Published in final edited form as:
Proc COMPSAC. 2020 July ; 2020: 664–673. doi:10.1109/COMPSAC48688.2020.0-182.

Regularization of Deep Neural Networks for EEG Seizure Detection to Mitigate Overfitting

Mohammed Saqib[1], **Yuanda Zhu**[2], **May Dongmei Wang**[1], **Brett Beaulieu-Jones**[3]

[1.]Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA

[2.]Electrical Engineering, Georgia Institute of Technology, Atlanta, GA

[3.]Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Abstract

Seizure detection is a major goal for simplifying the workflow of clinicians working on EEG records. Current algorithms can only detect seizures effectively for patients already presented to the classifier. These algorithms are hard to generalize outside the initial training set without proper regularization and fail to capture seizures from the larger population. We proposed a data processing pipeline for seizure detection on an intra-patient dataset from the world's largest public EEG seizure corpus. We created spatially and session invariant features by forcing our networks to rely less on exact combinations of channels and signal amplitudes, but instead to learn dependencies towards seizure detection. For comparison, the baseline results without any additional regularization on a deep learning model achieved an F1 score of 0.544. By using random rearrangements of channels on each minibatch to force the network to generalize to other combinations of channels, we increased the F1 score to 0.629. By using random rescale of the data within a small range, we further increased the F1 score to 0.651 for our best model. Additionally, we applied adversarial multi-task learning and achieved similar results. We observed that session and patient specific dependencies were causing overfitting of deep neural networks, and the most overfitting models learnt features specific only to the EEG data presented. Thus, we created networks with regularization that the deep learning did not learn patient and session-specific features. We are the first to use random rearrangement, random rescale, and adversarial multitask learning to regularize intra-patient seizure detection and have increased sensitivity to 0.86 comparing to baseline study.

Keywords

Seizure Detection; EEG; Deep Learning; Regularization

Corresponding author: May Dongmei Wang maywang@gatech.edu.

Appendix
The full code repository and scripts used to run the experiments is kept on https://github.com/Saqibm128/eeg-tasks.

I. Introduction

Epilepsy, a disease where patients suffer repeated seizures, is a major condition that affects more than 3 million US citizens [1]. Epilepsy can be caused from a variety of conditions but the resulting seizures are characterized by prolonged abnormal electrical activity in the brain which interferes with normal brain function and therefore normal function of the individual [2, 3]. A seizure can be caused by many root causes, such as concussions, fevers, photic stimulation or other possibly unknown factors [3]. These multiple factors interact and affect epilepsy and its prognosis, as shown in a recent study linking Sudden Unexpected Death in Epilepsy (SUDEP) with other top diagnostic codes from an insurance claims dataset [4]

Causes of epilepsy can vary, which can make treatment difficult, but a first diagnostic step is to use an electroencephalogram, also called an EEG, to view electrical activity through a series of electrodes placed on the scalp [2]. EEGs can reveal the specific location of a seizure, what the patterns of spread and retreat throughout a seizure episode are, how the seizure moves between different areas of the brain, and unique patterns before a seizure (i.e. pre-ictal) as well as during a seizure (i.e. ictal) [2]. Seizures can vary significantly from patient to patient, with various different structures, from focal spikes in patient-specific channels, to multiple spikes spreading across all channels being possible indications of seizure activity [5]. Though clinicians are trained to identify these markers, interrater agreement is near 46% on EEG, with sensitivities of 0.6 to 0.7, and between 4 to 8 false alarms per day [6]. Ultimately, only clinical observation of the patient in addition to EEG readings can prove a seizure versus a seizure-like state [5].

Seizure detection is a clinically relevant task due to the need for accurate algorithms capable of correctly identifying seizures in patients. Clinicians need to collect hours of EEG data for epileptic patients but have to annotate it in clinical practice to identify the best segments for further human analysis to inform future treatment [2]. When a seizure is automatically detected for patients, clinicians can intervene and prevent sudden unexpected death during the episode [7], which provides additional value. There has been interest in developing detection algorithms for these goals in the past. Consequently, a wide variety of human-engineered signal markers have been developed for automatic seizure detection, including entropy [8], coherence [9], and frequency and wavelet based features [10].

Besides signal markers, data-driven seizure detection has been explored in literature using various tools, including deep learning. Deep learning approaches with a 13 layer deep convolutional neural network (CNN) were able to achieve accuracies above 88%, with a sensitivity of 0.95 and specificity of 0.9 [11]. However, the researchers used a dataset of only 15 patients, which could mean that the model will not generalize to the larger populace [11]. Other researchers claimed results with sensitivity near 0.308 and a false alarm rate of 7 per day on the Temple University Hospital EEG Corpus (TUSZ) [12]. The authors claimed this dataset to be more challenging due to many hundreds more patients, lower signal to noise ratio (SNR), and more EEG channels per patients [12]. The study showcased the use of CNNs combined with a recurrent neural network (RNN) long short-term memory (LSTM) network and was shown to be applicable to other datasets as well.

In particular, seizure detection is challenging due to class imbalance and the relative uniqueness of the seizure phenotype between patients [13]. The relative rareness of events may increase false alarm rates, and in clinical environments, any alarm which has a high false alarm rate may be deemed useless and ignored. Additionally, for intra-patient seizure detection, varying seizure phenotypes may reduce the sensitivity of the algorithm since key biomarkers for detecting seizure in one patient of the testing set may not exist in the training set.

There has been significant work on seizure detection algorithms trained with the same set of patients compared to a different set of patients. In general, due to unique patterns and properties of EEG for each patients, creating an algorithm that can generalize to other groups can be challenging [13]. EEG tasks are harder to generalize between patients due to highly patient and session-specific patterns contained in the extremely data-dense EEG which can cause machine learning algorithms to fail. One work has suggested that subjects account for “32% of the variance, systems for 9% of the variance, and repeated sessions for each subject-system combination for 1% of the variance” in EEG records [14]. This means that previous best approaches have worked best with preexisting patient data.

Researchers in the past have circumvented issues around dependencies on patient and session bias on the data by creating patient-specific predictors that require previous data from patients. Researchers for one study were able to showcase that a patient-specific SVM predictor using hand-engineered features for a limited cohort of 16 patients could decrease false alarms to below 0.2 per hour for a seizure prediction task [10]. While patient specific predictors are clearly the state of the art, these models cannot be applied for patients not part of the initial train set.

Approaches that attempt intra-patient seizure detection explicitly have used multiple techniques to create an effective classifier. EEG researchers for many tasks have used data augmentation and regularization methods such as the addition of Gaussian noise and the use of overlapping windows [13], as well as the use of norm penalties on classifiers [15]. Regularization is an important tool for other EEG research fields as well. Researchers showcased the use of an adversarial multitask learning strategy for an EEG biometrics task to reduce session bias for patient identification [16], though there are no similar tasks for seizure detection in literature. Other researchers have used data augmentation by swapping hemispheres of channels symmetrically to create new examples [17].

In this work, we will compare different models and propose a novel deep learning framework for generalizable intra-patient seizure detection. We will use features automatically extracted primarily for seizure detection to create a session identification model and use its performance as a proxy for overfitting. In the past, we have already researched deep learning neural network for processing temporal dataset [18]. In this work, we extended this to continuous EEG waveform data. From literature and our first neural network project, generalization to patients outside of the train set is a hard task. We will tackle this challenge using unique strategies to reduce overfitting by targeting this unwanted bias. To our knowledge, this is the first use of random rescale, random rearrange, and adversarial learning for seizure detection in an intra-patient population.

We organize the paper as follows: in section 2, we introduce the Temple University Hospital EEG data set. In section 3, we discuss our preprocessing analysis, the initial setup for simple services, and the more in-depth strategies we applied for training our neural network. In section 4 and 5, we showcase results of both our traditional machine learning and deep learning experiments, a neural network without any regularization, and the results of our regularization techniques. In section 5 and 6 we discuss broader issues of why performance increased for certain approaches and how we plan to continue work.

II. Dataset

The Temple University Hospital EEG Corpus (TUSZ) is “the largest publicly available unencumbered dataset of EEG recordings”, containing thousands of EEGs in various reference systems [19]. This dataset consists of 592 patients, with 1185 collective EEG recording sessions. All EEGs were recorded with the 10-20 system, which is a 21-channel electrode format shown in Figure 1; each channel was sampled at 250, 256, or 512 Hz. The seizure corpus subset of the dataset includes labels which annotate each file over time for seizure sessions.

The Temple University Hospital EEG Corpus includes seizures from a variety of patients and is presented in a hierarchical format, as shown in Figure 2. Individual patients may come in multiple times to the hospital for multiple recording sessions, which are then split into consecutive non-contiguous token files of 20 to 60 minutes of recorded signal. With each token file, the dataset provides a time-indexed annotation labeling seizure and non-seizure events. Our approach included all the seizure types given within the dataset (generalized, focal, simple partial, complex partial, absence, tonic, clonic, tonic-clonic, atonic, and myoclonic). For this work, nevertheless, we simplify our scope to only detect seizure vs non-seizure classes, though there is a wide diversity of seizure phenotypes in the dataset.

EEG data has unique challenges to preprocess correctly, accentuated by the scale of our dataset. Extracranial EEG data signal is usually within the range of 20 to 100 μV [20], a much lower range than other bio-signals, which allows other noise to easily disrupt the signal. The relatively weak signal is susceptible to low frequency muscle artifact noise and higher frequency line noise [21]. In addition, the separation of the electrodes from the source of the voltage by layers of tissue, skull, and meninges prevents high frequency signal above 50 Hz from accurately measurement with extracranial electrodes [20].

