been reported to be effective in analyzing EEG signals [15, 16]. In Method 1, we applied Fast Fourier Transform (FFT) to each W_l seconds of clip having O seconds overlap across all EEG channels. Next, we took $log_{10}()$ of the magnitudes of frequencies in the range $1-f_{max}$ Hz. After this operation, the dimension of each training sample becomes $(N, f_{max} * W_l * 250)$ where N is the number of TCP montage channels. For Method 2, first FFT is applied to each W_l seconds of clip having O seconds overlap across all EEG channels. Next, the output of FFT is then clipped from 1 to f_{max} Hz and normalized across frequency buckets. The correlation coefficients (N,N) matrix is calculated from this normalized matrix of $(N, f_{max} * W_l * 250)$. Real eigenvalues are calculated on this correlation coefficients matrix with complex eigenvalues made real by taking the complex magnitude. We only considered the upper right triangle of the (N,N) correlation coefficients matrix (since it is symmetric) and sorted by the eigenvalues magnitude.

Table 1. Seizure Type Statistics for v1.4.0

Seizure Type	Seizure Number	Duration (Seconds)	Patient Number
Focal Non-Specific (FNSZ)	992	73466	109
Generalized Non-Specific (GNSZ)	415	34348	44
Complex Partial (CPSZ)	342	33088	34
Absence (ABSZ)	99	852	13
Tonic (TNSZ)	67	1271	2
Tonic Clonic (TCSZ)	50	5630	11
Simple Partial (SPSZ)	44	1534	2
Myoclonic (MYSZ)	3	1312	2

Table 2. Seizure Type Statistics for v1.5.2

Seizure Type	Seizure Number	Duration (Seconds)	Patient Number
Focal Non-Specific (FNSZ)	1836	121139	150
Generalized Non-Specific (GNSZ)	583	59717	81
Complex Partial (CPSZ)	367	36321	41
Absence (ABSZ)	99	852	12
Tonic (TNSZ)	62	1204	3
Tonic Clonic (TCSZ)	48	5548	14
Simple Partial (SPSZ)	52	2146	3
Myoclonic (MYSZ)	3	1312	2

For classification, we used the following algorithms: k-Nearest Neighbors (k-NN), Stochastic Gradient Descent (SGD), XGBoost, and Convolutional Neural Networks (CNN). For the first three algorithms, we used HyperOpt [17] to choose the best hyperparameters. For CNN models, we used the popular ResNet50 [18] model and retrained the final layer for this task. Since different seizure type data is highly imbalanced, during the training of CNN, we will randomly sample the same number of seizure data from each seizure type to ensure a balanced input for each batch of training.

For cross validation, in v1.4.0, TNSZ and SPSZ classes only contain data from 2 patients therefore, patient-wise cross validation will not yield statistically meaningful results. Hence previous work in the field [19, 20] chose to apply 5-fold seizure-wise cross validation, in which the seizures from different seizure types will be equally and randomly allocated to 5 folds. In this scenario train and test datasets can contain different seizure samples from the same patient. Since version v1.4.0 of the dataset has been used for evaluation studies by multiple researchers [19, 20] we also include baseline results of our methods for v1.4.0 to allow a direct performance comparison to these studies. In v1.5.2 of the dataset all 7 selected seizure types comprise data from 3 or more patients, which allows statistically meaningful 3-fold patient-wise cross validation. In this scenario, train and test datasets will always contain seizure samples from different patients. This approach makes it more challenging to boost model performance but has higher clinical relevance as it supports model generalization across patients. For each seizure type, we randomly and equally allocate patients into each fold. We started with seizure types covering less patients and moved on to seizure types carried by more patients. For each seizure type, we exclude patients allocated to previous seizure types. Since datasets of individual patients comprise a different number of seizures, each fold's seizure number can vary largely. Hence we also investigated the impact of selecting different random seeds on the total number of seizures per fold and found that this had essentially no effect on the seizure number for each fold which varied only by plus-minus 3 seizures.

