EEG-GCNN: Augmenting Electroencephalogram-based Neurological Disease Diagnosis using a Domain-guided Graph Convolutional Neural Network

Neeraj Wagh NWAGH2@ILLINOIS.EDU
Department of Bioengineering, University of Illinois at Urbana-Champaign

Yogatheesan Varatharajah VARATHA2@ILLINOIS.EDU
Department of Bioengineering, University of Illinois at Urbana-Champaign

Editors: Emily Alsentzer[⊗], Matthew B. A. McDermott[⊗], Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy[‡], Stephanie L. Hyland[‡]

Abstract

This paper presents a novel graph convolutional neural network (GCNN)-based approach for improving the diagnosis of neurological diseases using scalp-electroencephalograms (EEGs). Although EEG is one of the main tests used for neurological-disease diagnosis, the sensitivity of EEG-based expert visual diagnosis remains at ∼50%. This indicates a clear need for advanced methodology to reduce the false negative rate in detecting abnormal scalp-EEGs. In that context, we focus on the problem of distinguishing the abnormal scalp EEGs of patients with neurological diseases, which were originally classified as 'normal' by experts, from the scalp EEGs of healthy individuals. The contributions of this paper are three-fold: 1) we present EEG-GCNN, a novel GCNN model for EEG data that captures both the spatial and functional connectivity between the scalp electrodes, 2) using EEG-GCNN, we perform the first large-scale evaluation of the aforementioned hypothesis, and 3) using two large scalp-EEG databases, we demonstrate that EEG-GCNN significantly outperforms the human baseline and classical machine learning (ML) baselines, with an AUC of 0.90.

Keywords: EEG, Early diagnosis, Neurological disease, Graph CNN

1. Introduction

Neurological disorders (NDs) are diseases of the nervous system involving the brain, spinal cord, nerves, and muscles, and affect ∼1 billion people worldwide (W.H.O., 2006). EEG is one of the main diagnostic tests in neurology, where the visual identification of abnormal brain activity in a brief scalp EEG recording session (20–60 minutes) indicates the potential for NDs. However, it is very common to record EEGs that do not contain visible abnormalities; for example, 50% of the EEGs recorded from patients with seizures are deemed "normal" based on expert visual review (Smith, 2005). Such scenarios can cause delays in delivering clinical care and put patients at continued risk for injuries and comorbidities (Bouma et al., 2016). Thus, there is a critical need to develop EEG-based analytical tools that can enable a more rapid diagnosis of NDs.

The task of visually identifying abnormal EEG is challenging due to multiple reasons: abnormal discharges may not occur during a short EEG session; they may originate in deeper brain structures like

the cingulate, hippocampus, or insula; they may be activated only during sleep, which was not recorded; they may involve too small an amount of cortex to be measurable on the scalp; or subtle abnormalities are highly likely to evade inspection by the naked eye (Ebersole and Leroy, 1983). However, a recent study showed that deviations in EEG spectral characteristics, such as the alpha rhythm, can help distinguish epilepsy patients from healthy individuals even when visual classification was not possible (Varatharajah et al., 2020). The study also showed that the spatial patterns of the abnormalities are localized to specific brain regions impacted by the disease. Motivated by those findings, we sought to perform a large-scale evaluation of distinguishing "normal" EEGs of patients with NDs from EEGs of healthy individuals using state-of-the-art machine learning (ML) techniques.

In this paper, we present EEG-GCNN, a graph convolutional neural network (GCNN)-based approach that achieves state-of-the-art performance in classifying "normal" EEGs of patients with NDs versus EEGs of healthy individuals. The key contributions of our paper are the following: 1) we present the first large scale evaluation of the task at hand using the EEGs of 208 healthy individuals and 1,385 patients with NDs; 2) the proposed model includes a novel graph representation for EEG data using spatial and functional connectivity measures; and 3) EEG-GCNN achieves an AUC of 0.90 on the held-out test set and significantly outperforms human and classical ML baselines (10% improvement).

2. Related Work

Clinical relevance: Prior research related to the diagnosis of epilepsy has focused on visual identification of common epileptic abnormalities, such as interictal spikes and sharp waves (Hauser et al., 1982). Furthermore, the majority of the existing ML-based approaches have targeted automating the visual identification of normal and abnormal EEGs, using expert labels as ground truth (Schirrmeister et al., 2017; Roy et al., 2018; Alhussein et al., 2019). However, visual identification of abnormal EEGs is $\sim$ 50% sensitive, and therefore does not provide reliable ground truth labels. A clinically more important question is whether ML can distinguish between healthy EEGs and EEGs of patients that do not contain any visually identifiable abnormalities. A study by Varatharajah et al. (2020) provides strong evidence for this hypothesis. They found that deviations in brain health markers like the alpha rhythm help distinguish patients from healthy individuals using a modest sample and hand-tuned features. However, a large-scale evaluation of this hypothesis using state-of-the-art ML approaches has not been performed.