III. Methodology

A. Preprocessing

Voltage is a difference of potential between points, so different specific baseline references can define different “montages” of EEGs. The TUSZ presents data from separate montage systems, including a linked-ear reference and an average reference system. We chose the average reference system due to its large volume of records existing in this set, more than the other reference sets available. We resampled all records to 250 Hz, which, by Nyquist Theorem, meant that we would be limited to frequencies below 125 Hz. The frequency of the EEG signal lies between 1Hz and 50 Hz, with line noise at 60 Hz and low frequency muscle and heart artifacts (EMG and ECG) interfering near 1 Hz [20]. Accordingly, we filtered out noise from below 1 Hz and above 50 Hz with a fifth order Butterworth bandpass filter.

We then segmented each file into either 1-second or 4-second segments of non-overlapping data from the record and assigned values of seizure or non-seizure based on the annotations. Segments immediately before or after a signal were removed due to possible information leaking or ambiguity of the signal, as shown in Figure 3.

We also chose to define train, validation sets from the already existing train set in the TUSZ (0.80, 0.20 proportions, respectively), since there was no existing validation set in the data. We used the predefined test set from the TUSZ as the test set for both the traditional ML and the deep learning approach.

B. Traditional Machine Learning Approach

We used random forest, XGBoost, and logistic regression for our traditional ML approach. We trained on hand-engineered features generated from the entirety of the 4-second segments such as frequency power from the Fast Fourier Transform (FFT), entropy, and coherence. Frequency power from the FFT was generated by gathering the total powers of sets of frequency ranges corresponding to alpha, beta, theta, and delta waves [22, pp. 10–13]. Coherence was chosen from pairs of channels to determine the overall sum of coherences. For our traditional approach, we used a random search of the hyperparameter space, with the best hyperparameters chosen based on the validation set.

C. Deep Learning Approach

We trained deep learning algorithm on 4-second segments only. We used multiple layers of CNNs followed by a Long Short-Term Memory layer (LSTM) followed by feedforward layers with dropout. We also added Gaussian noise of 2 μV as a standard data augmentation technique and applied re-referencing to each 4-second segment, which we found was similar to other previous approaches [13]. We remove outliers that are defined as a standard deviation of more than 100 μV. These outliers are too noisy to have detectable EEG patterns. Because we only have 21 channels, we also used max pooling only over the time segments to avoid removing information about channels that could have been used in upper layers of the CNN. In other words, we did not reduce the channel dimension during feature extraction while still reducing the time dimension. We used early stopping of 20 epochs and saved the best model based on the validation F1 score.

We used random down-sampling as part of the pipeline for the traditional machine learning algorithms to balance seizure classes; positive 4-second seizure segments made up by only 12% of the training set. Then for each epoch, we only kept 24% of the total data (12% seizure to 12% non-seizure). Regarding neural networks, we randomly resampled each epoch for class balance, while still presented as many examples as possible to the network.

D. Session Identification

CNN layers are known for extracting highly complicated features, which cause overfitting similar as other networks. We discovered that there existed significant overfitting on the data. Based on previous literature, we attempted to confirm whether the features extracted by the CNN specifically fitted to a patient/session/system combination unique to a specific recording session [14]. Then we added a session identification layer on a parallel neural network that shared the same CNN feature extraction layers as in the original neural network. We did a one out of 550 session identification on all sessions from the training set with this layer. We used sessions instead of patients because a patient may come in for a recording session to multiple wings and departments of a hospital, which could represent additional bias. However, a session is a unique combination of a patient at a specific time with a specific hospital department and could capture more of these over-specific biases. Layers for the session identification network were “frozen” except for the final identification layer. This was to prevent the shared feature extraction layers from updating and interfering with the seizure detection task. As a result, feature extraction should be primarily driven by the main network, while the secondary network would measure the incidental overfitting of the features. We directly measure if the feature set created by the neural network could be used to predict exact sessions from the training set as the network trained. Extremely high session identification indicates a failure to learn generalizable seizure detection features; if so, this could signify that we had extracted features geared more towards a specific patient population, instead of generalizable to a patient population outside of the training set. We presented all the final train session identification accuracy for each model in our results.

E. Regularization through Random Rescale and Random Rearrangement of Minibatch

Based on our initial CNN/LSTM experiment, we randomly shuffled the channels for each minibatch with our deep learning model. Because the order changed on each minibatch, there was no set order of channels that neural network could depend on to memorize. This also forced the network to learn multiple combinations of spatial features instead of a specific combination from a single ordering of EEG channels.

We then randomly scaled EEG channels on each minibatch within a fixed scale to create additional data as another experiment. We take a number within this specified range, defined as 1/x to x, and multiply all data in the minibatch to create new examples scaled to this random number.

We used these experiments to test whether data augmentation could provide additional data to assist the network generalize to a broader independent test set, and whether these methods would increase or decrease performance for patients not in the train set. An increase in performance would indicate effective regularization by targeting an aspect of EEG that is specific to patients and sessions while a decrease would indicate that we were targeting an aspect of EEG data that was already generalizable throughout the population.

F. Adversarial Multi-Task Learning

We created an adversarial multitask learner by attaching a layer for session identification with negative weight to the original neural network for seizure detection, as shown in Figure

4. This layer was updated with weights shared from the parallel non-interfering session identification network for each minibatch but then negatively back-propagated throughout the network, creating adversarial training on each mini-batch for seizure detection and against session identification for the feature extraction CNN layers. We attempted multiple different seizure and negative session weightings to see whether we could reduce overfitting experienced by the network. We wanted to test if there with differences in how extracted features would generalize for seizure detection in comparison to how much they had fit on the know train set session specific features.

A higher seizure weight ratio should lessen the adversarial effect, while a lower weight ratio should emphasize feature extraction that cannot incidentally identify sessions. Our paper is the first to attempt this on seizure detection, against patient-specific features to try to generalize to a broader patient set.

IV. Traditional Learning Results

We decided to use a 1-second segment first for our traditional learner and compared to a 4-second segment for the seizure detection task, as shown in Table I and Table II. We chose the classifier with the best F1 score and extracted feature importance in Figure 5.

There was surprisingly little to no change in the traditional classifiers when presented with features extracted from the 4-second segments. The inability of the classifiers to learn higher granularity features and the reliance of features over the entire segment instead of smaller events within the segment caused lower performance. A more complicated algorithm for longer segments is likely necessary to capture the additional detail. XGBoost is still superior in performance for the 4-second data but degrades with respect to F1 score from 1-second data.

V. Deep Learning Results

We used a CNN-2D LSTM model, as shown in Figure 6 for all classifiers. We also include the final session identification for training sessions in our train set as an indication of overfitting of the model. Results without regularization were disappointing due to overfitting, as shown in Table III. As shown in Figure 7, the classifier overfits with very little increase in validation accuracy. The final training session identification accuracy, as shown in Table III, was 0.62, which further suggests that this model overfit. Our experiments afterwards attempted to target session identification as a proxy for overfitting. If we were able to decrease it, we would expect the F1 score to increase as well.

A. Randomly Rearranging Minibatch

Using random rearrangement can increase F1 scores in a test set. Only when we attempt to randomly rearrange the order of the channels did we create a classifier which could approach the traditional classifier with respect to F1 score. The training curve increases monotonically to an extremely high F1 score in Figure 8, while the validation curve appears to increase somewhat monotonically until epoch 10, unlike in Figure 7.

As shown in Table IV, final train session identification accuracy has decreased by 40% from the network without random rearrangement, which further indicates that we were able to regularize efficiently. Furthermore, sensitivity increased by 0.24, and AUC increased by 0.07 though specificity decreased by 0.09.

B. Randomly Rearranging Minibatch + Random Rescale Each Minibatch

We also used a random rescale on each minibatch in a range limited between 1/ (rescale factor) to (rescale factor). We use this range such that higher factors represent more extreme deviations from the original data. We do this to minibatches as a whole and samples to keep relative relationships between channels. We present results for a neural network using both random rescale on each minibatch and a randomly rearranged channel in Table V.

The best classifier in terms of F1 score was a classifier with a rescale factor of 1.5, which meant that each mini-batch was altered for each epoch by a random rescale factor ranging from

1-(1/1.5) to 1.5. However, this classifier also had the lowest specificity, which would increase false alarm rates in patients. For a classifier using both randomly rearranged channels for each minibatch, and a randomly rescaled signal, we found a learning curve with a validation F1 score which increases rapidly before settling on chaotic behavior after 10 epochs, in Figure 9.

We also considered what may have caused an increase of random rescale factor from 1.5 to 2 to decrease F1 score performances. As shown in Figure 10, the neural network may begin to have overfit again, as evidenced by the increase in session identification accuracy.

With both random rescale and random rearrange on each minibatch, there is a small but clear increase in performance with a rescale factor of 1.5 (minibatches scale randomly between 1/1.5 and 1.5). This represented the best model we were able to train.