Due to the heavy imbalance of the dataset, we used a weighted-F1 scoring metric. This metric was applied to each seizure type's F1 metric as shown in the following

equation and created an average score according to the ratio of the total number of samples of a specific seizure type vs. the total number of seizure samples of all types combined in the dataset.

$$Weighted\_F1 = \sum_{n=1}^{7} \frac{\alpha_n \times F1_n}{7}$$

Here, $\alpha_n = \frac{Number\ of\ Seizure\ Type\ n}{Total\ Seizure\ Number}$ and $F1_n$ is a seizure type n's F1 score.

[21] reports a seizure type classification accuracy of 88.3% using convolutional neural networks, transfer learning and an earlier (February 2018) version of the TUSZ seizure corpus which contains 1163 different seizures subdivided 70%/30% into training and testing datasets respectively. [22] describes a classification experiment for detecting GNSZ, FNSZ, TCSZ seizure types and normal brain activity using 120 training samples and 90 test samples which achieved 91.4% classification accuracy. To the best of our knowledge, our work using the latest version v1.5.2 of the TUSZ seizure corpus (released in May 2020 and containing 3047 seizures) is the first seizure type classification study that provides a performance baseline for patient-wise cross validation.

IV. Experiments and Results

To explore the design space in an efficient manner, we chose the two computationally fastest classifiers from Sec. III namely k-NN and SGD classifier and generated their weighted-F1 scores using both pre-processing methods for both cross validation splits. For f_{max}, W_l, and O i.e. the pre-processing hyperparameters, we generated results for all combinations of f_{max} = {12, 24, 48, 64, 96} Hz, W_l = {1, 2, 4, 8, 16} secs, and O = {$0.5W_l$, $0.75W_l$} secs. The best hyperparameters of k-NN and SGD for each combination were automatically discovered by running Hyperopt for 100 iterations.

The above experiment served two purposes. Firstly, it allowed us to understand how the performance of the system varies with f_{max}, W_l and O separately which is shown in Figure 1 and Figure 2. Upon inspecting the top row of both figures, we find that while the performance is higher at mid-f_{max} of 24 and 48 Hz, it drops at extreme frequencies. This probably happens since at lower f_{max} we lose relevant information [23] and at higher f_{max} the number of dimensions increases, and the classifiers suffer from the curse of dimensionality. The second row of both figures suggests that the performance decreases when W_l increases. The third row of both figures suggests that the performance increases when O increases. We speculate that this happens since both the decrease of W_l and increase of O lead to more samples in the training set.

Table 3. v1.4.0 5-fold seizure-wise cross-validation results on the four top performing hyperparameter sets for each pre-processing method.

	f_{max}	W_l	O	$k-NN$	SGD	$XGBoost$	CNN
Method 1	48	1	$0.75W_l$	**0.884**	0.695	0.817	0.714
	24	1	$0.75W_l$	0.883	0.621	0.844	0.722
	96	1	$0.75W_l$	0.880	0.724	0.745	0.718
	24	1	$0.5W_l$	0.879	0.604	0.766	0.713
Method 2	48	1	$0.75W_l$	**0.901**	0.807	0.851	*NA*
	24	1	$0.75W_l$	0.900	0.783	0.858	*NA*
	24	1	$0.5W_l$	0.895	0.752	0.819	*NA*
	96	1	$0.75W_l$	0.890	0.806	0.866	*NA*

Table 4. v1.5.2 3-fold patient-wise cross-validation results on the four top performing hyperparameter sets for each pre-processing method.

