Features importance in seizure classification using scalp EEG reduced to single timeseries

Sébastien Naze[1], Jianbin Tang[1], James R. Kozloski[2], Stefan Harrer[1]

[1] *IBM Research Australia*
sebastien.naze@gmail.com; jbtang@au1.ibm.com; sharrer@au.ibm.com

[2] *Computational Biology Center, IBM T.J. Watson Research Center, Yorktown Heights, New York, USA.*
kozloski@us.ibm.com

***Abstract*—Seizure detection and seizure-type classification are best performed using intra-cranial or full-scalp electroencephalogram (EEG). In embedded wearable systems however, recordings from only a few electrodes are available, reducing the spatial resolution of the signals to a handful of timeseries at most. Taking this constraint into account, we tested the performance of multiple classifiers using a subset of the EEG recordings by selecting a single trace from the montage or performing a dimensionality reduction over each hemispherical space. Our results support that Random Forest (RF) classifiers lead most efficient and stable classification performances over Support Vector Machines (SVM). Interestingly, tracking the feature importances using permutation tests reveals that classical EEG spectrum power bands display different rankings across the classifiers: low frequencies (delta, theta) are most important for SVMs while higher frequencies (alpha, gamma) are more relevant for RF and Decision Trees. We reach up to 94.3% $\mp$ 5.3% accuracy in classifying absence from tonic-clonic seizures using state-of-art sampling methods for unbalanced datasets and leave-patients-out 3-fold cross-validation policy.**

***Index Terms*—TUH, tonic-clonic, absence seizures**

I. Introduction

Epilepsy manifests through seizures which occurs uncontrollably [1]. Several types of seizures exist based on semiology, symptomatic experience and electrophysiological signatures [2]. Patients with epilepsy can display several seizure types [3], and the monitoring of seizures for forecasting and detection is a subject of intense research [4], [5]. Recent advances in the field use elaborate methods from machine learning to analyse EEG timeseries and automatically extract the most relevant features from the signals to perform the detection or classification task [6], [7].

An issue of those deep learning methods lies in the lack of interpretability of the abstract features learned by the deep neural network to perform the task [8]. Using manually engineered features can help for interpretation but typically these perform sub-optimally on electrophysiological recordings [9], [10], therefore highlighting a trade-off challenge between efficiency and interpretability.

Another challenge in patient monitoring is the movement away from the hospital settings and towards recording spontaneous seizures from a wearable device at home or in daily life [11], [12] . A major drawback of these wearable recordings is that their spatio-temporal resolution is further constrained in order for the device to be minimally inconvenient for the patient [13], [14].

Here, we confront those two challenges by reducing the spatio-temporal resolution of EEG signals to single timeseries per hemisphere and training classifiers on these series using engineered features. A combination of preprocessing methods, sampling algorithms and classifier types is explored systematically. The importance of each feature for each classifier is assessed using the best combination of preprocessing steps.

II. Materials & Methods

A. Dataset

We used the EEG Corpus from the Temple University Hospital (TUH) dataset [15]. The data was recorded using scalp EEG with 20 electrodes following the standard international 10-20 system, and timeseries were analyzed using the longitudinal transverse bipolar montage. All patient recordings which sampling frequency were different from 256 Hz were resampled at 256 Hz prior to preprocessing.

B. Pre-processing and feature extraction

Samples of pre-ictal and ictal periods were created using 4s sample size (as in [16]). Periods shorter than the sample size were discarded. A combination of several preprocessing steps were performed on each sample:

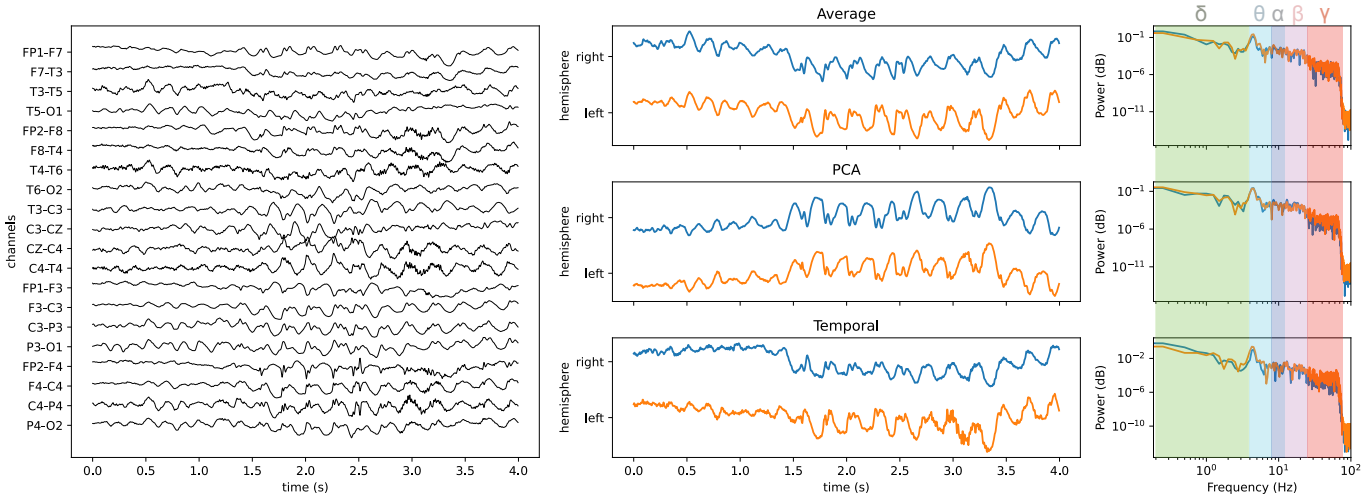

Fig. 1. Example of a 4s EEG sample (left) and its reduction to one timeserie per hemisphere (middle) using averaging over channels, PCA, or a subset of temporal channels. Power spectral features are used for α, β, γ, δ and θ bands (right).

Transformation to single timeseries per hemisphere: the signals from electrodes of each hemisphere were reduced to a single timeseries by either taking *(a)* the average over electrodes from the right and left hemisphere separately; *(b)* the strongest vector of the principal component analysis over the left and right hemisphere separately; or *(c)* the difference between electrodes (F7-T3 and F8-T4 for left and right hemispheres, respectively).

Normalization: the signal was then normalized for each sample using a z-score normalization over the sample size, i.e. $X = \frac{(x-\mu_x)}{\sigma_x}$, and further zero-centered using the moving-average over a 1s sliding window.

Fig. 1 illustrates the preprocessing of EEG samples. Features are computed as the average power over the alpha (α, 8-12 Hz), beta (β, 12-25 Hz), gamma (γ, 25-80 Hz), delta (δ, 0-4 Hz) and theta (θ, 4-8 Hz) frequency bands.

C. Sampling of imbalanced classes

Class imbalance can lead to skewed classification accuracies towards the class with most samples [17]. We alleviate this problem by sampling from the classes during training and testing phases using several methods from the `imblearn` library [18]:

- *Random Under Sampling* is a method that picks random samples (without duplicates) from the majority class(es) until the number of samples equals the minority class. This method is therefore referred to as unbiased from the sample distribution.
- *Cluster Centroids* uses the samples from the majority class that are closest to the class center determined by K-mean clustering. This methods is therefore biased towards samples that represent the average of the class.
- *Near-Miss* under-sampling is a method that takes the samples from the majority class which are on average the closest to the samples from the minority class [19]. This method makes the classification more difficult as samples are selected to be least discriminable between classes.