C. Adversarial Multitask Learning

As another investigation into whether it would be possible to reduce overfitting through regularization, we also attempted an adversarial multitask learning task, with results in Table VI. We attempted various combinations of loss weights for session identification and seizure detection to see if there were major differences in seizure detection performances. The adversarial model was able to use stronger relative adversarial weight (negative session weight) to reduce session identification. This is reflected in the table as the session identification accuracy decreasing as the relative weighting of the adversarial main task to the seizure detection task increases. However, F1 score did not appear to increase as much with our approach as with randomly rearranging the channels on each minibatch. There is one result at a seizure weight of 10 and session weight of −1 which has a high F1 score that may represent an optimal balance. This may mean that the network may use the data augmentation from random rearrangement more effectively than with an adversarial multitask model.

D. Adversarial Multitask Learning with Randomly Rearranged Channels

We also used adversarial multitask learning while using a randomly rearranged channel, as shown in Table VII. The two techniques together appeared to temper the effect of adversarial training and increase session identification even though both approaches were used. Even though session identification in the train set does not go below 0.3, the F1 score in the test set does not degrade much for many of our experiments.

Using random rearrangement shows improvements to seizure detection F1 score for every combination of seizure weight and adversarial session identification weight. However, session identification accuracy increases more than if random rearrangement or adversarial training were used separately. Furthermore, the F1 score does not increase beyond our best classifier using only random rescale and random rearrange, though specificity shows a clear improvement.

E. Adversarial Multitask Learning with Randomly Rearranged, Randomly Rescaled Channels

Finally, we also used both random rescaling and random rearrangement of channels on each minibatch to change F1 score performance as shown in Table VIII. Using an adversarial multitask learning system does not dramatically increase or decrease performances. There are only small changes in overall results depending on the weightings of seizure detection and adversarial session identification. For some sample learning curves, there appears to be a period where session identification can remain relatively suppressed, as shown in Figure 11. We found that the feature set extracted by the CNN layers does eventually increase the session identification score as the network converges. In fact, the network with the highest F1 with all three regularization strategies had a session identification accuracy of 0.482.

VI. Discussion

Our work is the first, to our knowledge, to attempt several new techniques with regards to seizure detection:

- Use of session identification to measure overfitting from patient and session specific biases
- Use of random rearrange as a form of data augmentation
- Use of random rescale as a form of data augmentation
- Use of adversarial multitask learning to target session identification

We used traditional learners to set a baseline for seizure detection F1 scores and to find the most important features. The results of the traditional learner in the literature suggested that coherence between pairs of channels could indicate seizure, as an unhealthy correlated state of certain brain areas [9]. These are spatial features dependent on channels, which represent interactions between physical brain areas separated throughout space. We also found that ensemble methodologies such as RF and XGBoost easily outperformed linear methodologies such as Logistic Regression, as well as our non-randomly rearranged CNN and could approach high F1 scores of our other CNN models. Ensemble learners are known for reducing overfitting through weak learners trained on subsets of the data, which may explain the performance increase.

Based on our initial deep learning experiment, we were surprised to see that the CNN could extract features for a 1 of 550 session identification task, despite being explicitly trained for seizure detection. Our result indicated that generalization to another set outside of the train set would be a key challenge for CNN feature extraction layers. For example, our first baseline network in Figure 7 failed to generalize because it extracted session and patient specific features not generalized seizure detection features. We saw both a high session identification accuracy of 0.602 and a dismal performance for seizure detection with an F1 score of 0.56 in Figure 7. The high training session identification was a target to reduce overfitting and helped inform later approaches.

Session identification training was an important tool to leverage throughout our experiments for reducing the type of overfitting we saw with the first neural network. Traditionally, overfitting can be detected for deep networks when the validation curve stopped rising, but for the TUSZ, due to the approach we took, we could see overfitting immediately after even a few batches. We believe that reusing the same sessions and generating multiple 4-second segments from each may have caused overfitting within a batch. The network can be presented many multiple examples of the same session taken at differing times but with similar identifying features. Even if the segments were non-overlapping and could represent different aspects of EEG overtime, segments from the same session were not independent.

Many of our most performant experiments had lower session identification in exchange for better generalization for seizure detection. We could see that for these models, as shown in Figure 8, the validation score can increase more with each additional epoch compared to with no technique targeting session identification. To accomplish this goal, we reduced session identification capability for the extracted feature set using techniques such as random rearrange and random rescale. In addition, we found that adversarial multitask learning is an important tool for directly affecting session identification. Such techniques aimed to identify the signal that could have been used to uniquely identify EEG segments specific to a session. We were especially interested in seeing whether seizure detection performance would suffer if we altered an EEG with these methods.

We first directly changed the data to regularize the network. We began with an approach of random rescaling and random rearrangement of the channel input to partially “destroy” aspects of input data that caused overfitting, such as a maintained channel order and a maintained EEG magnitude. Intuitively, directly increasing random aspects of the signal should reduce performance for both seizure detection and session identification. Instead, we found experimental results that we forced the network to consider new examples not in the original dataset, while removed identifying features that could correlate with any specific session. This increased seizure detection sensitivity and F1 score.

The significant decrease of session identification accuracy from Table III to Table IV suggests that random rearrangement is clearly the most performant regularization strategy due to its presentation of spatial features that may not be originally represented in the unchanged data. A single patient may only have seizures localized as a series of patterns in a few distinct channels, but random rearrangement helps replicate this pattern in multiple combinations that can generalize to other patients with seizures in completely different channels. This increases sensitivity of seizure detection classifiers in our test set, which is the primary driver of increase in our F1 score.

We based randomly rescaling minibatches as another form of regularization in combination with random rearrange that uses data augmentation. Our F1 score for seizure detection increased from 0.63 to 0.65, suggesting that the two approaches increased regularization together (at a rescale factor of 1.5). Furthermore, we reduced session identification to 0.08 with this rescale factor. However, the session identification and seizure detection performance are highly sensitive to the rescale factor used. For example, with a rescale factor of 2, (which allows minibatches to be randomly scaled between 0.5 and 2 for every epoch), the train session identification appears to have increased to 0.541 compared to lower values of rescale, while decrease seizure detection F1 score of 0.62. This may mean that features for seizure detection may have been affected more than features for session identification, which may have caused the network to extract the latter.

We applied random rescale and random rearrange as effective means to reduce overfitting by extracting spatially and session invariant features from the input data. On the other hand, our adversarial learning approach attempted to extract patient-invariant and session-invariant features to improve generalization. Instead of directly manipulating the input data, we affected the process by which the neural network trained by penalizing increases in session identification. We originally believed that all these regularization methods could help increase F1 seizure detection scores, but we found that F1 scores remained close to 0.6 for most of our choices of hyperparameters. However, adversarial multitask learning did increase specificity compared to use of random rescale and random rearrange, without drastically reducing F1 score.

Our regularized seizure detection models are competitive in sensitivity and F1 score. We choose the work from Golmohammadi as a baseline, as they worked on the same task with the same dataset. The researchers claim that models trained on the TUSZ have lower performance compared to other datasets due to issues in the data “representative of common clinical issues”, so a comparison to their results is most appropriate [11]. The researchers could only achieve a sensitivity of 0.30 when they optimize for the highest specificity of 0.97 and achieved a sensitivity of 0.39 when they allowed a lower specificity of 0.76 [11]. Our processing pipeline and application of regularization could increase sensitivity in the intra-patient seizure detection to 0.86 compared to this baseline. Unfortunately, our specificity is not as competitive, which means that our models will output more false alarms than some of the approaches in literature. This decrease in specificity may be due to increase in sensitivity for such a rare event. This lower specificity compared to other models may also be due to using only 4-second segments instead of longer segments. Accordingly, we believe that we can increase this in future experiments.

We believe further investigation into other labels in the dataset describing key aspects of the seizure events, such as the seizure subtype and the text clinical notes could help provide further approaches in the future to combat overfitting and to continue work with remaining challenges. There needs to be further consideration on what would be the best approach for increasing specificity and therefore decreasing false alarms for patients outside the train set, as well as what is practically possible for intra-patient seizure detections. Future work shall also include more data augmentation and regularization methods to improve generalization, other segment sizes instead of just 4 seconds, and additional preprocessing to further clean data. Additional analysis on spatial dependencies is needed to investigate features learned and extracted by a CNN classifier. To be more specific, features learned and extracted from fixed order channels shall be compared against randomly rearranged channels on each minibatch of the input data. We suggest that such an experiment will confirm if overfitting occurs if features are uniquely identifiable to single or groups of patients only in the train set.

Nonetheless, we were able to create performant seizure detection algorithms, despite numerous challenges. We processed hundreds of EEG signals from various patients and applied both hand engineered and CNN features to various models to work towards seizure detection. We also used new regularization methods to solve our models' failure to generalize to other patients. We saw that spatial features could be important for seizure detection in traditional ML algorithms but could also be overfit on with deep learners. We found and tackled intra-patient variation from patients we could not train on by using the session identification score of patients we did train on and used random rearrange, random rescale, and adversarial training to reduce it.

VII. Conclusion

We showed that spatial dependencies are among the most important features learned by traditional ML techniques for EEG seizure detection analysis. However, these features could cause overfitting in a deep neural network by becoming associated with a specific patient. Session identification during training helps provide an additional view into overfitting on session specific features but is not the only cause of overfitting. We also used regularization beyond measuring the session identification. We created spatially and session invariant features by enforcing our networks to rely less on exact combinations of channels and signal amplitudes, but instead to learn new dependencies towards seizure detection. We are the first to use random rearrangement, random rescale, and adversarial multitask learning to regularize intra-patient seizure detection, and have increased sensitivity to 0.86 from a baseline study with a slightly lower specificity. We discovered that combining random rescale could further increase performance. However, we also found that adversarial learning was not effective in combination with the other regularization methods. Further analysis remains as future work.

