Engineering Nonlinear Epileptic Biomarkers Using Deep Learning and Benford's Law

Joseph Caffarini (Caffarini@neurology.wisc.edu)
University of Wisconsin-Madison Department of Neurology
Klevest Gjini
University of Wisconsin-Madison Department of Neurology
Brinda Sevak
University of Wisconsin-Madison Department of Neurology
Roger Waleffe
University of Wisconsin-Madison Department of Computer Science
Mariel Kalkach-Aparicio
University of Wisconsin-Madison Department of Neurology
Melanie Boly
University of Wisconsin-Madison Department of Psychiatry
Aaron F Struck
University of Wisconsin-Madison Department of Neurology

Research Article

Keywords: Encoder, Entropy, Benford's Law, Convolutional, Fully Connected, Dense, Neural Network, Dimensionality Reduction, Feature Selection, Deep Learning, iEEG, Classifier, Random Forest, Epilepsy, Gini Importance, Seizure Identification, Biomarkers

Posted Date: December 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1105250/v1

Abstract

In this study we designed two deep neural networks to encode 16 feature latent spaces for early seizure detection in intracranial EEG and compared them to 16 widely used engineered metrics: Epileptogenicity Index (EI), Phase Locked High Gamma (PLHG), Time and Frequency Domain Cho Gaines Distance (TDCG, FDCG), relative band powers, and log absolute band powers (from alpha, beta, theta, delta, gamma, and high gamma bands.

The deep learning models were pretrained for seizure identification on the time and frequency domains of one second single channel clips of 127 seizures (from 25 different subjects) using “leave-one-out” (LOO) cross validation. Each neural network extracted unique feature spaces that were used to train a Random Forest Classifier (RFC) for seizure identification and latency tasks. The Gini Importance of each feature was calculated from the pretrained RFC, enabling the most significant features (MSFs) for each task to be identified. The MSFs were extracted from the UPenn and Mayo Clinic's Seizure Detection Challenge to train another RFC for the contest. They obtained an AUC score of 0.93, demonstrating a transferable method to identify interpretable biomarkers for seizure detection.

Introduction

The gold-standard for seizure detection remains expert opinion. Dependence on human interpretation discourages the use of chronically implanted Intracranial Electroencephalography (iEEG) devices such as Neuropace® because the high volume of data required to train the device renders manual iEEG interpretation impractical and error prone, making automated seizure detection a necessity. Therefore, reliable automated methods capable of improving efficiency and accuracy of seizure detection could significantly impact therapy adequacy assessment and optimize time-to-treatment.

There are several iEEG analysis metrics that have been used to detect seizures. The work at hand focuses on the methods derived from the Fast Fourier Transform (FFT) of an iEEG segment. Specifically, we used band powers, Epileptogenicity Index (EI) and Phase Locked High Gamma (PLHG). Analysis of band powers refers to the comparison between frequency bands in the power spectrum (alpha, beta, gamma, and theta) with different functional properties.[18] EI is the ratio of high frequency power to low frequency power. PLHG uses the high and low frequency components of the Hilbert Transform to quantify ictal phase-amplitude coupling. These measures are useful in identifying epileptogenic zones and are intuitive because they use predefined frequency bands.[1,2] They also have grounding in the current understanding of the neurophysiology underlying seizure generation.[25] The effectiveness and simplicity of the selected features make them benchmarks for comparing other engineered metrics for neurological applications, such as approximate entropy, sample entropy, and fuzzy entropy.[4,21,22,25,26] In this study, the Euclidean distance (sometimes referred to as the Cho – Gaines distance) was chosen instead of entropy because EEG recordings have been shown to follow Benford’s Law[16], making them potentially relevant to seizure detection. The Cho-Gaines distance was specifically applied to both the time and frequency domains (TDCG, FDCG).[9] Altogether, these metrics are advantageous because they can be easily parallelized on a

graphics processing unit (GPU) to extract digit information not directly accessible to a neural network or Fourier Transform.

In addition to engineered features, deep neural network (DNN) feature spaces were examined for potential biomarkers because they can be GPU accelerated, do not require explicitly designed features, and can abstract higher level details from training data. Such properties make them useful for real time perceptual applications, as demonstrated in numerous studies.[5,7,8,10,12,17,20,21,23] However, DNNs are essentially a "black box" approach, making them difficult to approve for clinical use because clinicians cannot interpret how the network generates its predictions. Our goal was to address shortcomings in interpretability by creating an encoder for learning seizure-specific features and selecting the most relevant for a particular task. Inspired by autoencoders[13], we designed each neural networks to have an encoder sub network for learning feature representations and a classifier sub network in lieu of a decoder for classifying the encoded feature spaces.