D. Classification

We assessed several types of classifiers to discriminate between absence and tonic-clonic seizures. Here, we briefly summarize the different characteristics of each classifier:

- *Support Vector Machines (SVMs)* try to find the maximally separating hyperplanes between samples from 2 classes (and do so repeatedly for each combination of classes in multi-class classification). The kernel is a similarity function that scores the distance between samples in feature space. We assess the score of SVMs using a linear kernel (simplest case) and a non-linear radial-basis function (RBF) kernel. Since our feature space is much smaller than our number of samples, the method was not prone to over-fitting and the regularization parameter was set to $C = 1$.
- *Decision trees* are hierarchical structures (connected nodes) whereby each branching represents conditions over features which aggregate through the depth of the tree to define class labels. The conditions for the split criterion are found by maximizing the entropy at each node of the tree. We set the maximum depth of the tree to be the number of features (5) and a minimum of 2 samples is necessary to split a node into branches.
- *Random Forests* are a set of many decision trees which outputs are combined to form the best weighted conglomerate. Individual trees are created at random based on feature values at the beginning and then refined through training.

E. Cross-validation

We performed our analysis using 3-fold cross validation scheme with leave-patient-out policy. This means that for

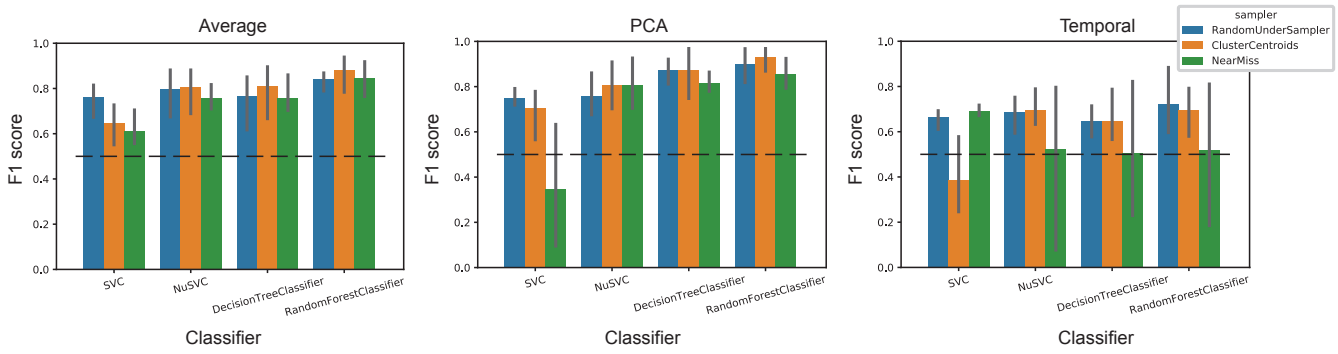

Fig. 2. Overall accuracy of each classification scheme computed by F1-score across preprocessing methods. The horizontal dashed line is the score under the null hypothesis, i.e. when the class is predicted randomly.

each split, all samples from a patient are either in the training set or the test set, without overlap. This prevents over-fitting on the specific patient trace which is commonly observed in standard n-fold cross validation policies ([6], [7]). We used the Stratified Group K-fold splitting module from `scikit-learn` [20], whereby classes are seizure labels and groups are individual patients.

F. Feature importance

The importance of each of the spectral band features to the overall performance of trained classifiers is assessed through a random permutation test over samples (rows), one feature (column) at a time. For each permutation, the classifier is re-trained using the permuted feature values and the score is compared to the originally trained classifier's score [21]. The score's difference is averaged over 100 permutations to give a value between 0 (low importance) to 1 (high importance), and was performed using the `scikit-learn` library [20].

III. Results

We created a pipeline combining sample creation from raw EEG, preprocessing, feature extraction, dataset splitting for 3-fold cross validation, sampling, classification and extraction of feature importances. We first assessed the performance of each classifier using a combination of preprocessing steps and data samplers. We then analyse which features play a major role in discriminating between seizure types.

A. Performance of the different classifiers across preprocessing methods

Figure 2 shows the performance of the 4 classifiers using the 3 samplers and 3 different dimensionality reduction techniques at preprocessing . We observe that performing a PCA over the EEG instead of using a subset of temporal electrodes systematically increases the accuracy of the downstream classification (mean +17.4%, SD 8.8%). It is also observed that the PCA does better than averaging across all electrodes per hemisphere (mean +0.9%, SD 3.7%).