Our experiments confirm findings from literature that the increased noise and the high inter-session variability can cause the deep neural network to easily overfit on feature-rich data. In particular, the TUSZ is among the more challenging datasets sourced from a real-world environment. Thus, proper use of regularization can help mitigate many issues and increase sensitivity for seizure detection.

Acknowledgment

We want to acknowledge Bio-Medical Informatics and Bioimaging Lab for providing significant support for the project. We want to acknowledge DBMI and the summer internship at SIBMI for providing an extraordinary opportunity to continue work on this project and allowing us to expand the scope. We also want to acknowledge Dr. Mauricio Villamar and Dr. Kasper Claes for first-hand knowledge of the clinical flow for epilepsy and for additional direction and advice during this project.

This research was funded by NIH grant U54HG007963, T15LM007092, and NSF EAGER1651360, Children's Healthcare of Atlanta and Georgia Tech Partnership Grant, Petit Institute Faculty Fellow Fund, and Carol Ann and David D. Flanagan Faculty Fellow Research Fund to May D Wang

References

[1]. Zack MM and Kobau R, "National and state estimates of the numbers of adults and children with active epilepsy—United States, 2015," MMWR Morb. Mortal. Wkly. Rep, vol. 66, no. 31, p. 821, 2017. [PubMed: 28796763]

[2]. Noachtar S and Rémi J, "The role of EEG in epilepsy: A critical review," Epilepsy Behav., vol. 15, no. 1, pp. 22–33, 5 2009, doi: 10.1016/j.yebeh.2009.02.035. [PubMed: 19248841]

[3]. Fisher RS et al., "ILAE official report: a practical clinical definition of epilepsy," Epilepsia, vol. 55, no. 4, pp. 475–482, 2014. [PubMed: 24730690]

[4]. Zhu Y, Wu H, and Wang MD, "Feature Exploration and Causal Inference on Mortality of Epilepsy Patients Using Insurance Claims Data," in 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), 2019, pp. 1–4, doi: 10.1109/BHI.2019.8834638.

[5]. Fisher RS, Scharfman HE, and deCurtis M, "How can we identify ictal and interictal abnormal activity?," Adv. Exp. Med. Biol, vol. 813, pp. 3–23, 2014, doi: 10.1007/978-94-017-8914-1_1. [PubMed: 25012363]

[6]. Amorim E et al., "Performance of spectrogram-based seizure identification of adult EEGs by critical care nurses and neurophysiologists," J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc, vol. 34, no. 4, pp. 359–364, 7 2017, doi: 10.1097/WNP.0000000000000368.

[7]. Van de Vel A et al., "Non-EEG seizure-detection systems and potential SUDEP prevention: State of the art," Seizure, vol. 22, no. 5, pp. 345–355, 6 2013, doi: 10.1016/j.seizure.2013.02.012. [PubMed: 23506646]

[8]. Srinivasan V, Eswaran C, and Sriraam N, "Approximate Entropy-Based Epileptic EEG Detection Using Artificial Neural Networks," IEEE Trans. Inf. Technol. Biomed, vol. 11, no. 3, pp. 288–295, 5 2007, doi: 10.1109/TITB.2006.884369. [PubMed: 17521078]

[9]. Mormann F, Lehnertz K, David P, and Elger CE, "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients," Phys. Nonlinear Phenom, vol. 144, no. 3–4, pp. 358–369, 10 2000, doi: 10.1016/S0167-2789(00)00087-7.

[10]. Shoeb A, Bourgeois B, Treves ST, Schachter SC, and Guttag J, "Impact of patient-specificity on seizure onset detection performance," presented at the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007, pp. 4110–4114.

[11]. Acharya UR, Oh SL, Hagiwara Y, Tan JH, and Adeli H, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," Comput. Biol. Med, vol. 100, pp. 270–278, 9 2018, doi: 10.1016/j.compbiomed.2017.09.017. [PubMed: 28974302]

[12]. Golmohammadi M, Ziyabari S, Shah V, Obeid I, and Picone J, "Deep Architectures for Spatio-Temporal Modeling: Automated Seizure Detection in Scalp EEGs," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 745–750, doi: 10.1109/ICMLA.2018.00118.

[13]. Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, and Faubert J, "Deep learning-based electroencephalography analysis: a systematic review," ArXiv190105498 Cs Eess Stat, 1 2019 [Online]. Available: http://arxiv.org/abs/1901.05498. [Accessed: 25-Aug-2019]

[14]. Melnik A et al., “Systems, Subjects, Sessions: To What Extent Do These Factors Influence EEG Data?,” Front. Hum. Neurosci, vol. 11, p. 150, 2017, doi: 10.3389/fnhum.2017.00150. [PubMed: 28424600]

[15]. Zhao Y et al., “Improving Generalization Based on l1-Norm Regularization for EEG-Based Motor Imagery Classification,” Front. Neurosci, vol. 12, p. 272, 2018, doi: 10.3389/fnins.2018.00272. [PubMed: 29867307]

[16]. Ozdenizci O, Wang Y, Koike-Akino T, and Erdogmus D, “Adversarial Deep Learning in EEG Biometrics,” IEEE Signal Process. Lett, vol. 26, no. 5, pp. 710–714, 5 2019, doi: 10.1109/LSP.2019.2906826. [PubMed: 31814690]

[17]. Deiss O, Biswal S, Jin J, Sun H, Westover MB, and Sun J, “HAMLET: Interpretable Human And Machine co-LEarning Technique,” ArXiv Prepr. ArXiv180309702, 2018.

[18]. Saqib M, Sha Y, and Wang MD, “Early prediction of sepsis in EMR records using traditional ML techniques and deep learning LSTM networks,” presented at the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 4038–4041.

[19]. Obeid I and Picone J, “The Temple University Hospital EEG Data Corpus,” Front. Neurosci, vol. 10, 2016, doi: 10.3389/fnins.2016.00196. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2016.00196/full. [Accessed: 17-Aug-2019]

[20]. Malmivuo J and Plonsey R, “Bioelectromagnetism,” Med. Biol. Eng. Comput, vol. 34, pp. 9–12, 1996. [PubMed: 8857306]

[21]. Puce A and Hämäläinen MS, “A Review of Issues Related to Data Acquisition and Analysis in EEG/MEG Studies,” Brain Sci., vol. 7, no. 6, 5 2017, doi: 10.3390/brainsci7060058. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5483631/. [Accessed: 30-Jan-2020]

[22]. Sanei S and Chambers JA, EEG signal processing. John Wiley & Sons, 2013.

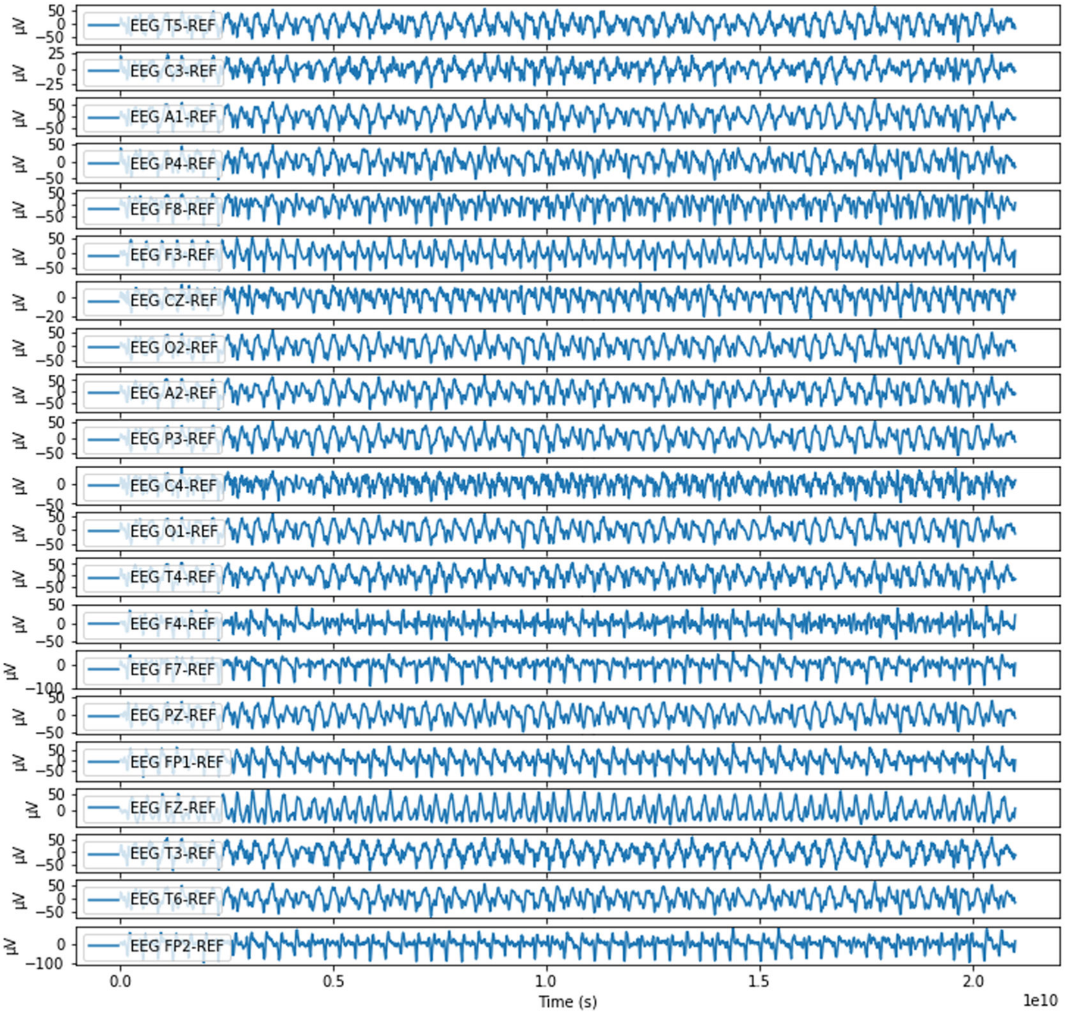

Figure 1:

An example 20 second record of a patient undergoing generalized seizure. The data is a highly dense, highly sampled time series of 21 interrelated channels; both spatial and temporal dependencies are important factors to consider.