Results

Encoder Training for Seizure Specific Features:

Three DNNs were trained with architecture described in **Figure 1a-c**. The first two "black box NNs" use an encoder-classifier setup designed to automatically *encode* EEG data into learned features and then *classify* seizure probability based on these features. The third DNN contains only a classifier, replacing the encoder with hand engineered features (called the EMC), and serves as a control for assessing the performance of the first two. Each model was pretrained on the "seizure detection" task and given a one second iEEG segment to predict the seizure probability. We used single channel segments to force the models to learn ictal-specific features and so they could be applied to recordings with different electrode configurations and lengths. After training, the black box neural networks were adept to extract features from iEEG data by computing the output of the encoder. The two black box networks were named after their encoders: a Convolutional Neural Network encoder or Time Domain Black Box (TDBB), and Power Spectrum Neural Network encoder or Frequency Domain Black Box (FDBB). The architectures of each model are described in **Figure 1ab** and the pretraining classifier performance for all architectures is summarized in **Figure 1c**.

If the information captured by the DNNs was characteristic of seizure activity, then we expected all models to exhibit similar performance on the "seizure identification" task because they have identical classification sub networks. As shown in **Figure 1c,** the black box networks performed about the same as the EMC network on the pretraining task, with the FDBB scoring a slightly higher AUC score than the TDBB, and slightly worse than the EMC (AUC 0.836 ± 0.31 vs 0.823 ± 0.35 and 0.837 ± 0.31, respectively). The similar performance between all neural networks indicates that the black box feature spaces contain information that is just as relevant to seizure activity as the engineered metrics ensemble.

Feature Space Comparisons:

The pretraining stage demonstrated that each DNN can extract characteristic biomarkers for seizure activity. The next goal was to study the significance and interpretability of the extracted features from the pretrained models and to see whether they could be used for a new "latency" task: determining if a seizure is within 15s of onset. To achieve this goal, two Random Forest Classifiers (RFCs), one for each task, were created each using an ensemble feature set made by concatenating the features from the TDBB, FDBB, and EMC. The RFCs allowed the importance of all features to be evaluated for each task. The results of this analysis are summarized in **Figure 2bc** for seizure identification. After the most significant features (MSFs) were identified, an additional RFC was trained for each task using only the MSFs, allowing for the contributions of the MSFs to model accuracy be evaluated (by comparison with the accuracy of the RFCs using all features).

As shown in **Figure 2b**, the RFC using only MSFs performed essentially the same as the RFC using all features, illustrating that the ensemble feature set can be reduced without sacrificing performance. **Figure 2c** summarizes the Gini Importance of each feature within the feature ensemble during RFC for seizure identification. The most significant features from each individual feature vector (Engineered or EMC, FDBB or p, and TDBB or z) were identified as Log Beta Absolute Band Power, p0, and z9, respectively.

Once the MSFs were identified, the overall relationships between all features were examined using a correlation matrix to determine if the DNNs were learning features resembling the hypothesis driven, engineered features previously used for seizure detection. As shown in **Figure 2d**, when each of the black box feature spaces were compared to the engineered feature set, they were most correlated to the Log Absolute Band Powers, indicating the DNNs learned nonlinear combinations of the Log Absolute Band powers. When compared to each other, a few of the z features were highly correlated with many p features, demonstrating that the higher order abstractions learned by the DNNs are based on clinically relevant signal properties. Surprisingly, little to no correlation was seen between the DNN feature sets and nonlinear features like EI and PLHG, indicating that the models learned a different, unspecified relationship between band powers. The correlation of the DNN feature sets with each other and the Log Absolute Band Powers illustrates that they generated nonlinear combinations of clinically relevant features, potentially forming new biomarkers for use in new machine learning tasks.