The difference in sampling method accounts to 9.9% (SD 2.8%) of classification performance. Cluster centroids resulted in the highest performances. Surprisingly, a near miss sampling can outperform a random sampling (or perform similarly) when using a dimensionality reduction technique such as PCA or averaging. Also, linear SVM is more sensitive to the sampling methods that non-linear SVM, decision tree and random forest classifiers.

B. Feature importance differs across classifiers

Lastly, we assess the contribution of each feature to the classification accuracy. This is performed by randomly permuting row entries of a feature column and re-training the classifier to assess its new accuracy using the permuted feature values. The drop in accuracy is indicative of the feature importance of the column being shuffled. Since preprocessing using PCA and cluster centroids sampling reached the best classification accuracy, we show feature importances for each classifiers using this preprocessing methods but similar results were observed using averaging across electrodes and other sampling methods.

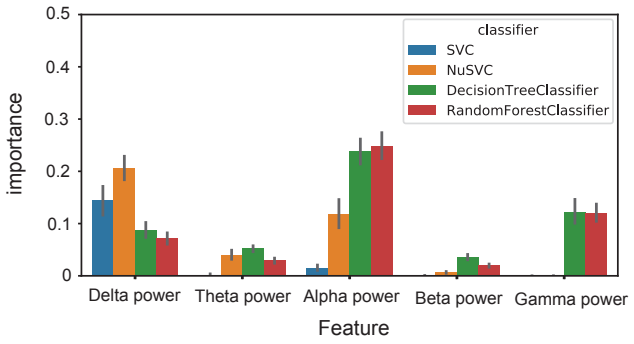

Fig. 3. Feature importance across classifiers.

Figure 3 shows that the feature importance for the delta frequency band (0-4 Hz) exceeds all others for SVM classifiers, while higher frequency bands (alpha and gamma) play the major role for decision tree and random forest classifiers. This indicates that higher frequencies in the EEG play a subtle but essential role in improving the classification performances

across seizure types and across the classification methods that we explored.

IV. Conclusion

We assessed several pre-processing methods in the context of EEG signal classification between tonic-clonic and absence seizures. By systematically comparing classification performances across timeseries normalization schemes, sampling from imbalanced classes and dimensionality reduction techniques, our results demonstrate that applying a PCA over the whole EEG signals leads to better outcomes than using only a subset of electrodes or averaging across electrodes. This indicates that sacrificing temporal precision for spatial integration of the signals across the scalp is beneficial for this seizure type classification task. It is clear that in the designing of wearable systems for patients monitoring using EEG, recordings from many electrode spatially distributed over the scalp gives better classification outcomes than using a single electrode. This is especially important for patients experiencing a wide range of seizure types or when the seizure semiology evolves across the span of the disorder (e.g. transitioning from absence to tonic-clonic [22]).

Our work also paves the route for more interpretable results of machine learning outputs. This is especially important for medical applications, since a mechanistic understanding of the AI systems can permit better comprehension of the processes at play in patients. Others have reviewed features of interest for seizure detection and classification [10], [23]. Our results indicate that while low frequency EEG component (i.e. delta band) is most relevant for classification using SVM, more elaborate classifiers devote greater importance to the higher frequencies (alpha and gamma bands). Since the interplay of low and high frequency discharges during seizure is complex and specific to certain seizure types, our results suggest that each classifier type is picking up on those features differently. A next step will involve modeling of seizure dynamics [24], and optimization of model parameters to reproduce the different classes synthetically as conceptually introduced in a previous study on classifying transcranial magnetic stimulation responses [9].

V. Acknowledgments

We are indebted to Joseph Picone, Iyad Obeid and their team of researchers, clinicians, engineers and technicians at Temple University who collected and curated the dataset.

References

[1] P. Jiruska, M. D. Curtis, and J. G. R. Jefferys, Eds., *Modern Concepts of Focal Epileptic Networks*, 1st ed. Academic Press, Jul. 2014.

[2] V. K. Jirsa, W. C. Stacey, P. P. Quilichini, A. I. Ivanov, and C. Bernard, "On the nature of seizure dynamics," *Brain*, vol. 137, no. 8, pp. 2210–2230, 2014.