Technical relevance: Prior studies have proposed graphical-model-based approaches to encode the rich spatial and temporal information content of EEG data (Varatharajah et al., 2017). However, those studies have hard-coded the spatio-temporal relationships based on domain knowledge and therefore, lack the ability to learn from data. Recent studies have addressed this limitation using GCNNs, albeit being limited to the application of emotion recognition (Song et al., 2018; Wang et al., 2018). They have proposed a dynamical graph convolutional neural network model (DGCNN) with an adjacency matrix reflecting the functional coupling between EEG electrodes. While EEG-GCNN and DGCNN share some commonalities, there are notable differences. First, DGCNN does not fully exploit the power of graph representation as it tries to binarize the adjacency matrix during training. Whereas, EEG-GCNN captures

brain connectivity through edge weights on a fully-connected graph. Second, DGCNN makes extensive use of hand-tuned features, whereas, EEG-GCNN applies minimal feature engineering. Finally, based on evidence that the pathological changes induced by chronic neurological diseases are spatially related and show temporally localized functional patterns (Hyun et al., 2011), we employ a connectivity measure reflecting both the spatial and functional relationships between brain regions.

3. Data

We support our experiments by pooling together two publicly available large scalp EEG databases: 1) the Temple University Hospital EEG (TUH EEG) Corpus (Obeid and Picone (2016)), which contains clinical EEG recordings of patients with NDs and 2) the Max Planck Institute Leipzig Mind-Brain-Body (MPI LEMON) Dataset (Babayan et al. (2019)), which contains resting-state recordings from healthy participants.

TUH EEG: This dataset comprises of >30,000 EEG recordings collected at TUH starting from 2002. The recordings vary in terms of patient ages, diagnoses, medications, channel configurations, and sampling frequencies. A subset of recordings in TUH EEG have been broadly annotated by experts as either "normal" or "abnormal", and have been released as a derived dataset called the TUH EEG Abnormal Corpus (TUAB). For our experiments, we only utilize the TUAB recordings that are annotated as "normal" while ignoring those labeled as "abnormal", leading to a total of 1385 EEGs from 1385 distinct patients.

MPI LEMON: This dataset represents a cross-sectional sample of 228 healthy individuals from Leipzig, Germany. The sample comprised two age groups: young adults (ages 20-35) and older adults (ages 59-77). EEG recordings were made using 62 electrodes in the 10-10 sensor configuration with a sampling rate of 2500Hz, for a total of 216 participants. Each subject's session is made up of 16 trials, each 60 seconds long: 8 eyes-closed and 8 eyes-open. We included data from both trials in our experiments. The raw data were corrupted for 8 subjects, which resulted in a useful set of healthy EEGs from a total of 208 healthy subjects.

4. Data Preprocessing & Feature Engineering

The overall flow of EEG preprocessing and feature extraction is depicted in Figure 1.

Preprocessing: We employed a very minimal level of preprocessing, with each raw recording being transformed as follows: (1) a common subset of bipolar montage electrodes were selected from the raw channels, (2) the recording was resampled to 250Hz, followed by (3) a highpass filter at 1Hz, and finally (4) a notch filter at the power-line frequency of 50Hz. We emphasize that neither were routine physiological EEG artifacts like eye blinks or muscle movements explicitly suppressed nor were bad channels rejected. Implementation was done using routines provided by the MNE-python library (Gramfort et al. (2013)).

Channel selection: Because the EEG data of healthy individuals were recorded using a 62-channel 10-10 system, we selected a subset of channels that matched the 10-20 system used to record the EEG data from epilepsy patients. Within the EEG data of selected channels from the healthy and patient populations, we selected 4 bipolar pairs of electrodes from each hemisphere, producing 8 channels of EEG data for each participant. Thus, our analysis involves the following bipolar channels: F7-F3, F8-F4, T7-C3, T8-C4, P7-P3, P8-P4, O1-P3, and O2-P4.