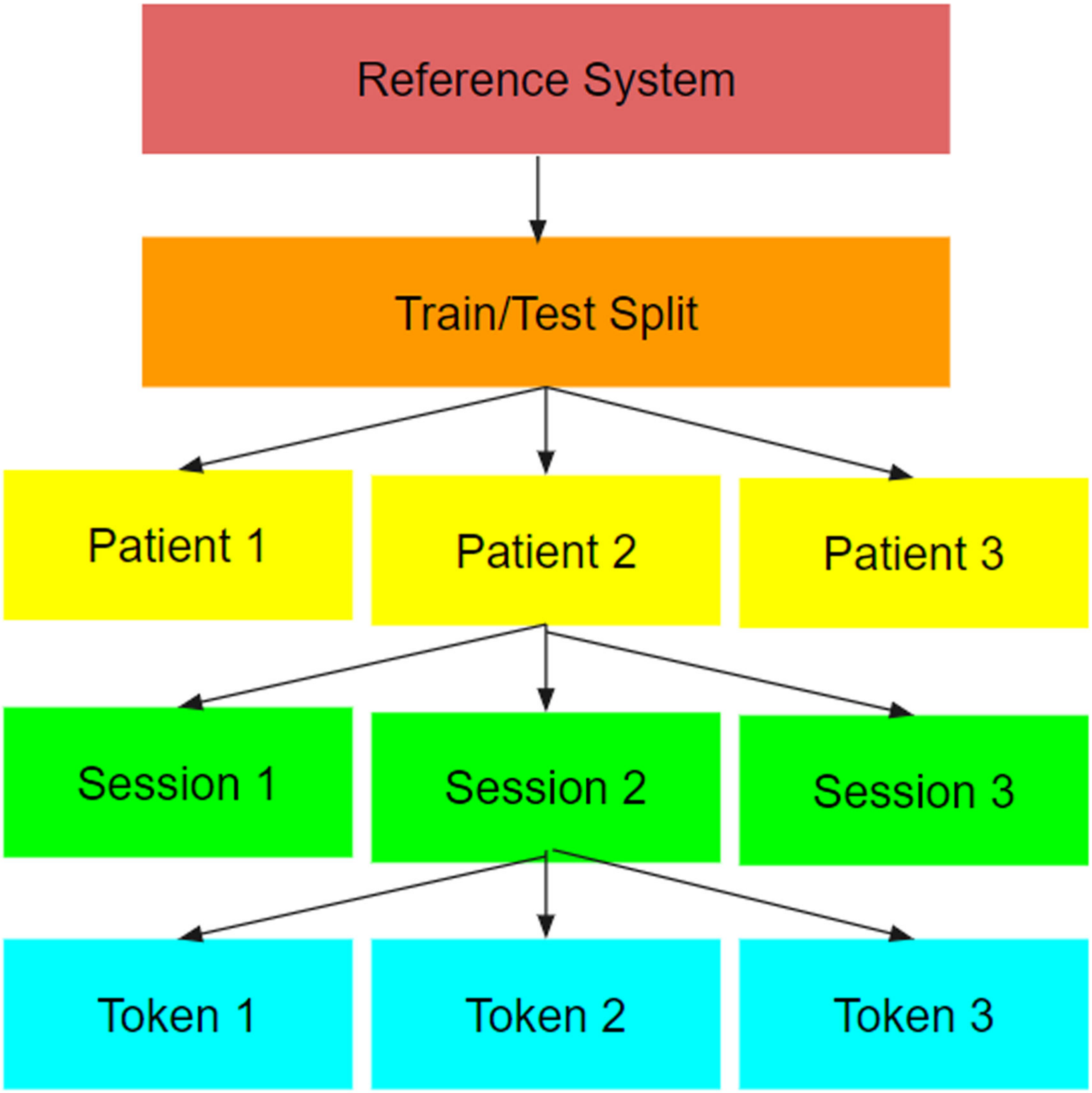

Figure 2:
A simplified view of the hierarchy of structures that we read from. This hierarchy splits data by the reference system used to gather the data, followed by a train test set, followed by patients, then sessions, then individual token files.

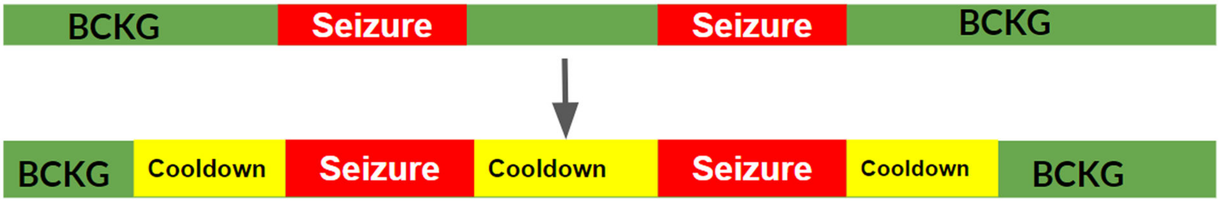

Figure 3:

All EEG files came with time-based annotations to identify if/when a seizure occurred in the segment, which we alter slightly.

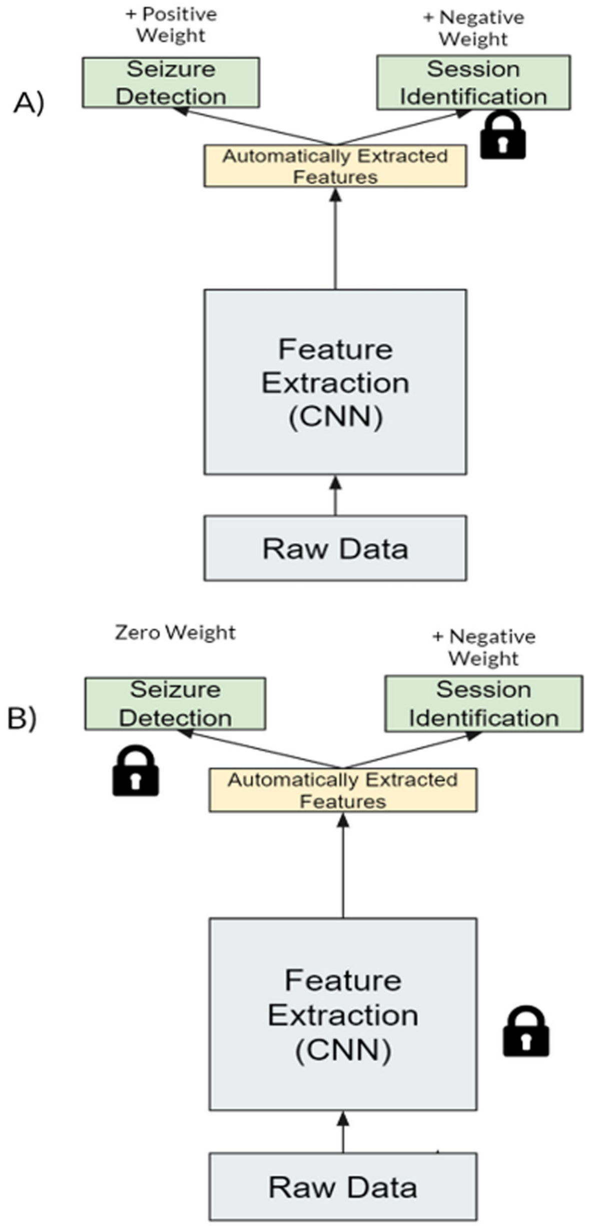

Figure 4:

Adversarial multi-task learning approach for seizure detection. We used two neural networks with two task, weights that shared and updated between the networks, and “frozen” layers forced to remain the same after each minibatch update to create a multi-task adversarial network. A) The first network attempts to update the feature extraction for seizure detection and against session identification without updating the weights for the final session identification layer. B) The second network updates session identification weights by predicting positively for sessions, without changing the other weights.