	f_{max}	W_l	O	$k-NN$	SGD	$XGBoost$	CNN
Method 1	96	1	$0.75W_l$	0.466	0.432	**0.561**	0.524
	24	1	$0.75W_l$	0.437	0.384	0.559	0.530
	48	1	$0.75W_l$	0.467	0.407	0.526	0.525
	24	1	$0.5W_l$	0.423	0.390	0.512	0.504
Method 2	48	1	$0.75W_l$	0.401	0.469	**0.542**	*NA*
	96	1	$0.75W_l$	0.418	0.459	0.535	*NA*
	24	1	$0.5W_l$	0.392	0.452	0.530	*NA*
	24	1	$0.75W_l$	0.412	0.462	0.524	*NA*

Secondly, this design space exploration using simple classifiers revealed which combination of hyperparameters works best for both pre-processing methods. We select the four top performing sets of hyperparameters and perform 5-fold seizure-wise cross-validation for v1.4.0 and 5-fold patient-wise cross-validation for v1.5.2 on all the classifiers. Note that CNNs cannot be used to process the data from pre-processing method 2 as it does not produce 2D data. As before, hyperparameters have been chosen by running Hyperopt for 100 iterations. Table 3 shows the four top performing hyperparameter sets' average weighted-F1 scores for both pre-processing methods of v1.4.0 5-fold seizure-wise cross-validation. Table 4 shows the four top performing hyperparameter sets' average weighted-F1 scores for both pre-processing methods of v1.5.2 3-fold seizure-wise cross-validation.

Results shown in Table 3 and Table 4 demonstrate the feasibility of using machine learning techniques for automated seizure type classification. The best performing model types were k-NN achieving a weighted-F1 score of 0.901 for v1.4.0 and XGBoost reaching a weighted-F1 score of 0.561 for v1.5.2. We speculate that the reason for the more complex XGBoost algorithm being the best performer on v1.5.2 is the fact v1.5.2 contains more seizures and thus training samples than v1.4.0.

Seizure-wise cross-validation yielded a higher weighted F1 score than patient-wise cross-validation which indicates that it is more challenging to build machine learning models that generalize across patients than to build models which generalize across seizure types. When deploying such models as part of real world patient monitoring scenarios, continuous online re-training and updating of models during device operation will allow