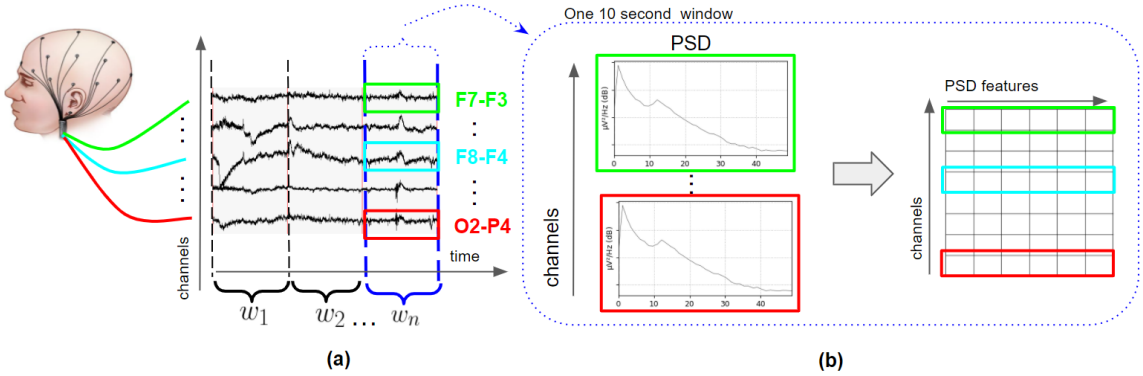

Figure 1: Schematic representation of the feature extraction process.

Windowing: We divided the preprocessed recordings into contiguous non-overlapping windows of 10s each. Each window consists of an EEG recording from eight bipolar channels as defined earlier. We make a simplifying assumption that the signal in each 10s window is independent of other windows in the same recording. It is important to note that while ML training is done using windows, window predictions are aggregated to form subject predictions. This setup is illustrated in Figure 1(a).

Classification data: We labeled TUAB recordings as "diseased" (patient), and MPI LEMON as "healthy". Each window for a subject was assigned the same label as the parent recording. This results in a total of 203,616 diseased windows and 21,718 healthy windows. We highlight a target class imbalance in the dataset at two levels: 1) recording-level imbalance ratio of ∼7:1 (diseased:healthy) and 2) window-level imbalance ratio of ∼9:1 (diseased:healthy). The handling of imbalance at the window-level (relevant for ML training) is discussed later in the manuscript.

Features: The frequency content of the windowed EEG signals, obtained through the Power Spectral Density (PSD), was summarized into the dominant brain wave bands defined as follows: delta (1-4Hz), theta (4-7.5Hz), alpha (7.5-13Hz), lower beta (13-16Hz), higher beta (16-30Hz), and gamma (30-40Hz). We extracted the total band power from each band for each of the 8 montage channels, leading to a feature matrix of shape (8 channels x 6 features) for each window, as shown in Figure 1. Figure 2 illustrates the differences between the two classes based on the extracted features using boxplots. Note that the features from channels in left and right hemispheres were averaged to generate combined features for each region. In addition, the features were z-scored for visual clarity.

5. Model Description

In this section, we describe the motivation behind EEG-GCNN and mathematically formulate some of its distinct features. Overall, the aim of EEG-GCNN is to learn representations of EEG activity in a way that incorporates functional coupling (coordinated-firing activity) of distributed regions of the brain and the structural coupling (through physical white-matter tracts) between distant brain regions.

Our proposed model (shown in Figure 3) consists of several parts. First, it describes the 10-second EEG window using a graph structure, which is given as input to the GCNN model. Second, graph convolutions are performed on the input data to generate node-level embeddings. Third, an averag-

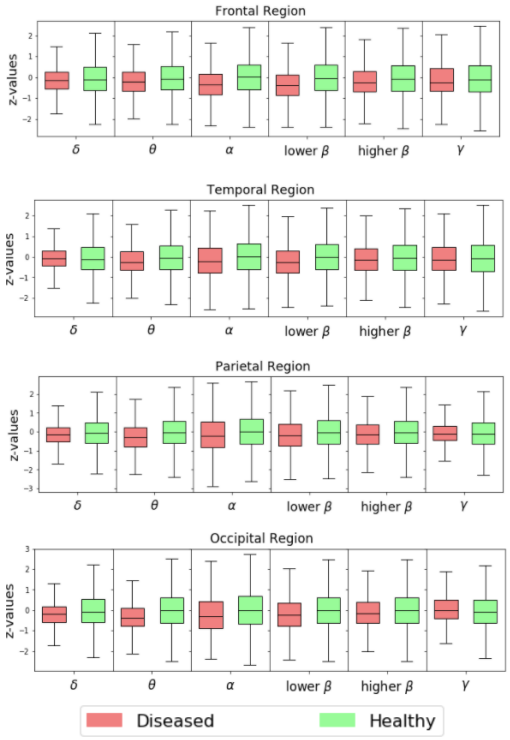

Figure 2: Box-plots representing the group differences of spectral features.

ing operation is performed across the nodes to generate graph-level embeddings. The graph-level embeddings are then provided as inputs to the fully-connected network, which predicts the output class. We describe each of these steps in the following sections.