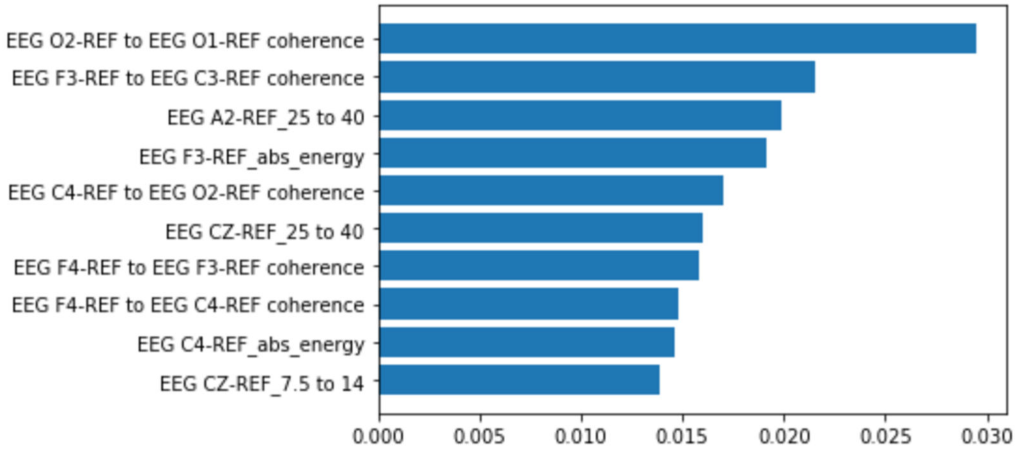

Figure 5:

Feature importance extracted for all features from the best 1 second XGBoost classifier. Coherence measurements were among the most important features for predicting seizures in this classifier

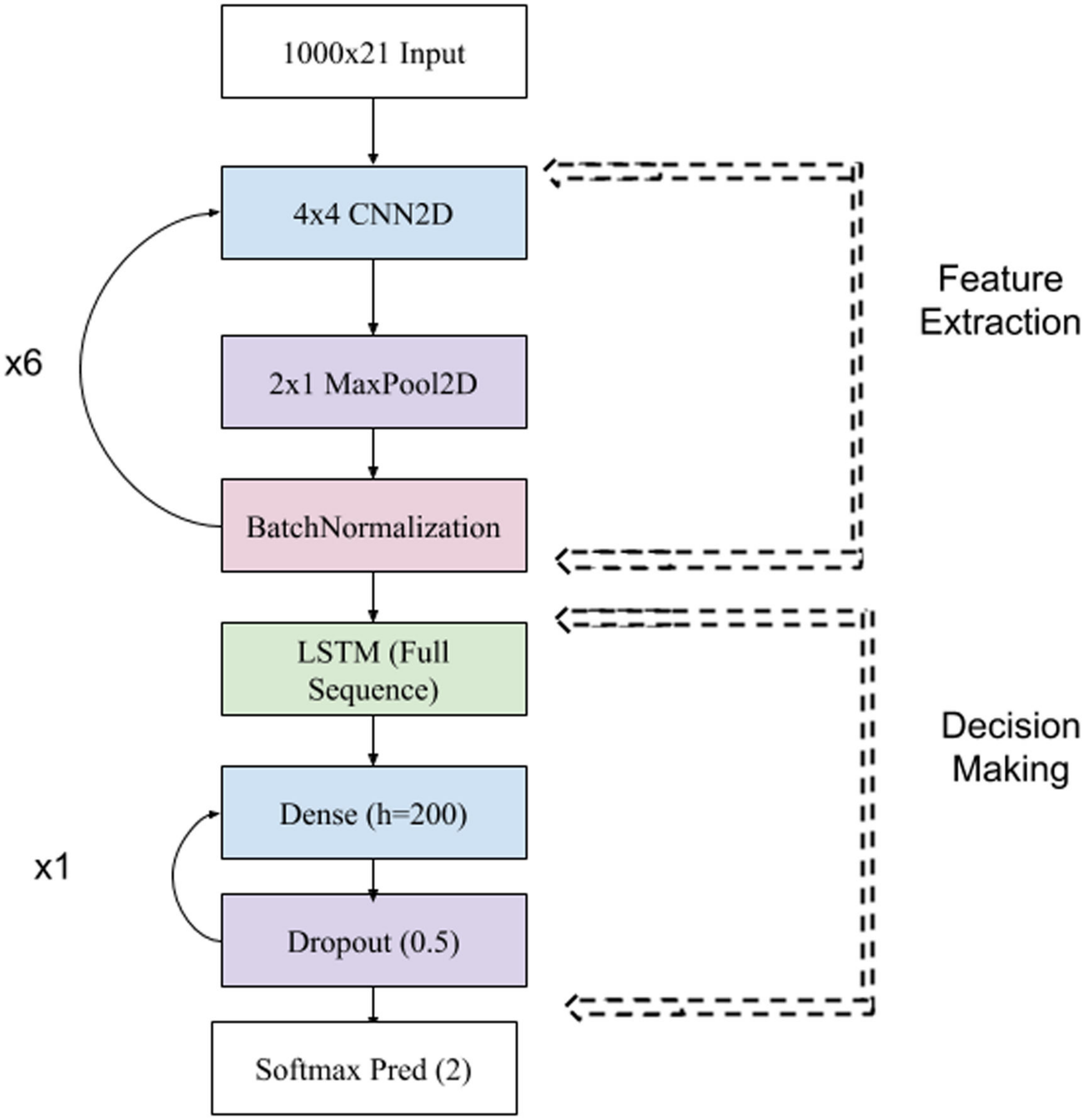

Figure 6:
Architecture of our CNN2D/LSTM. We ran multiple architectures but found that the following had consistently strong results. We used ADAM with an initial learning rate of 0.0005. We found in practice that our networks were most likely to converge if we also applied a learning rate decay of 0.9 each epoch. We used an RELU activation function except for the final layer, where we used softmax. The network trains in parallel another session identification dense layer (not shown), which connects to the feature extraction layers of the network and which is prevented from updating weights to the network.

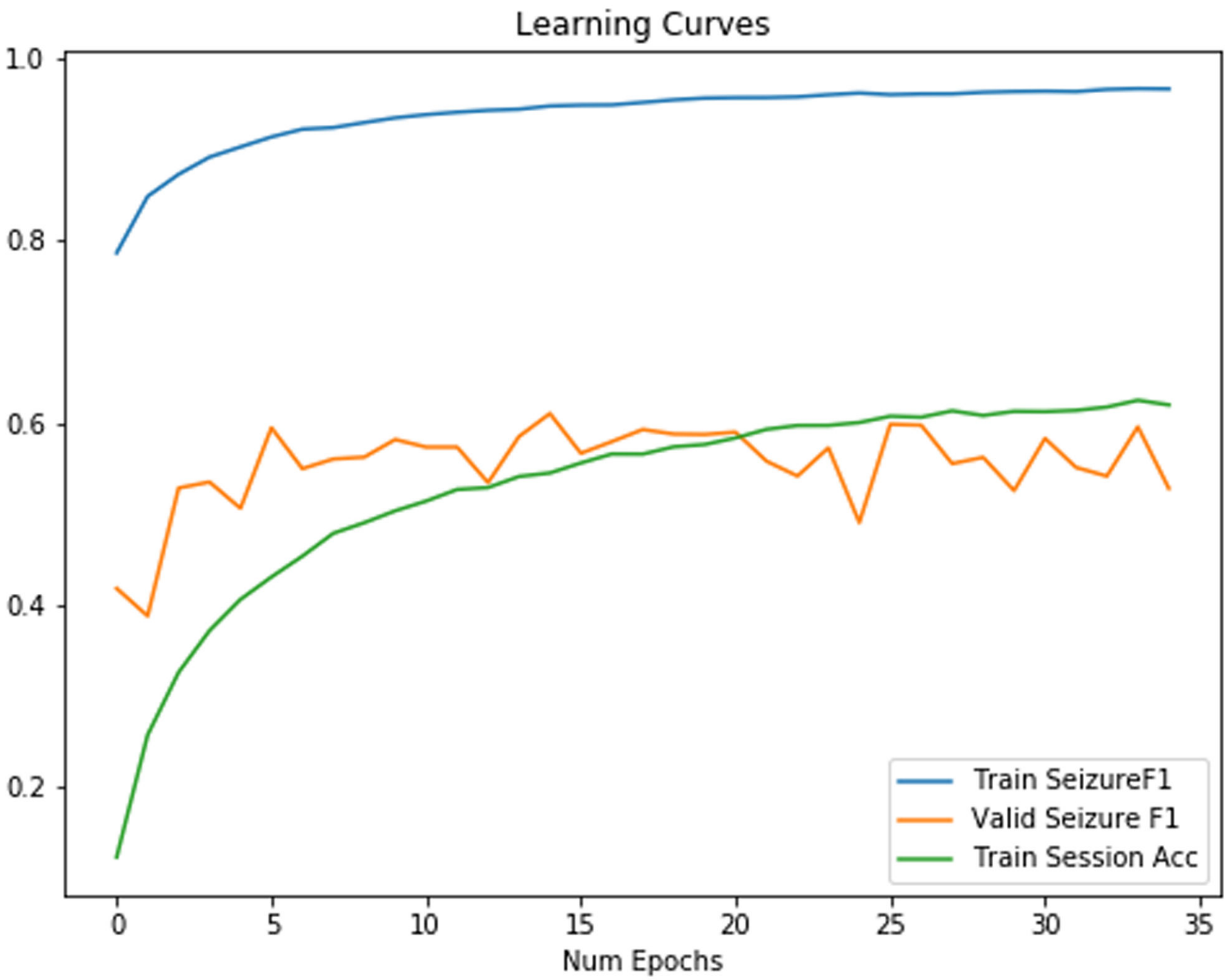

Figure 7:

Training/Validation F1 Score for a baseline neural network. Session identification accuracy increases while valid seizure F1 fails to increase.