Kaggle Challenge Attempts:

After the MSFs were identified, we evaluated their transferability by using them to identify seizures on a completely unseen dataset. We also compared the MSF classifiers with a purely deep learning approach to evaluate the selected biomarkers. The TDBB was used as a control for the Kaggle competition because it performed slightly better overall than other neural networks on single channel classification, as shown in **Figure 1c**. The channel controlled TDBB obtained a test AUC score of 0.77, and the subject specific TDBBs obtained a test AUC score of 0.87. The former attempt is likely more robust because it was trained on a broader dataset and avoids overfitting patient specific information. Next, the MSFs were extracted from the Kaggle dataset and used to train two RFCs, one for identification and another for latency. This reduced feature set greatly outperformed the deep learning models and control ensembles

by achieving a maximum score of 0.93. The Kaggle test scores for each of the MSFs, least significant features (LSFs), and all features as the controls are included in **Figure 3.**

Discussion

The encoder – classifier architecture allowed for the development of feature embeddings for generating nonlinear feature spaces for classification. The TDBB and FDBB both were able to encode latent spaces capable of achieving state of the art performance in seizure identification.

The pretrained encoders from both DNNs were able to generate feature spaces highly transferable to new datasets. The MSFs from these pretrained models were selected and applied to the Kaggle Challenge, where they achieved an AUC score 0.93. The strong performance on the Kaggle competition demonstrates that MSFs identified during pretraining are transferable to new datasets, and that DNNs are capable of extracting biomarkers from iEEG recordings. It is important to note that the MSFs from DNNs change depending on the size of the dataset and cross-validation method (LOO vs by channel, or sample) used to pretrain the models. This finding suggests that to accurately transfer features, extra care is required to prevent overfitting and data leakage during pretraining. As a general practice to make a generalizable seizure detection algorithm, it is critical that the training and validation datasets use completely different subjects, not just different seizures, or channels from the same subject. It is common to overfit to patient specific seizure factors that make transferability limited. Transferability was confirmed in this experiment by using a publicly available database only after fully developing the models.

There are many other seizure identification experiments using neural networks and engineered features[10,23], but direct comparison to many of the models is difficult because of inconsistencies between the performance metrics chosen, how the models are trained, how the scores are reported (maximum achieved vs average score), and the composition of the dataset. Data composition matters because most projects report the accuracy, which has recently been shown to be inconsistent, due to its dependency on the data composition and the model architecture.[10,11] Regardless of the metric, performance can be inflated if **i)** there is a large class imbalance between ictal and preictal data points, **ii)** data leakage between validation or testing groups exists, or **iii)** the validation scores are reported from cross validation instead of a separate holdout set. These inflation-fostering conditions are difficult to grasp without source code. Often score reporting lacks consistency in describing whether the result refers to the achievement of a maximum or an average score. This could imply the reporting of outliers rather than expected performance.

To overcome the issues identified as obstacles in model comparison, our models were compared to others that completed similar tasks with some comparable metrics. A useful source for assessing performance was the UPenn and Mayo Clinic's Seizure Detection Challenge, where we scored an AUC 0f 0.93 using MSFs, and 0.95 when using all features. In the former case, our MSF ensemble was significantly smaller than the features used by other, higher-ranking submissions while using a simpler

RFC model (30 estimators vs 3000 estimators) to generate predictions. These differences made our MSF ensemble more computationally efficient to train on longer recordings without significantly hurting performance. Additionally, our models were able to achieve scores within the ranges of those reported by other state of the art methods while using a more defined feature set and shorter time window.[5,10,12]

Based on comparisons to similar models, we can conclude that our proposed method was able to obtain state of the art performance in a simple identification task and can be applied to longer datasets to fully assess its utility in early seizure detection. The best features from deep learning models can be identified and selected using Gini Importance with minimal preprocessing for use in other machine learning tasks. Using this method, we found that the simplest features, such as the Beta and Gamma Log Absolute Band Powers, are the most important, and may be foundational to engineer novel epileptic biomarkers for early seizure detection.

Methods

1-Subjects and IEEG Database

Approval was obtained from the University of Wisconsin Institutional Review Board, informed consent was obtained from all human subjects prior to participation. All methods were conducted in accordance with the Declaration of Helsinki.

After approval of the respective IRBs, the raw EEGs from 25 pre-surgical epilepsy patients were visually inspected. The seizure onset time and channels involved with seizure onset were marked and confirmed by a board-certified neurophysiologist (Boly M). Only channels involved at seizure onset were used for further analysis.