Graph structure: Suppose that EEG data of a subject are recorded through M channels. Initially, the data is discretized by dividing the recording duration into N windows. We represent the interactions between the channels at a window n as a dynamic graph $G_n = (V, E_n)$, where V is the set of $|V| = M$ channels and $E_n \subset V \times V$ is the set of undirected links between channels. Data on G_n can be represented by a feature matrix $X_n \in \mathbb{R}^{m \times d}$, where d denotes the input feature dimension per channel. The edge set E can be represented by a weighted adjacency matrix $A \in \mathbb{R}^{m \times m}$.

Connectivity measures: The adjacency between channels i and j is denoted as A_{ij} and is a combination of spatial (A^s_{ij}) and functional (A^f_{ij}) connectivity.

$$A_{ij} = \frac{1}{2}\left(A^s_{ij} + A^f_{ij}\right) \tag{1}$$

Specifically, the edge weight between each pair of nodes is made of two components: 1) the geodesic distance between the two electrodes when the standard 10-20 electrode configuration is mapped to a unit sphere (a proxy for spatial brain connectivity) and 2) the coherence values between the timeseries signal of the two electrodes (a proxy for functional brain connectivity). While the coherence values naturally lie in $[0, 1]$, the geodesic distances were standardized into the same range. Both these measures are then averaged to generate edge weights that lie within $[0, 1]$.

The geodesic distance A^s_{ij} between two points on a sphere of radius r with coordinates (x_i, y_i, z_i) and (x_j, y_j, z_j) in Cartesian coordinate space is defined as:

$$A^s_{ij} = \arccos\left(\frac{x_i x_j + y_i y_j + z_i z_j}{r^2}\right)$$

The spectral coherence A^f_{ij} between the two channel timeseries i and j, with cross-spectral density S_{ij} and power spectral densities S_{ii} and S_{jj}, is defined as:

$$A^f_{ij} = \frac{\mid E[S_{ij}] \mid}{\sqrt{E[S_{ii}].E[S_{jj}]}}$$

Graph convolutions: We use the spectral graph convolution propagation rule defined by (Kipf and Welling, 2016). Suppose there are L number of graph convolutional layers and let $l = 0, 1, \ldots, L-1$ denote the layer number. Each graph convolution produces a feature transformation of its inputs as described below, where $W^{(l)}$ denotes the weights of layer l, $H^{(l)}$ denotes the output of

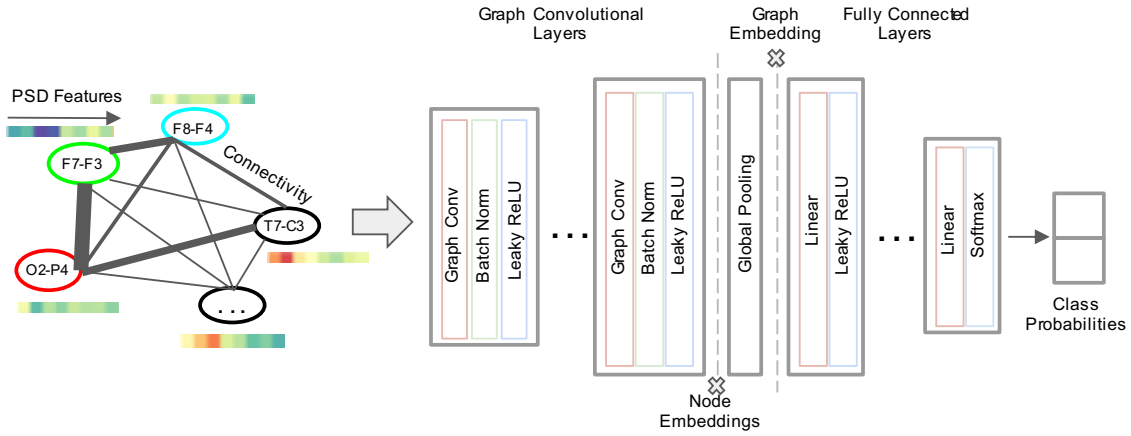

Figure 3: Graph representation of EEG data and the EEG-GCNN model architecture.

layer l (with $H^{(0)} = X_n$), and D denotes the diagonal degree matrix of graph G_n. Note that the degree matrix is trivial in our case because the graph is fully connected.

$$H^{(l+1)} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \qquad (2)$$

EEG-GCNN is aimed at graph-classification and therefore, node embeddings ($H^{(L)}$) are aggregated at the end of graph convolutions to form an embedding of the whole graph.