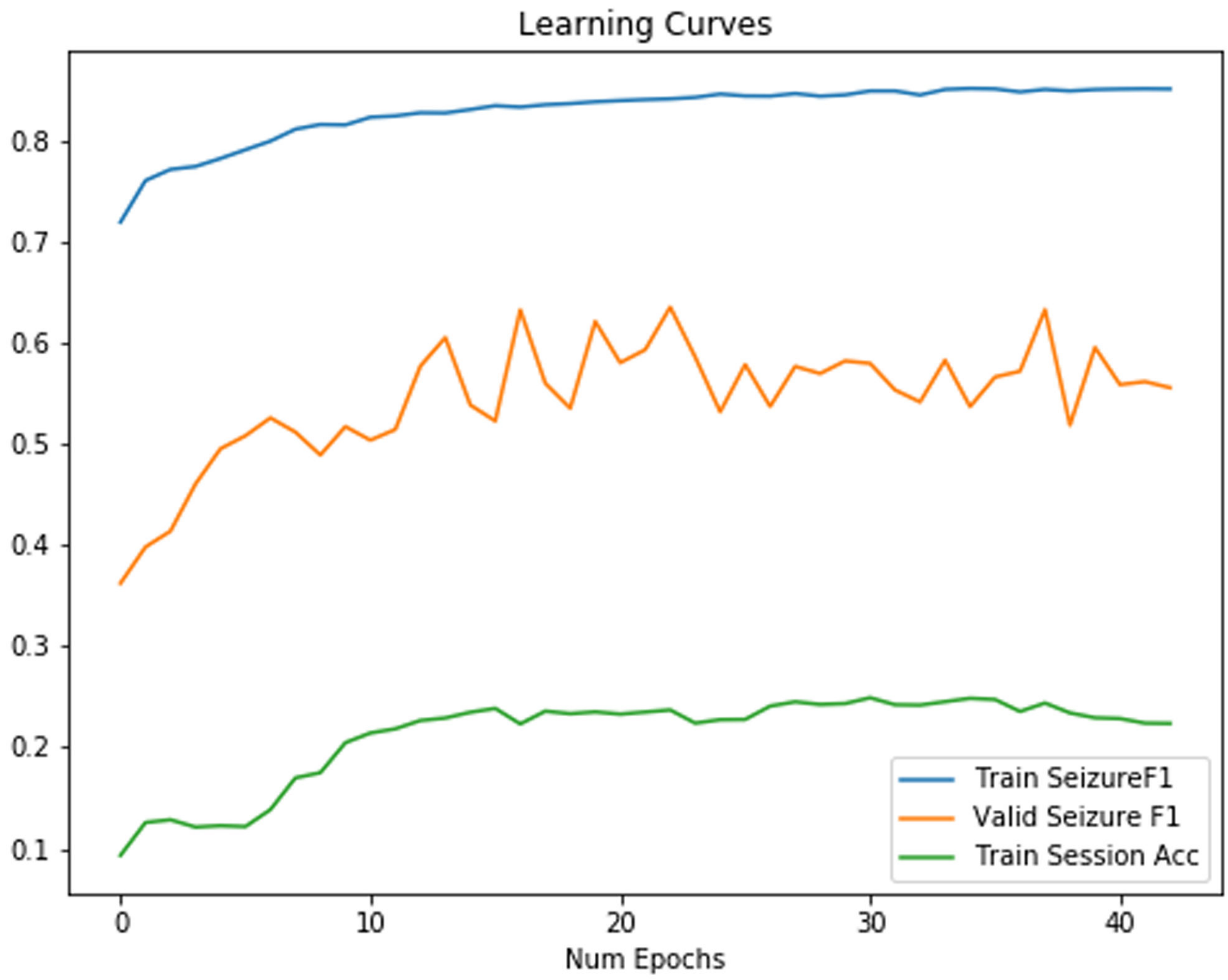

Figure 8:
Training/Validation F1 Score for a neural network with randomly rearranged channels on each minibatch. We found that our CNN2D/LSTM model was able to generalize somewhat to a set of patients outside of the training set.

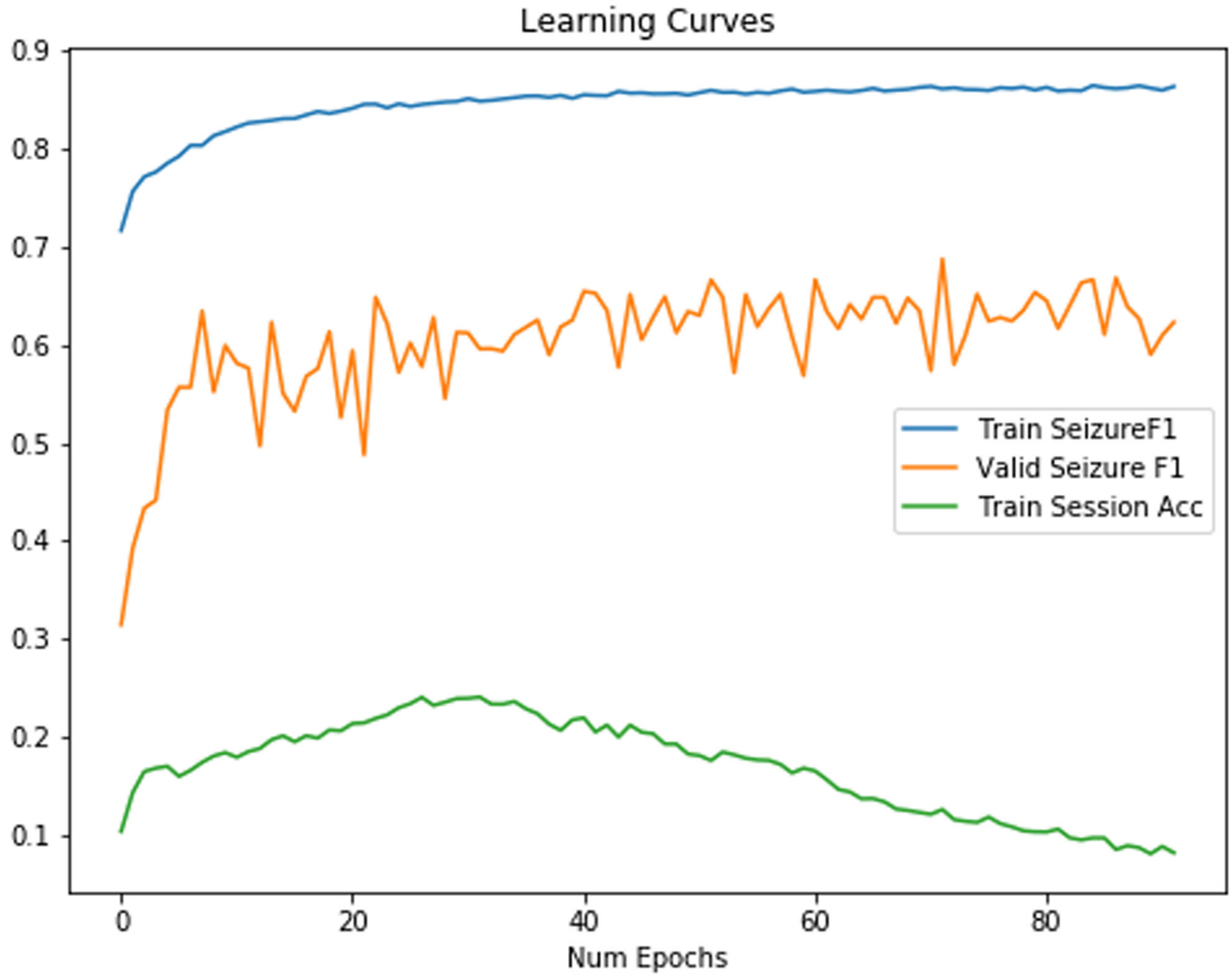

Figure 9:

Training/Validation F1 Score for a neural network with randomly rearranged channels on each minibatch and a random rescale factor of 1.5. These learning curves show that validation seizure F1 score increases rapidly before showing overfitting. Interestingly, the session identification accuracy increases before decreasing.

