APS360 FINAL PROJECT: SEIZURE DETECTION WITH A CNN/LSTM MODEL

Andrew Sun
Student# 1010851008
andrewr.sun@mail.utoronto.ca

Yawen Zhang,
Student# 1010878668
yawenyw.zhang@mail.utoronto.ca

Dario Serrano
Student# 1009984304
dario.serrano@mail.utoronto.ca

Vanessa Huo
Student# 1010890840
vanessa.huo@mail.utoronto.ca

ABSTRACT

This report summarizes our group's research and development of an AI model capable of detecting seizures in Electroencephalogram (EEG) signals. We discuss relevant literature and similar models, before describing our data processing and model architecture. We end by showing that our model achieves strong results on never-before-seen data, and bring up possible ethical concerns. —-Total Pages: 8

1 INTRODUCTION

Epilepsy affects millions globally, yet current treatment options such as anti-epileptic drugs (AEDs) often fail to control seizures in about one-third of patients, making reliable, real-time seizure prediction a crucial clinical challenge to overcome Clinic. Therefore, our project goal is to investigate the application of Recurrent Neural Networks (RNN) for the analysis of (EEG) data to make accurate seizure or non-seizure predictions. RNNs are particularly well-suited to this task due to their ability for modeling temporal dependencies and retaining information over time.

2 BACKGROUND AND RELATED WORK

Seizure detection is a well-explored field of machine learning, with several papers and models on the subject. This focus is due to the challenging yet impactful nature of seizure detection. Currently, as there is no blanket cure for seizures, treatment focuses on detection and mitigation in order to minimize harm (Kerr et al., 2024). An accurate model for predicting or detecting seizures allows for improved statistics regarding treatment efficacy, removing patients from dangerous situations, and administering short-term medication (Batista et al., 2024) (Kerr et al., 2024).

2.1 RELATED PAPERS

Several papers test the effectiveness of certain ML models at this task, including one on Simple Vector Machines (SVMs), one on Recurrent Neural Networks (RNNs) and feature extraction, and one on Graph Convolutional Neural Networks (GCNs) and Long Short Term Memory cells (LSTMs) (Batista et al., 2024), (Zhu et al., 2024), (Kuang et al., 2024).

The SVM model aims to predict seizures 5 minutes in advance, breaking down the pre-seizure period into 3 overlapping events. This model increases sensitivity and performs above random chance for 62% of patients, as opposed to 49% without the breakdown. However, due to only predicting 49% of seizure occurrences, this model is lacking in a real-world context (Batista et al., 2024).

The RNN model uses wave transforms such as the Short-Term Fourier Transform and the Wavelet Transform to extract features from before and during seizures. The time and frequency features are

then fed into an encoder before the RNN classifies them, expressing an overall positive if 24 out of 30 consecutive outputs are positive. This model aimed to identify seizures at least 5 minutes in advance, but no more than 30 minutes in advance, and was able to do so with an F1[1] score of 98% (Zhu et al., 2024).

Lastly, the GCN and LSTM model produces the highest accuracy, reaching an average of around 99%. It was trained on preprocessed data from the Boston Children's Hospital, and uses the GCN to extract spatial features, and the LSTM to extract temporal features. This paper notes that CNNs and RNNs lose information when transforming data into lower dimensions, something that this approach avoids (Kuang et al., 2024).

Additionally, enough work has been done in this field for studies to be done reviewing the models and datasets available. Starting with "EEG datasets for seizure detection and prediction- A review", this paper notes several key features of the datasets available for seizure prediction. The number and type of EEG channels vary from dataset to dataset, which directly impacts the number of input features. This limits the generalization of models trained on one dataset, as their input is incompatible with other methods of EEG recording. Additionally, seizure datasets are highly imbalanced, as seizures occur less than 1% of the time, which can reduce the performance of neural networks. While balancing methods can be used to mitigate this, doing so results in data that is less representative of real-world scenarios. Overall, the variance across datasets limits the development of general models that can function for multiple datasets without significant retuning (Wong et al., 2023).