Window-level predictions: The graph embedding is then provided as input to a fully-connected network which produces output $Y_n \in [0, 1]$, which represents the probability that the n^{th} window was recorded from a patient with ND.

Deriving subject-level predictions: To estimate whether an entire EEG recording was recorded from a patient, we use a maximum likelihood estimation based on the window-level predictions. We model the window-level predictions of a subject S_i as independent observations made from a Bernoulli trial with an unknown probability π_i, where π_i is the probability that the subject S_i is a patient. Then, an estimate of π_i that maximizes the likelihood function $\prod_{n=1}^{N(i)} \pi_i^{Y(n)}(1-\pi_i)^{(1-Y(n))}$ after N windows is given as $\hat{\pi}_i = \frac{\sum_{n=1}^{N} Y_n}{N}$.

6. Evaluation Setup

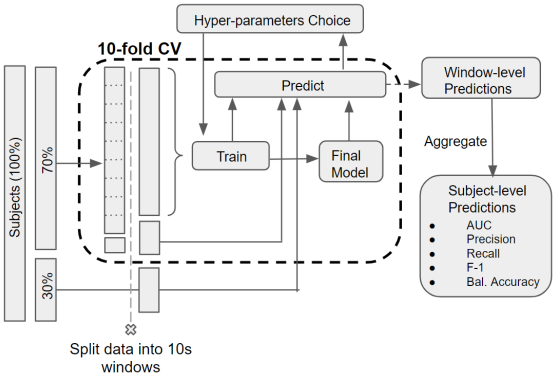

Figure 4: Evaluation procedure. We ensured that the data in training, validation, and testing sets were from disjoint sets of subjects.

The evaluation procedure is depicted in Figure 4. Table 1 summarizes divisions into training, validation, and testing sets. A 10-fold cross-validation (CV) routine was applied to the training set for each model to get robust estimates of model performance on validation and test sets. Note that each of the 10 CV folds maintained a disjoint set of participants in the resulting train and validation sets. This was done to ensure the validation performance reflects the performance on unseen subjects. Additionally, each model we compared was trained using the exact same 10 CV folds. We note that, a) the train-

Level	Train+Val (70%)	Test (30%)
Subjects	1,115 (P: 964, H: 151)	478 (P: 421, H: 57)
Windows	156,556 (P: 140,841, H: 15,715)	68,778 (P: 62,775, H: 6,003)

Table 1: Training, validation, and test sets. Abbreviations: P – Patient, H – Healthy.

ing and prediction are done at window-level, b) window-level predictions are aggregated to obtain subject-level predictions, and c) all CV folds share the same hyperparameters.

Implementation of ML models: We implemented several ML models for comparison as described in Section 7. The shallow variant of EEG-GCNN comprises only of 2 Graph Convolution layers (output dimensions: 64, 128) and a Global Mean Pooling layer. Notably, the shallow EEG-GCNN has no hidden Linear layers. In contrast, the deep variant of EEG-GCNN, is composed of 5 Graph Convolution layers (output dimensions: 16, 16, 32, 64, 128), a Global Mean Pooling layer, and 2 hidden Linear layers (hidden dimensions: 30, 20). The baseline Fully-connected Neural Network comprises 2 hidden Linear Layers (hidden dimensions: 64, 32). The Random forest baseline comprises 100 decision tree learners, with 4 features considered for node splitting, and a maximum tree depth of 15. The bagging fraction was set to 20% of the training set, and the complexity parameter for tree pruning was set to 0.015. For random forests, the predicted class probability of a sample is computed as the fraction of samples of that class present in the leaf node, averaged over all the trees in the forest. Model implementations were done using the Scikit-learn, PyTorch, and PyTorch Geometric libraries (Pedregosa et al. (2011), Paszke et al. (2019), Fey and Lenssen (2019)). The models were trained on a Linux server with 64GB memory and two NVIDIA Titan Xp GPUs.

Handling class imbalance: The imbalance in the training data was handled by using a weighted cross-entropy loss function for the neural network models and evaluating a weighted Gini impurity score during random forests tree construction. The class weights were set to the inverse of the window count for that class, leading to a higher loss value for mistakes on the minority class.