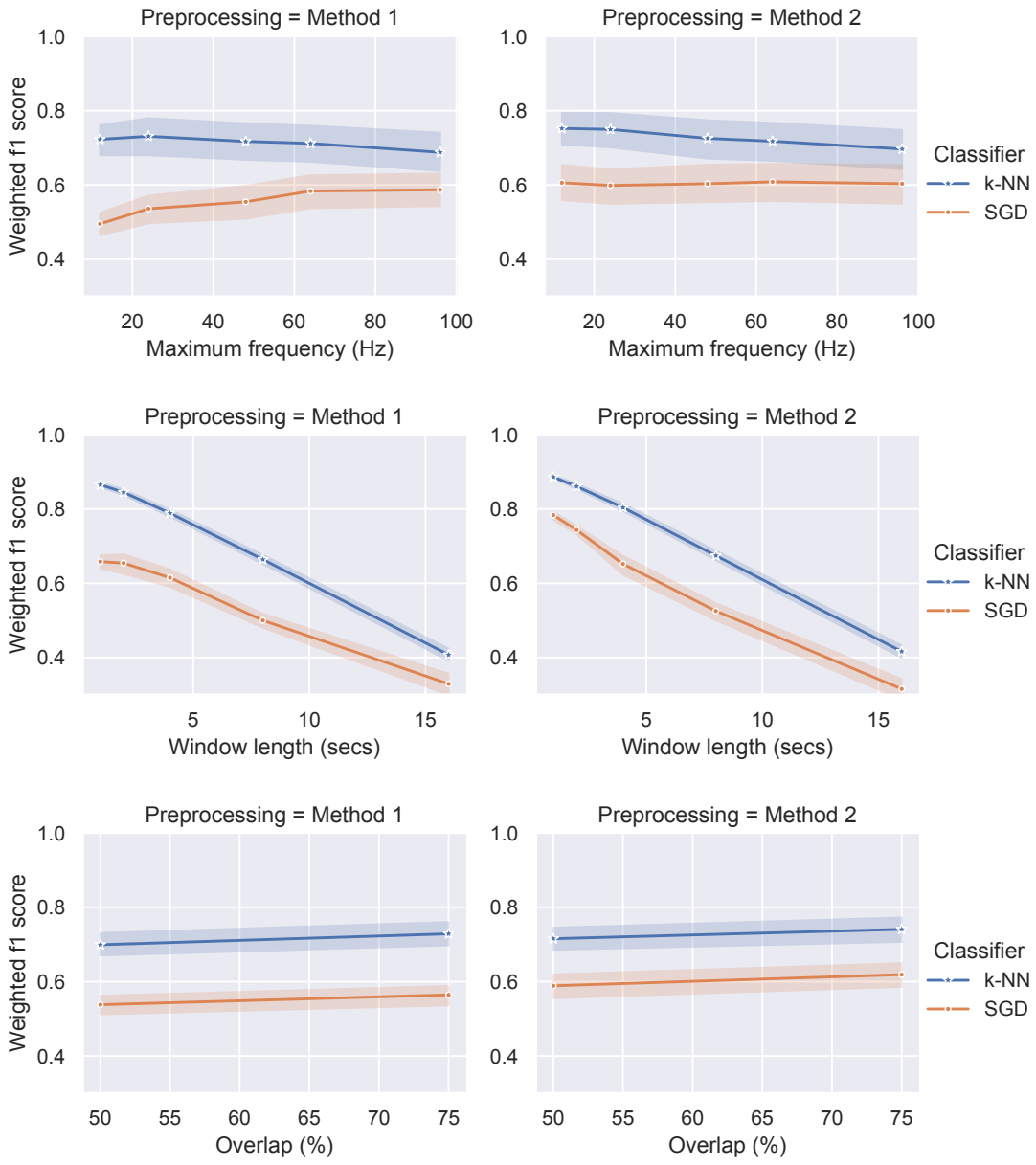

Figure 1. In this figure, we show how the weighted-F1 score varies with f_{max} (top row), W_l (middle row), and O (bottom row) for both pre-processing techniques on k-NN and SGD classifier for v1.4.0

to customize initially trained models to specific disease expressions of individual patients. Such real-time model re-training will also allow for adjusting detection and classification models to changing disease expressions in individual patients over time.

Automated detection and classification of seizures is the first step towards building digital seizure diary technology towards overcoming the severe limitations of manual diaries[2]. Digital seizure diaries could be used to support patient-specific seizure suppression and disease management systems. The methods described in this paper may play an important role for building digital seizure diary technology in the future.

V. Conclusion

In this study, by performing seizure-wise and patient-wise cross-validation for a variety of machine learning models applied to EEG data from epilepsy patients we demonstrated that machine learning techniques can be used to automatically classify different types of epileptic seizures. We hope that automatic classification of seizure types will improve long-term patient care, en-