The selected dataset consisted of 710 channels taken from 127 different seizures across 25 individuals: 13 seizures from five patients were from the University of Wisconsin; 82 seizures from 13 patients were from the Epilepsiae Database[15]; 32 seizures from seven individuals were provided by Mayo Clinic.[14] Data from the Epilepsiae database was prefiltered with a 50Hz notch filter, and data from Mayo Clinic and the University of Wisconsin were prefiltered with a 60Hz notch filter. Each lead containing seizures was resampled to 400Hz and cropped to two minutes in length, with seizure onset being one minute into the segment. The preictal labels were assigned to every point before seizure onset, and the ictal labels were assigned to every point at and after seizure onset, creating an even amount of preictal and ictal data points. Overall, there were 15,240 seconds of data, consisting of 50% preictal and 50% ictal/post-ictal points, where each point is a one second clip.

2 - Computing Engineered Features

A – Computing Digit Distributions and Cho Gaines Distance:

The entire dataset was broken into one second epochs, and the power spectrum of each epoch was computed using the magnitude of the FFT. The first nonzero digits were counted and normalized to generate a probability distribution over the extracted digits.

The leading nonzero digits were extracted from each time point using the following element-wise formula:

$$l = \left\lfloor \frac{|e|}{10^{\lfloor \log_{10}(|e|) \rfloor}} \right\rfloor \qquad (1)$$

Where e represents an arbitrary sized tensor.

The results were set to zero whenever the function was undefined, i.e., where e = 0. The specific digits between one and nine were counted and normalized to generate a probability distribution over the extracted digits within the window. The Cho-Gaines Distance (a.k.a. Euclidian Distance[9]) was used to compare observed probability distributions (**Equation 3)**, with expected probability distributions as calculated using the formula from Benford's 1937 work[3], as shown in **Equation 2**.

$$P_d = \log_{10}\left(\frac{d+1}{d}\right) \quad (2)$$

Where d is a digit from {1,2,3,4,5,6,7,8,9}.

$$D = \sqrt{N * \sum_{i=1}^{9} (e_i - o_i)^2} \quad (3)$$

Where N is the EEG sample rate, e_i is the expected frequency that i is the first digit, o_i is the observed frequency where i is the first digit.

B - Band Powers:

The delta (2 to 4 Hz), theta (4 to 8 Hz), alpha (8 to 12 Hz), beta (12 to 30 Hz), gamma (30 to 80 Hz), and high gamma (80 to 150 Hz) band powers were computed by using Welch's method.[2,6,18] Band powers were then normalized by the total signal power to create the relative band powers, and the log of the Absolute Band Powers were also used as features. The DC component (f = 0 Hz) was dropped before computing relative band powers.

C - Epileptogenicity Index (EI):

The formula for computing the EI is shown in **Equation 4:**[2]

$$EI = \frac{P_\beta + P_\gamma + P_{\gamma'}}{P_\theta + P_\alpha} \quad (4)$$

Where P are the relative band powers of the α, β, γ, γ', and θ are the alpha, beta, gamma, high gamma, and theta bands, respectively.

D – Phase Locked High Gamma (PLHG):

The formula for computing PLHG is shown in **Equation 5**:[1]

$$\text{PLHG} = \left| \frac{1}{n} \sum_{1}^{n} A_{HFO} * e^{\Phi_{LF} - \Phi_{HFO}} \right| \quad (5)$$

Where A_{HFO} is the instantaneous signal envelope of the high frequency oscillations, and is the difference between the low and high frequency instantaneous phases. LF in this case means from the theta, alpha, and beta bands, where the HFO means from the gamma and high gamma bands.

3 - Neural Network Design:

Three neural networks were designed to learn from one second segments of iEEG data: A fully connected network for learning features from the Relative Welch Power Spectrum (FDBB), a convolutional neural network for extracting time domain features (TDBB), and a fully connected network for classifying engineered metrics (EMC). Each DNN had distinct encoder and classifier sub networks to allow for the networks to be used in dimensionality reduction. The classifier sub network was the same in all three models to allow for the comparison between different feature spaces. The output of the encoder sub networks was used for feature analysis. The dropout in the classifier sub network was 75% for FDBB, 90% for the TDBB, and 40% for the EMC. All networks were designed to work on any arbitrary number of electrodes by changing the input channels parameter. This allowed the same network architecture to be applicable to more datasets, such as the 2015 Kaggle competition "UPenn and Mayo Clinic's Seizure Detection Challenge", Kaggle contest. The architectures for all three models are included in **Figure 1ab**.