Evaluation metrics: Model generalization performance was evaluated based on the receiver operating characteristic (ROC) curve made from subject-level class probabilities (calculated by averaging window-level probabilities). We report the average AUC scores obtained across the 10 CV folds. Additionally, we pick an optimal decision threshold using Youden's J statistic (Youden (1950)) (i.e., the threshold that maximizes the sum of sensitivity and specificity) and use it to report the precision, recall, F-1, and balanced accuracy scores. While precision, recall, and F-1 scores are calculated treating the "patient" EEGs as the positive class, balanced accuracy is calculated as the average of the recall of each class.

6.1. Hyperparameter Tuning

In the following, we highlight the heuristics and strategies used in the tuning process. The exact values of the hyperparameters chosen for the final models can be found in the model definition files of released software.

Neural networks: The use of batch normalization made training vastly more robust to specific choices of hyperparameters. Nonetheless, layers with more trainable parameters required stronger dropout regularization. Layer depth was increased until performance gains were marginal. The initial learning rate of Adam optimizer was set to 0.1 and then decayed by a factor of 10 at regular intervals. The training process was vi-

sualized extensively to determine saturation and identify less favorable settings.

Random forests: Random forests was tuned through 3 rounds of iterative grid search (1512 configurations) on a fixed validation set. Starting from a wide uniform grid, next rounds focused on favorable grid regions. Performance was seen to be sensitive to 5 variables: total estimators in the forest, the maximum depth of tree, the cost-complexity pruning parameter, the bagging fraction, and the minimum number of samples at leaf nodes. The best performing configuration was then used in 10-fold CV.

7. Experiments & Results

Model comparisons: We employ the evaluation procedure discussed above to compare the performance of a shallow and deep EEG-GCNN architecture against two classical ML baselines - fully-connected neural networks (without graph convolutions) and random forests. The results of this comparison on held-out subjects are presented in Table 2 and Figure 5 displays the corresponding ROC curves. Our results indicate that GCNN-based approaches outperform both the ML baselines ($\sim 10\%$ improvement in AUC and balanced accuracy). Additionally, a Kolmogorov-Smirnov test between GCNN models and ML baselines rejected the null hypothesis that the subject-level probabilities (averaged across CV folds) are drawn from the same distribution ($p < 0.05$). This finding suggests that the differences in AUC are statistically significant. However, based on the optimal threshold chosen, our results indicate that random forests provides improved recall and F1 score despite providing lower AUC and precision.

To highlight chance-level performance and provide additional context for interpreting model comparisons in light of imbalanced data, we provide the performance of two trivial classifiers. We consider two blind classifiers i.e., the classifiers that were not trained with data: 1) a classifier that predicts the positive class with the label imbalance probability of $\sim$0.86, whose results are reported after 1000 simulations (referred to as trivial classifier 1) and 2) a classifier that always predicts the majority class (i.e., patient) regardless of the input (referred to as trivial classifier 2). While trivial baselines achieve a balanced accuracy and AUC of 0.50, the ML models, particularly EEG-GCNN, shows a marked improvement over all baselines.

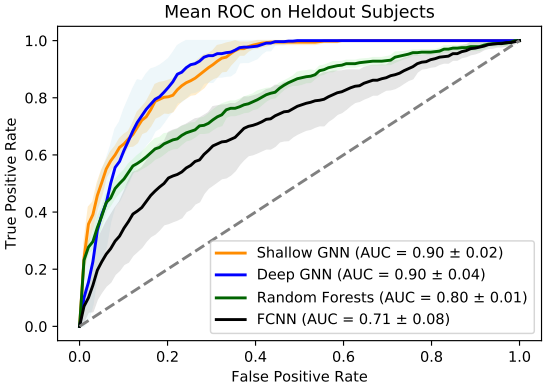

Figure 5: The mean ROC curves, where the shaded region indicates 95% confidence intervals.

Shallow or Deep Model: We evaluated the effect of GCNN network depth on performance with shallow and deep EEG-GCNN variants. Our results suggest that larger depth provides only a marginal improvement in performance over shallower variants. This observation is in line with multiple previous GCNN studies that report marginal reductions in performance with increasing network depth beyond just 2-4 layers (Zhao and Akoglu, 2019).