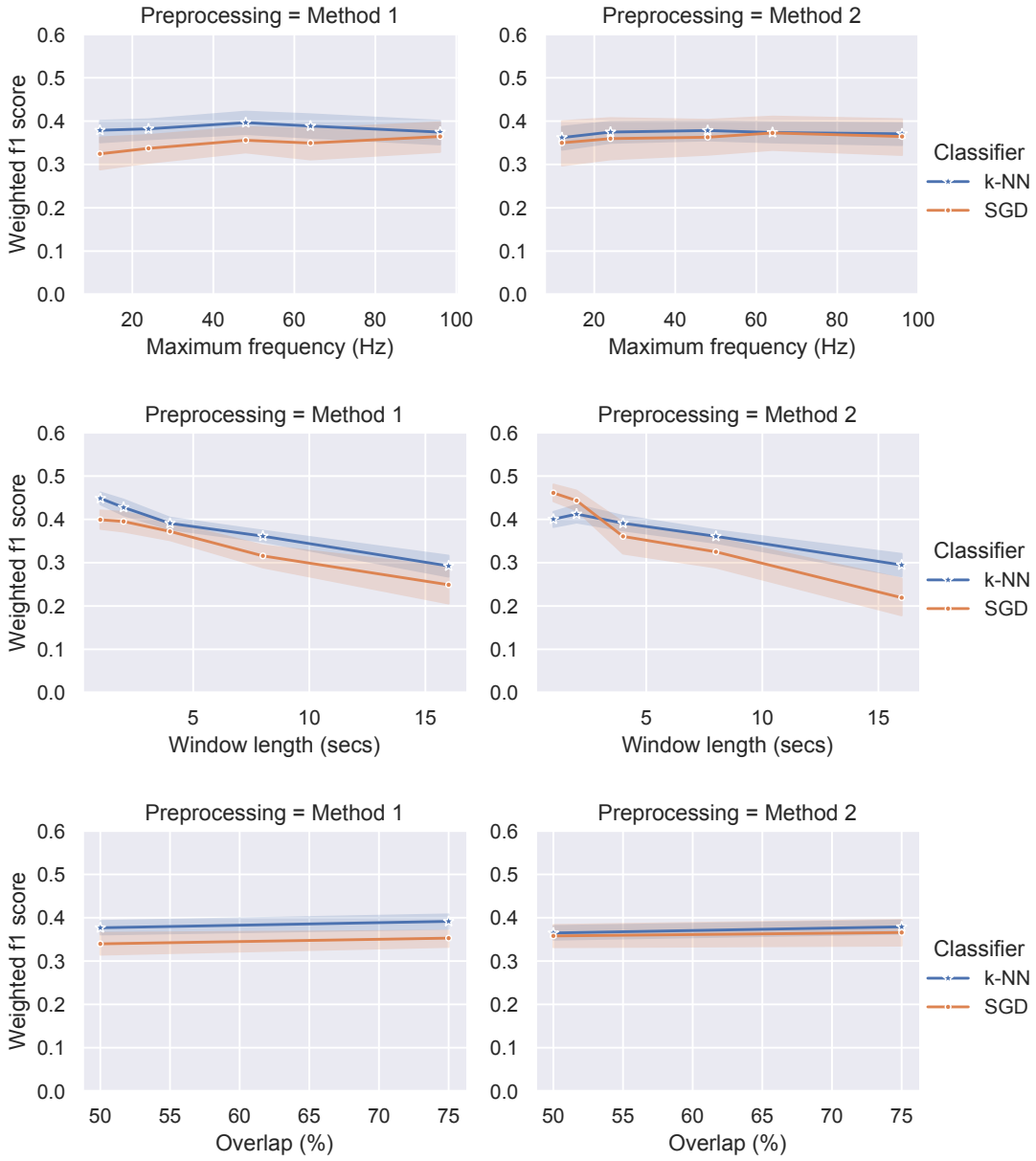

Figure 2. In this figure, we show how the weighted-F1 score varies with f_{max} (top row), W_l (middle row), and O (bottom row) for both pre-processing techniques on k-NN and SGD classifier for v1.5.2

abling timely drug adjustments and remote monitoring. To promote research in this topic, we have released our data pre-processing code for v1.4.0 and v1.5.2 [24], and we also plan to release the machine learning model code which we developed to generate the presented results.

References

[1] M. Patty Obsorne Shafer RN, "About Epilepsy: The Basics," https://www.epilepsy.com/learn/about-epilepsy-basics, 2014.

[2] R. S. Fisher, D. E. Blum, B. DiVentura, J. Vannest, J. D. Hixson, R. Moss, S. T. Herman, B. E. Fureman, and J. A. French, "Seizure diaries for clinical research and practice: limitations and future prospects," *Epilepsy & Behavior*, vol. 24, no. 3, pp. 304–310, 2012.