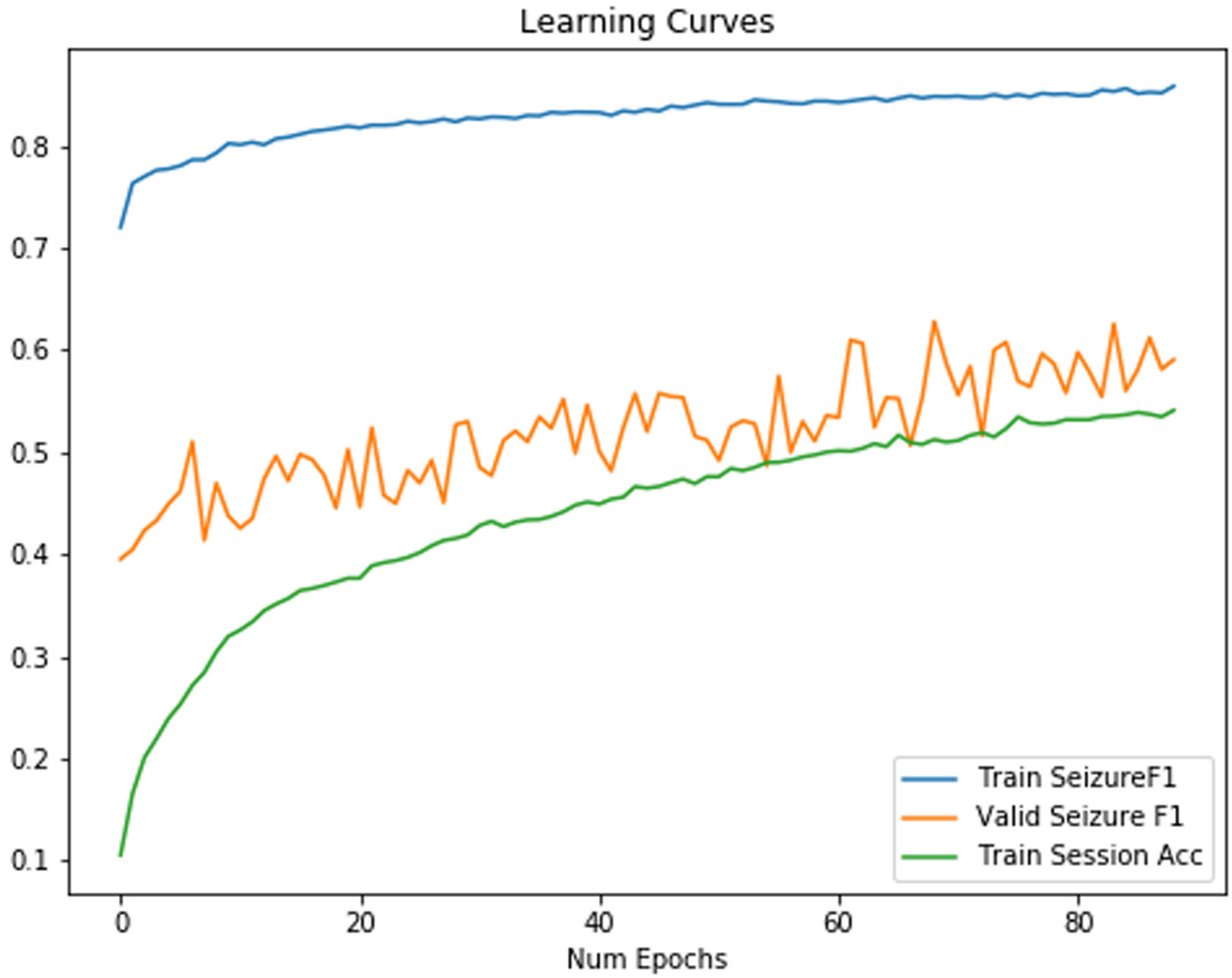

Figure 10:

Training/Validation F1 Score for a neural network with randomly rearranged channels on each minibatch and random rescaling on each minibatch of 2. There is a monotonically increasing training session identification again.

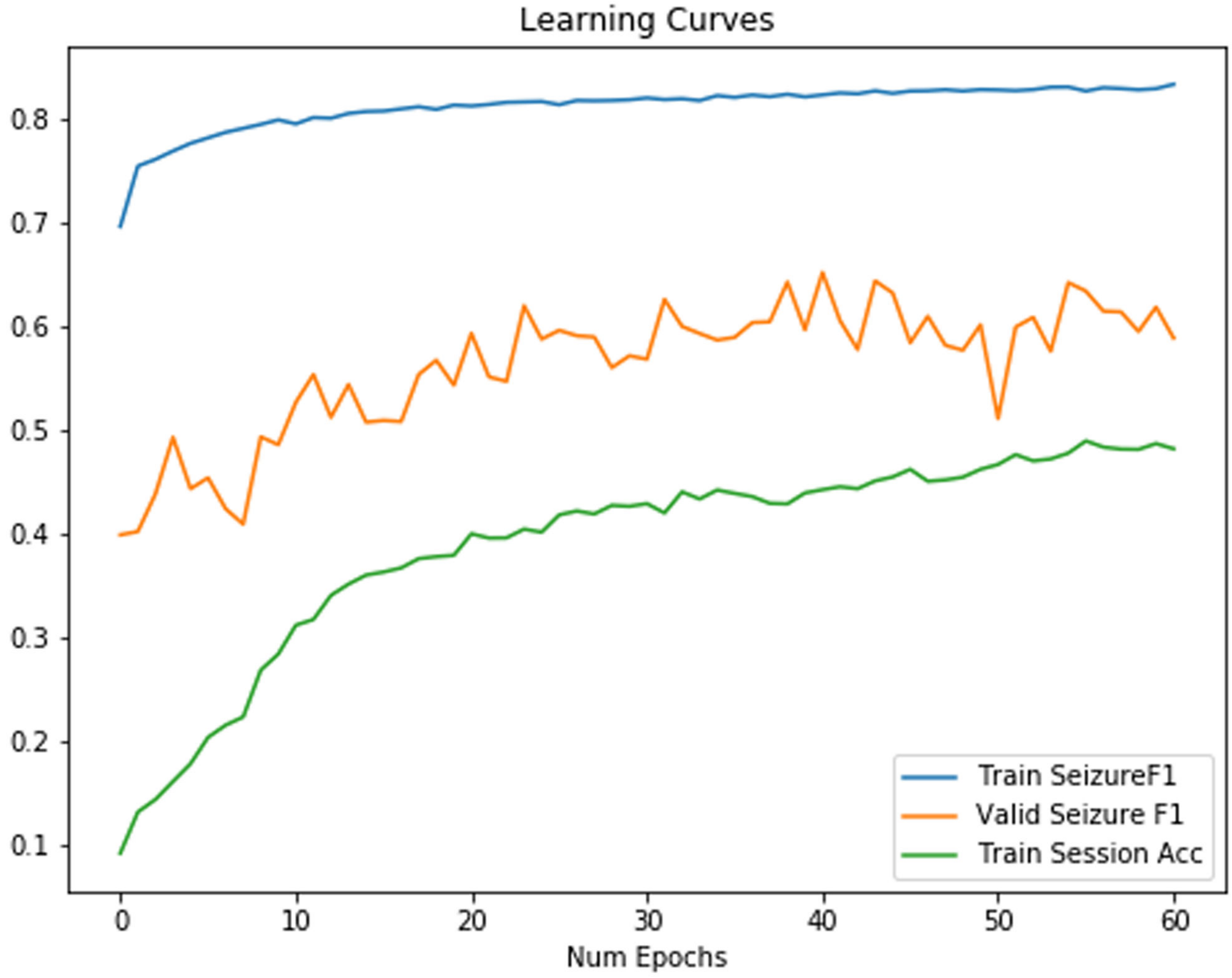

Figure 11:

Training/Validation F1 Score for a neural network with randomly rearranged channels on each minibatch and random rescaling on each minibatch of 2 with a seizure weight of 25 and patient weight of −1.

Table I.

Results for 1-second segments using a variety of Traditional Learning techniques

	AUC	Sensitivity	Specificity	F1 Score
Random Forest	0.6954	0.2955	0.9551	0.4331
Logistic Regression	0.4070	0.2848	0.5286	0.2833
XGBoost	0.7299	0.5769	0.8435	0.6347

Table II.

Results for 4-second segments using a variety of Traditional Learning techniques

	AUG	Sensitivity	Specificity	F1 Score
Random Forest	0.6534	0.7154	0.6532	0.5588
Logistic Regression	0.4028	0.7428	0.1273	0.3840
XGBoost	0.7518	0.5019	0.9064	0.5803

Table III.

Results for 4-second segments using a CNN/LSTM

AUC	Sens.	Spec.	F1 Score	Session Identification
0.716	0.544	0.874	0.556	0.620

TABLE IV.

Results for 4-second segments using a CNN/LSTM w/ Randomly Rearranged Channels

AUC	Sens.	Spec.	F1 Score	Session Ident.
0.785	0.782	0.786	0.629	0.224

TABLE V.

Results for 4-Second Segments Using Random Rearrange and Random Rescale

Rescale Factor	AUC	Sens.	Spec.	F1 Score	Session Ident.
1.05	0.732	0.709	0.830	0.625	0.261
1.1	0.718	0.765	0.785	0.619	0.179
1.5	0.758	0.866	0.695	0.652	0.082
2	0.726	0.727	0.814	0.622	0.541

Table VI.

Results for 4-second segments using an Adversarial Multitask CNN/LSTM

Seizure Weight	Session Weight	AUC	Sens.	Spec.	F1 Score	Session Ident.
50	−1	0.701	0.547	0.858	0.543	0.637
25	−1	0.687	0.440	0.884	0.483	0.251
10	−1	0.713	0.604	0.849	0.575	0.097
5	−1	0.717	0.578	0.864	0.571	0.059

Table VII.

Results for 4-second segments using an Adversarial Multitask CNN/LSTM with Randomly Rearranged Channels Each Minibatch

Seizure Weight	Session Weight	AUC	Sens.	Spec.	F1 Score	Session Ident.
50	−1	0.726	0.727	0.815	0.623	0.483
25	−1	0.757	0.674	0.872	0.644	0.499
10	−1	0.716	0.717	0.805	0.608	0.461
5	−1	0.716	0.730	0.799	0.611	0.368

Table VIII.

Results for 4-second segments using an Adversarial Multitask CNN/LSTM with Randomly Rearranged, Randomly Rescaled Channels Each Minibatch

Seizure Weight	Session Weight	AUC	Sens.	Spec.	F1 Score	Session Ident.
50	−1	0.711	0.712	0.800	0.601	0.545
25	−1	0.746	0.719	0.844	0.644	0.482
10	−1	0.755	0.596	0.894	0.613	0.437
5	−1	0.718	0.662	0.832	0.598	0.358