Large vs small training size: We observed an additional increase in performance of shallow EEG-GCNN when trained on only a tenth of the training data. This result is counter-intuitive in the deep learning paradigm and goes against conventional wisdom of increasing performance by using more

Model	AUC	Precision	Recall	F-1	Bal. Accuracy
FCNN	0.71 (0.08)	0.94 (0.02)	0.66 (0.11)	0.77 (0.08)	0.66 (0.07)
Random Forests	0.80 (0.01)	0.95 (0.01)	0.79 (0.08)	0.86 (0.05)	0.74 (0.02)
Deep EEG-GCNN	**0.90 (0.04)**	0.99 (0.00)	0.74 (0.08)	0.84 (0.06)	**0.85 (0.04)**
Shallow EEG-GCNN	**0.90 (0.02)**	0.99 (0.01)	0.72 (0.07)	0.83 (0.04)	0.83 (0.02)
Trivial Classifier 1	0.50 (0.02)	0.88 (0.01)	0.87 (0.02)	0.87 (0.01)	0.50 (0.02)
Trivial Classifier 2	0.50 (N/A)	0.88 (N/A)	1.00 (N/A)	0.94 (N/A)	0.50 (N/A)

Table 2: Results on the held-out set of 478 subjects. All metrics were calculated at the subject-level treating patient EEGs as the positive class. N/A indicates no variability.

data. However, a recent study (Nakkiran et al., 2019) shows that shallower models perform worse on larger data for a certain range of model complexity. While FCNN showed characteristics of overfitting, both shallow and deep EEG-GCNN models did not. Deep GNN remains unaffected while FCNN shows a drop in performance, as shown in Table 3.

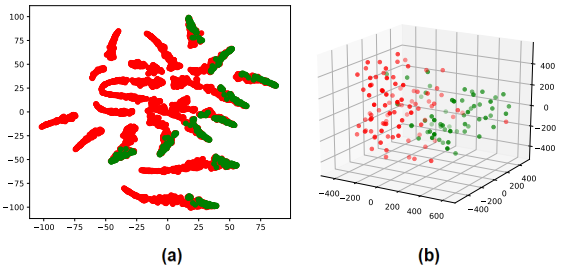

Figure 6: t-SNE maps of heldout test subject embeddings. Electronic zoom recommended for viewing. Green denotes a healthy subject, while red denotes a diseased patient. (a) 2D map of the embeddings from all 10 cross-validation models plotted together. (b) 3D map of one cross-validation fold.

Improved linear separability: Spectral features show faint distributional clues, if any, (Figure 2) while shallow EEG-GCNN (AUC: 0.90) dramatically outperforms a vanilla FCNN (AUC: 0.71). This is interesting because shallow EEG-GCNN does not include any hidden layers in the fully-connected part, suggesting that graph convolutions learn better feature representations. We perform a qualitative assessment of the 128-dimensional EEG-GCNN embeddings using the t-SNE algorithm (Maaten and Hinton (2008)), results of which are shown in Figure 6. Note that each point in the scatter plot is a subject (red denotes patient, green denotes healthy). Figure 6a shows the t-SNE maps of the embeddings from all 10 CV models plotted together, i.e., each subject is plotted 10 times, and Figure 6b shows the t-SNE maps of the embeddings of a single CV fold. We find that the EEG-GCNN embeddings, in general, show better separability compared to spectral features.

8. Discussion & Future Work

Conventional analysis of EEG relies on expert annotations of various phenomena (e.g., awake and sleep, artifacts, bad channels). Such annotations are time consuming, costly, susceptible to human error, and clearly not scalable. We presented a fully-automated approach based on graph neural networks that does not require expert annotations of specific events. As such, our approach provides multiple benefits: 1) it enables large scale studies, 2) it can eliminate individual biases, and 3) it can augment the visual review of EEGs by providing focused inputs and help reduce physician burnout (Verghese et al., 2018). Our future efforts will focus on developing end-to-end models with raw data/spectrograms as inputs and implementing saliency methods, as they can help identify novel EEG features and advance neurological disease research.

A limitation of our study is that the EEGs of two populations, healthy and pa-

Model	Train	Val	Test
FCNN	0.80 (0.08)	0.64 (0.03)	0.67 (0.07)
Shallow EEG-GCNN	0.96 (0.03)	0.94 (0.02)	**0.92 (0.03)**
Deep EEG-GCNN	0.93 (0.04)	0.92 (0.03)	0.90 (0.05)

Table 3: Training, validation, and test AUCs when trained using only a tenth of the data.

tients, were acquired using different systems under different conditions. Therefore, the difference between the acquisition systems/environments is a potential confounder in our analyses. To address this limitation as best we could, we undertook the same preprocessing steps for both the EEG datasets. Regardless, future studies including EEGs of both controls and patients recorded using the same acquisition system are necessary to eliminate this confounder and to elucidate the clinical value of our approach.