4 - Neural Network Pretraining:

All three neural networks were trained using leave-one-out (LOO) cross validation across 25 subjects, where all the data points from one patient were withheld for training. For each fold, models were trained using stochastic gradient descent with Adam optimizer and binary cross entropy loss until there was no improvement in the validation loss for ten epochs. Training and validation sets were also separated by subject, with eight training subjects being assigned to the validation set. The training set was shuffled for each fold to ensure different validation and training sets for each fold. Within each set, the data was broken up into separate, single channel 1sec segments. The AUC, Brier, PPV, NPV, Recall, and Accuracy were all recorded for each test set, and the model with the best overall scores was selected to attempt the Kaggle challenge. Cross validation was conducted with the same random seed, allowing the data points used in each fold to be identifiable and consistent across models.

5 - Feature Analysis:

For each fold, the test subject and associated pretrained model were selected to extract the FDBB and TDBB feature spaces, preventing data leakage during RFC training. The feature spaces were then analyzed using Gini Importance and a correlation analysis. Gini Importance is a computationally efficient splitting criterion used in RFCs that splits samples into nodes minimizing the impurity of each node during fitting, making it the most efficient way of evaluating features when using RFCs.[19] The correlation matrix between all feature combinations was computed to illustrate potential relationships between the DNNs and engineered metrics, and enabled us to determine if the DNN feature spaces were related to the engineered metrics.

6 - Feature Ensembles:

As summarized in Steps 2 and 3 in **Figure 2a**, leave-one-out (LOO) cross validation of a Random Forest Model across 25 subjects was used to generate interpretable models using all features for both seizure identification and latency determination tasks. The latency task is defined by the UPenn-Mayo Kaggle challenge and consists of identifying ictal segments from the first 15 seconds of the seizure as the positive group (12.5% of samples).[24]The dataset was expanded in the pretraining stage by splitting all channels for each subject and using a one second sliding window with 50% overlap to generate additional frames, as demonstrated in the winning Kaggle Submission. For the remaining 24 subjects within each LOO fold, an RFC (30 estimators, split by entropy, max depth of five) was trained using 100-fold cross validation on single channel FDBB, TDBB, and EMC. The Gini Importance of each feature was aggregated from the pretrained models from each internal cross validation fold and external LOO fold, making a score distribution from 2,400 importance calculations overall. Importance scores are represented in **Figure 2c**.

7 – Kaggle Challenge:

The deep neural network with the best AUC score on our seizure identification task was selected to attempt the Kaggle competition. The competition consisted of seizure and latency identification tasks, where contestants design a model for first identifying seizures, and another for determining if the seizure is within 15 seconds of onset (determine latency). Each subject had their own instance of the neural network module to prevent issues from mismatching channel counts. Each subject model was trained using stochastic gradient descent, Adam optimizer, and binary cross-entropy with five-fold cross validation by sample, where 25% of the validation set was held out for testing after each epoch. Additionally, another TDBB model was trained by projecting each sample point to be 16 leads, allowing for the same model to be applied to all subject data by controlling for the number of channels. These approaches were included because they resemble how neural networks are typically applied to real world problems and provide a baseline for comparison.

Dataset augmentation consisted of creating time points starting at each half second by combining the last half of the previous sample with the first half of the next, effectively using a sliding window with 50%

overlap. The final response was graded by the Kaggle auto grader, which generated an ROC score representing the average score of the latency task and identification task.

In addition to the TDBB classifiers, the TDBB features were extracted from the Kaggle Dataset by averaging all feature spaces from each of the 25 pretrained LOO folds to create a single average feature space. This is possible because each pretraining LOO fold has its own fully trained model, and there are 25 possible folds because each patient is held out once. From the averaged features, the three most important and single most important for both the identification and latency tasks were extracted from each channel, creating two separate feature sets tailored to each task. The same was done with the least important features to serve as controls. All feature sets were used to train Random Forest classifiers for each subject (30 estimators, entropy, max depth of five) to evaluate the transferability of our DNNs and determine if they can be used for feature engineering. The trained estimator from each fold of this cross-validation step was used to create a Kaggle entry to allow for a distribution of 30 potential scores for various feature ensembles. It is important to emphasize that the neural networks used to extract the black box features were pretrained on data unrelated to the Kaggle set.