[3] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[4] W. Zhu, L. Xie, J. Han, and X. Guo, "The application of deep learning in cancer prognosis prediction," *Cancers*, vol. 12, no. 3, p. 603, 2020.

[5] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[6] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk *et al.*, "A clinically applicable approach to continuous prediction of

future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.

[7] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[8] S. Roy, I. Kiral-Kornek, and S. Harrer, "Deep learning enabled automatic abnormal eeg identification," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 2756–2759.

[9] I. Kiral-Kornek, S. Roy, E. Nurse, B. Mashford, P. Karoly, T. Carroll, D. Payne, S. Saha, S. Baldassano, T. O'Brien *et al.*, "Epileptic seizure prediction using big data and deep learning: toward a mobile system," *EBioMedicine*, vol. 27, pp. 103–111, 2018.

[10] M. K. Siddiqui, R. Morales-Menendez, X. Huang, and N. Hussain, "A review of epileptic seizure detection using machine learning classifiers," *Brain Informatics*, vol. 7, no. 1, pp. 1–18, 2020.

[11] R. S. Fisher, J. H. Cross, C. D'souza, J. A. French, S. R. Haut, N. Higurashi, E. Hirsch, F. E. Jansen, L. Lagae, S. L. Moshé *et al.*, "Instruction manual for the ilae 2017 operational classification of seizure types," *Epilepsia*, vol. 58, no. 4, pp. 531–542, 2017.

[12] S. Harrer, B. Antony, P. Shah, and J. Hu, "Artificial intelligence for clinical trial design," *Trends in Pharmacological Sciences*, vol. 40, no. 8, pp. 577–591, 2019.

[13] 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.

[14] S. Ferrell, V. Mathew, M. Refford, V. Tchiong, T. Ahsan, I. Obeid, and J. Picone, "The temple university hospital eeg corpus: Electrode location and channel labels," *Institute for Signal and Information Processing Report*, vol. 1, no. 1, 2020.

[15] Y. Paul, "Various epileptic seizure detection techniques using biomedical signals: a review," *Brain informatics*, vol. 5, no. 2, p. 6, 2018.

[16] K. Schindler, H. Leung, C. E. Elger, and K. Lehnertz, "Assessing seizure dynamics by analyzing the correlation structure of multichannel intracranial eeg," *Brain*, vol. 130, no. 1, pp. 65–77, 2007.

[17] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: a python library for model selection and hyperparameter optimization," *Computational Science & Discovery*, vol. 8, no. 1, p. 014008, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] D. Ahmedt-Aristizabal, T. Fernando, S. Denman, L. Petersson, M. J. Aburn, and C. Fookes, "Neural memory networks for seizure type classification," *arXiv*, pp. arXiv–1912, 2019.

[20] U. Asif, S. Roy, J. Tang, and S. Harrer, "Seizurenet: Multispectral deep feature learning for seizure type classification," *arXiv preprint arXiv:1903.03232*, 2019.

[21] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "Eeg based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Networks*, vol. 124, pp. 202–212, 2020.

[22] I. R. D. Saputro, N. D. Maryati, S. R. Solihati, I. Wijayanto, S. Hadiyoso, and R. Patmasari, "Seizure type classification on eeg signal using support vector machine," *Journal of Physics: Conference Series*, vol. 1201, no. 1. IOP Publishing, 2019, p. 012065.

[23] R. S. Fisher, W. Webber, R. P. Lesser, S. Arroyo, and S. Uematsu, "High-frequency eeg activity at the start of seizures." *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, vol. 9, no. 3, pp. 441–448, 1992.

[24] S. Roy, "IBM Features For Seizure Detection," https://www.isip.piconepress.com/projects/tuh_eeg/downloads/ibm_seizure_preprocessed/, 2019.