Furthermore, scalp-EEG provides a rich representation of the underlying brain state with substantial spatial and temporal granularity. However, the presence of artifacts makes the decoding of underlying brain state a nontrivial task. We postulate that the representation of EEG data learned by our model can be used to describe the underlying brain state and has potential utilities in other areas of brain research such as sleep staging, brain computer interfaces, and neural state decoding. In the future, we will also investigate the possibility of classifying various brain states (e.g., sleep stages) using the core methodology developed in this study.

Code & data availability: The datasets used in this study are already publicly available. The final trained models, data set metadata, and code to reproduce Table 2 results are available at https://github.com/neerajwagh/eeg-gcnn.

9. Conclusion

We introduce EEG-GCNN, a novel GCNN architecture for multi-channel EEG data inspired by the clinical significance of brain connectivity patterns in neurological disease pathology. We apply EEG-GCNN on a large dataset of "normal" EEG recordings from 1593 subjects and present strong evidence for the ability to distinguish between "normal" EEGs of neurologically diseased individuals and the EEGs of healthy individuals (AUC: 0.90). Clinical use of the superior predictive capability of EEG-GCNN can shorten the traditional diagnosis process and help expert neurologists make more accurate diagnoses.

10. Acknowledgements

We would like to thank our Mayo Clinic collaborators Gregory A. Worrell, Benjamin H. Brinkmann, and Brent M. Berry, for providing neurological domain expertise and validating the clinical significance of our proposed hypothesis. In addition, we would like to thank the Mayo Clinic Neurology Artificial Intelligence Initiative and the Mayo Clinic Illinois Alliance for Technology-based Healthcare Research, for providing financial and logistical support for this research.

References

Musaed Alhussein, Ghulam Muhammad, and M Shamim Hossain. Eeg pathology detection based on deep learning. *IEEE Access*, 7:27781–27788, 2019.

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral

physiology in young and old adults. *Scientific data*, 6:180308, 2019.

HK Bouma, C Labos, GC Gore, C Wolfson, and MR Keezer. The diagnostic accuracy of routine electroencephalography after a first unprovoked seizure. *European Journal of Neurology*, 23(3):455–463, 2016.

John S Ebersole and Robert F Leroy. Evaluation of ambulatory cassette EEG monitoring: III. Diagnostic accuracy compared to intensive inpatient EEG monitoring. *Neurology*, 33(7):853–853, 1983.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.

W Allen Hauser, V Elving Anderson, Ruth B Loewenson, and Stella M McRoberts. Seizure recurrence after a first unprovoked seizure. *New England Journal of Medicine*, 307(9):522–528, 1982.

June Hyun, Myung Jae Baik, and Ung Gu Kang. Effects of psychotropic drugs on quantitative eeg among patients with schizophrenia-spectrum disorders. *Clinical Psychopharmacology and Neuroscience*, 9 (2):78, 2011.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. Deep learning enabled automatic abnormal eeg identification. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2756–2759. IEEE, 2018.

Robin Tibor Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. *arXiv preprint arXiv:1708.08012*, 2017.

S J M Smith. Eeg in the diagnosis, classification, and management of patients with

epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2):ii2–ii7, 2005. ISSN 0022-3050. doi: 10.1136/jnnp.2005.069245. URL https://jnnp.bmj.com/content/76/suppl_2/ii2.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 2018.

Yogatheesan Varatharajah, Min Jin Chong, Krishnakant Saboo, Brent Berry, Benjamin Brinkmann, Gregory Worrell, and Ravishankar Iyer. Eeg-graph: A factor-graph-based model for capturing spatial, temporal, and observational relationships in electroencephalograms. In *Advances in Neural Information Processing Systems*, pages 5371–5380, 2017.

Yogatheesan Varatharajah, Brent Berry, Boney Joseph, Irena Balzekas, Vaclav Kremen, Benjamin Brinkmann, Gregory Worrell, and Ravishankar Iyer. Electrophysiological correlates of brain health help diagnose epilepsy and lateralize seizure focus. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3460–3464. IEEE, 2020.

Abraham Verghese, Nigam H Shah, and Robert A Harrington. What this computer needs is a physician: humanism and artificial intelligence. *Jama*, 319(1):19–20, 2018.

Xue-han Wang, Tong Zhang, Xiang-min Xu, Long Chen, Xiao-fen Xing, and CL Philip Chen. Eeg emotion recognition using dynamical graph convolutional neural networks and broad learning system. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1240–1244. IEEE, 2018.

W.H.O. *Neurological disorders: public health challenges*. World Health Organization, 2006.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. *arXiv preprint arXiv:1909.12223*, 2019.