Declarations

Acknowledgements:

We acknowledge support from the following sources:

R01NS111022

K23NS112473

Lily's Fund for Epilepsy Research

Tiny Blue Dot Foundation

Author contributions:

Joseph Caffarini: Conceptualized and developed the neural networks and feature selection methods outlined in paper, developed code for the project, assisted in annotating iEEG data.

Klevest Gjini: Conceptualized the feature selection methods. Assisted in Data analysis, Collected iEEG data. Edited manuscript.

Brinda Sevak: Collected iEEG data and assisted in annotation.

Roger Waleffe: Conceptualized the neural network architectures. Edited manuscript.

Mariel Kalkach-Aparicio: Edited manuscript.

Melanie Boly: Collected and annotated iEEG data.

Aaron F. Struck: Conceptualized seizure detection problem, verified seizure annotations, edited manuscript, and provided clinical relevance to project.

Additional Information:

The authors do not have any conflicts of interest.

References

1. Bandarabadi, M. *et al.* Assessing Epileptogenicity Using Phase-Locked High Frequency Oscillations: A Systematic Comparison of Methods. *Frontiers in Neurology* **10,** (2019).
2. Bartolomei, F., Chauvel, P. & Wendling, F. Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG. *Brain* **131,** 1818–1830 (2008).
3. Benford, F. The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society* **78,** 551–572 (1938).
4. Bruhn, J., Röpcke, H. & Hoeft, A. Approximate Entropy as an Electroencephalographic Measure of Anesthetic Drug Effect during Desflurane Anesthesia. *Anesthesiology* **92,** 715–726 (2000).
5. Burrello, A., Schindler, K., Benini, L. & Rahimi, A. Hyperdimensional computing with local Binary Patterns: One-shot learning of SEIZURE onset and identification of ICTOGENIC brain regions Using Short-time iEEG Recordings. *IEEE Transactions on Biomedical Engineering* **67,** 601–613 (2020).
6. Crone, N. E., Sinai, A. & Korzeniewska, A. High-frequency gamma oscillations and human brain mapping with electrocorticography. *Progress in Brain Research* 275–295 (2006). doi:10.1016/s0079-6123(06)59019-3
7. Daoud, H. & Bayoumi, M. A. Efficient Epileptic Seizure Prediction Based on Deep Learning. *IEEE Transactions on Biomedical Circuits and Systems* **13,** 804–813 (2019).
8. Ditthapron, A., Banluesombatkul, N., Ketrat, S., Chuangsuwanich, E. & Wilaiprasitporn, T. Universal Joint Feature Extraction for P300 EEG Classification Using Multi-Task Autoencoder. *IEEE Access* **7,** 68415–68428 (2019).
9. Druică, E., Oancea, B. & Vâlsan, C. Benford's law and the limits of digit analysis. *International Journal of Accounting Information Systems* **31,** 75–82 (2018).
10. Gómez, C. *et al.* Automatic seizure detection based on imaged-EEG signals through fully convolutional networks. *Scientific Reports* **10,** (2020).
11. Huang, J. & Ling, C. X. Using AUC and accuracy in Evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* **17,** 299–310 (2005).
12. Hussein, R., Palangi, H., Wang, Z. J. & Ward, R. Robust detection of epileptic seizures using deep neural networks. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018). doi:10.1109/icassp.2018.8462029
13. Kingma, D. P. & Welling, M. An Introduction to Variational Autoencoders. (2019). doi:10.1561/9781680836233