[3] C. Bernard, S. Naze, T. Proix, and V. K. Jirsa, "Modern concepts of seizure modeling," *Int. Rev. Neurobiol*, vol. 114, pp. 121–153, 2014.

[4] I. Kiral-Kornek, S. Roy, E. Nurse, B. Mashford, P. Karoly, T. Carroll, D. Payne, S. Saha, S. Baldassano, T. O'Brien, D. Grayden, M. Cook, D. Freestone, and S. Harrer, "Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System," *EBioMedicine*, vol. 27, pp. 103–111, Jan. 2018.

[5] L. Kuhlmann, P. Karoly, D. R. Freestone, B. H. Brinkmann, A. Temko, A. Barachant, F. Li, G. Titericz Jr, B. W. Lang, and D. Lavery, "Epilepsyecosystem. org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG," *Brain*, vol. 141, no. 9, pp. 2619–2630, 2018, publisher: Oxford University Press.

[6] S. Roy, U. Asif, J. Tang, and S. Harrer, "Seizure Type Classification Using EEG Signals and Machine Learning: Setting a Benchmark," in *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec. 2020, pp. 1–6, iSSN: 2473-716X.

[7] U. Asif, S. Roy, J. Tang, and S. Harrer, "SeizureNet: Multi-Spectral Deep Feature Learning for Seizure Type Classification," in *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, ser. Lecture Notes in Computer Science, 2020, pp. 77–87.

[8] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, and et al., "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, Apr. 2018.

[9] S. Naze, V. Caggiano, Y. Sun, M. V. Lucas, A. Etkin, and J. R. Kozloski, "Classification of TMS evoked potentials using ERP time signatures and SVM versus deep learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 3539–3542.

[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, pp. 1–18, 2020.

[11] S. Beniczky, I. Conradsen, O. Henning, M. Fabricius, and P. Wolf, "Automated real-time detection of tonic-clonic seizures using a wearable EMG device," *Neurology*, vol. 90, no. 5, pp. e428–e434, 2018.

[12] K. Vandecasteele, T. De Cooman, Y. Gu, E. Cleeren, K. Claes, W. V. Paesschen, S. V. Huffel, and B. Hunyadi, "Automated epileptic seizure detection based on wearable ECG and PPG in a hospital environment," *Sensors*, vol. 17, no. 10, p. 2338, 2017.

[13] D. Sopic, A. Aminifar, and D. Atienza, "e-Glass: A Wearable System for Real-Time Detection of Epileptic Seizures," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[14] H. Ocak, "Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2027–2036, 2009.

[15] A. Harati, S. López, I. Obeid, J. Picone, M. P. Jacobson, and S. Tobochnik, "The TUH EEG CORPUS: A big data resource for automated EEG interpretation," in *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec. 2014, pp. 1–5.

[16] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, "EpilepsyGAN: Synthetic Epileptic Brain Activities With Privacy Preservation," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, pp. 2435–2446, Aug. 2021.

[17] M. K. Siddiqui, X. Huang, R. Morales-Menendez, N. Hussain, and K. Khatoon, "Machine learning based novel cost-sensitive seizure detection classifier for imbalanced EEG data sets," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 14, no. 4, pp. 1491–1509, 2020.

[18] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[19] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, 2003.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] S. Beniczky, G. Rubboli, A. Covanis, and M. R. Sperling, "Absence-to-bilateral-tonic-clonic seizure: A generalized seizure type," *Neurology*, vol. 95, no. 14, pp. e2009–e2015, Oct. 2020.

[23] P. Boonyakitanont, A. Lek-uthai, K. Chomtho, and J. Songsiri, "A review of feature extraction and performance evaluation in epileptic seizure detection using EEG," *Biomedical Signal Processing and Control*, vol. 57, p. 101702, Mar. 2020.

[24] S. Naze, C. Bernard, and V. Jirsa, "Computational Modeling of Seizure Dynamics Using Coupled Neuronal Networks: Factors Shaping Epileptiform Activity," *PLOS Computational Biology*, vol. 11, no. 5, p. e1004209, May 2015.