14. Kini, L. G., Davis, K. A. & Wagenaar, J. B. Data integration: Combined imaging and electrophysiology data in the cloud. *NeuroImage* **124,** 1175–1181 (2016).
15. Klatt, J. *et al.* The EPILEPSIAE database: An extensive electroencephalography database of epilepsy patients. *Epilepsia* **53,** 1669–1676 (2012).
16. Kreuzer, M. *et al.* Brain Electrical Activity Obeys Benford's Law. *Anesthesia & Analgesia* **118,** 183–191 (2014).
17. Li, J., Struzik, Z., Zhang, L. & Cichocki, A. Feature learning from incomplete EEG with denoising autoencoder. *Neurocomputing* **165,** 23–31 (2015).
18. Malik, A. S. & Amin, H. U. Designing an EEG Experiment. *Designing EEG Experiments for Studying the Brain* 1–30 (2017). doi:10.1016/b978-0-12-811140-6.00001-1
19. Nembrini, S., König, I. R. & Wright, M. N. The revival of the Gini Importance? *Bioinformatics* **34,** 3711–3718 (2018).
20. Qiu, Y., Zhou, W., Yu, N. & Du, P. Denoising Sparse Autoencoder Based Ictal EEG Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 1–1 (2018). doi:10.1109/tnsre.2018.2864306
21. Rasheed, K. *et al.* Machine Learning for Predicting Epileptic Seizures Using EEG Signals: A Review. *IEEE Reviews in Biomedical Engineering* **14,** 139–155 (2021).
22. Schwilden, H. Concepts of EEG processing: from power spectrum to bispectrum, fractals, entropies and all that. *Best Practice & Research Clinical Anaesthesiology* **20,** 31–48 (2006).
23. Shoeibi, A. *et al.* A comprehensive comparison of handcrafted features and convolutional autoencoders for epileptic seizures detection in EEG signals. *Expert Systems with Applications* **163,** 113788 (2021).
24. UPenn and Mayo Clinic's Seizure Detection Challenge. *Kaggle* (2014). Available at: https://www.kaggle.com/c/seizure-detection. (Accessed: 22nd June 2021)
25. Weiss, S. A. *et al.* Ictal high frequency oscillations distinguish two types of seizure territories in humans. *Brain* **136,** 3796–3808 (2013).
26. Xiang, J. *et al.* The detection of epileptic seizure signals based on fuzzy entropy. *Journal of Neuroscience Methods* **243,** 18–25 (2015).
27. Yang, Y. *et al.* Epileptic Seizure Prediction Based on Permutation Entropy. *Frontiers in Computational Neuroscience* **12,** (2018).

Figures

Figure 1

Neural Network Architectures and LOO Pretraining. a) Feature extraction pipeline. Example shows one second of single channel data being processed by pipeline. b) Neural network architectures. The power spectrum neural network (FDBB) was designed to extract features from the positive frequencies of the log Welch power spectrum. The convolutional neural network (TDBB) was designed to extract features from the time domain. The Classifier sub network common was to both neural networks and was used as the only component of the engineered metrics classifier (EMC). Using the same classifier network for all three models allows all feature spaces to be compared during feature selection. The variable C represents the number of channels in the input data, with C = 1 during the pretraining stage. c) Pretraining performance of neural network models. The TDBB was selected for the Kaggle Challenge because it appeared to edge out the other models in most of the metrics.

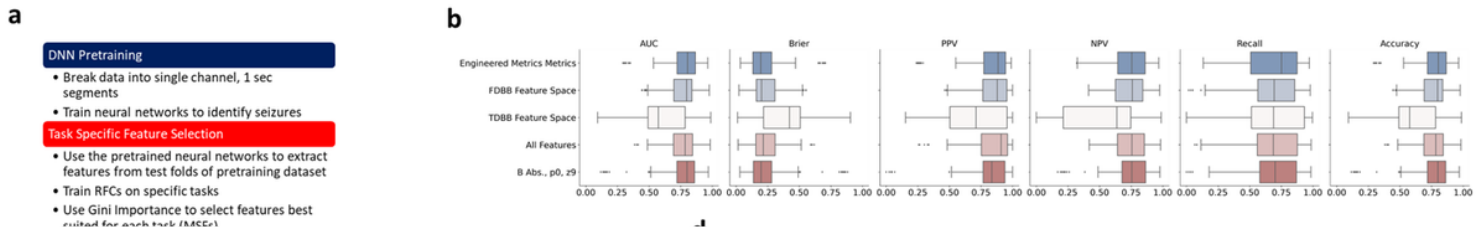

Figure 2

Process for identifying significant features during pretraining using random forest. a) Process outline for identifying MSFs. Ordered from top to bottom, steps 2 and 3 were applied to extract MSFs for identification and latency tasks. b) Comparative performance of feature fnsembles on single channel RFC pretraining. The scores for each LOO test fold are included. The most important features were able to achieve comparable performance to the more complete feature ensembles. c) Feature gini importance from single channel identification task. Each feature ensemble has one highly dominant feature. d) Correlation matrix between all feature ensembles. Both black box encoders learned to extract features correlating to the absolute band powers and weakly to the PLHG and EI. Some black box features were highly correlated to each other, suggesting that they are capturing related information.

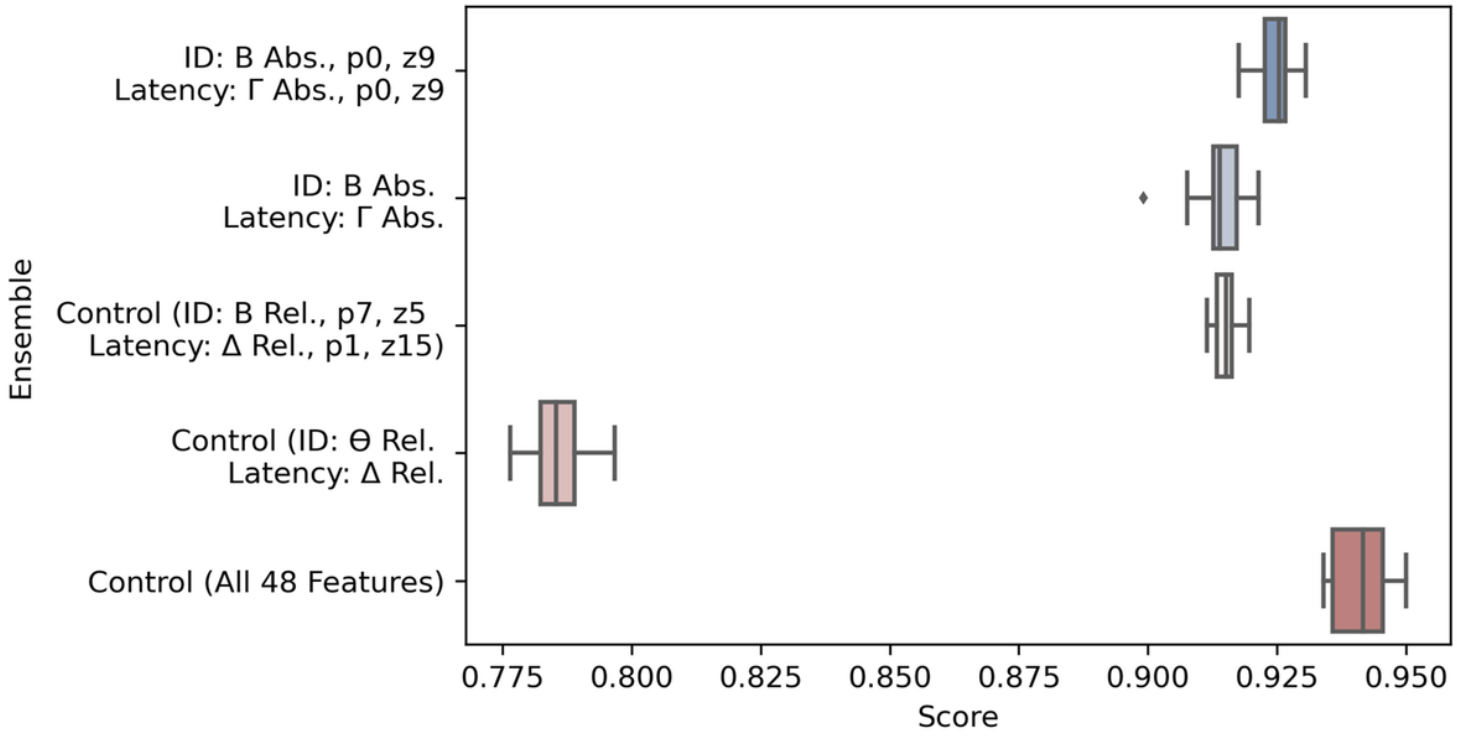

Figure 3

Kaggle test scores using most and least significant feature (MSF and LSF) ensembles. B Abs. and Γ Abs. are the log absolute band powers for the B and Γ bands, respectively. Δ Rel. and Θ Rel. are the relative band powers of the Δ and Θ bands. Variables starting with p are FDBB features, and variables starting with z are TDBB features. The Kaggle submissions using the 3 MSFs for each task outperformed both the LSF control groups but underperformed the control containing all 48 features. The model using the MSFs, i.e., the B Abs. and Γ Abs. was able to achieve the third highest score, and the control groups with the LSFs, was unable to beat the maximum scores of the other 3 groups. The control group without black box features performed significantly worse than the other models, indicating that black box features contribute more to classification than Gini Importance suggests, and even the least important black box features still have predictive value.