"The present and future of seizure detection, prediction, and forecasting with machine learning, including the future impact on clinical trials" is a paper that presents several unique insights on ML-based seizure detection. It presents the idea of seizure forecasts, models that give the probability of a seizure occurring soon, similar to a weather forecast. This approach works well with the continuous output of all models, eliminating the post-processing necessary to receive a binary answer. Additionally, this paper notes that the low frequency of seizures ($<$1%) necessitates metrics other than accuracy, lest a naive model that never predicts a seizure achieve an accuracy of 99%. These metrics include recall, the percentage of positive cases identified out of all positive cases, and precision, the ratio of true positives to all positives classified by the model (Kerr et al., 2024).

These papers provide a solid background for our project, with several insights and examples to account for during our work.

3 Ethical Concerns

Two key ethical concerns were identified and kept in mind throughout the model training:

1. **Patient Data Privacy Ethics:** There is risk in handling sensitive EEG without proper anonymization. We made sure to use CHB-MIT data that had patient metadata replaced with filler data to protect sensitive information.
2. **Limited Generalizability to Diverse Seizure Types:** Our model is built on a pediatric cohort from CHB-MIT and may not depict EEG patterns of all ages or seizure types, risking misclassification when used on unfamiliar patient profiles and data. This is a clear limitation that must be kept in mind when applying our model in any practical capacity. (Louis et al., 2016)

4 Data Processing

We used the "EEG Seizure Analysis Dataset", a subset of the Children's Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) EEG dataset. This dataset is publicly available on Kaggle and contains scalp EEG recordings from 22 pediatric subjects with seizures at the CHB-MIT Badea & Collaborator (2021). The data is provided in a .npz file, each containing two arrays:

- Signals of shape (N, 23, 256), where N is the number of signals, each signal is sampled at 256 Hz across 23 electrode recording channels.

[1]The F1 score is a metric for the efficacy of a model, defined as the harmonic mean of precision and accuracy (Kundu, 2022)

- labels of shape (N,), where each signal is labeled as 1 (seizure) or 0 (non-seizure).

We followed the below procedures to process the data:

1. **Data Loading:** Load the complete dataset files from Kaggle (eeg-predictive_train.npz and eeg-predictive_val.npz) with NumPy. Assign variables "x_all" and "y_all" respectively to the signal and labels. x_all should be an array of signals with shape (N, 23, 256) and y_all should be an array of labels with shape (N,).

2. **Data Splitting:** From sklearn.model_selection import train_test_split, then using the module, split "x_all" and "y_all" into 80% train, 10% validation and 10% test sets. Assign variable names of "x_train", "y_train", "x_val", "y_val" and "X_test", "y_test" to their respective datasets.

3. **Noise Removal:** Construct a Butterworth band-pass filter function using SciPy modules ("butter" and "filtfilt") to remove frequencies outside of the 0.5-40 Hz range, which is the typical range of the EEG frequency spectrum. The band-pass filter function is then applied to all the signal arrays such as "x_train", "x_val" and "x_test" to remove unwanted noise.

4. **Artifact Removal:** Artifacts are removed using PyWavelets to decompose each signal. Common parameters for EEG signals are employed including: wavelet="db4", mode="soft" and level=None. The coefficients of each wavelet are calculated using the multilevel decomposition platform – pywt.wavedec(). Sigma is calculated using the Median Absolute Deviation (MAD), where σ = np.median(np.abs(last_coefficient)) $\times \frac{1}{0.6745}$ The universal threshold is calculated to assign derived values for noise cutoff and it equals $\sigma \times \sqrt{2 \times \ln(\text{length of data})}$. Then, all the signals are denoised and reconstructed using the multilevel reconstruction – pywt.waverec() with the above variables, and the denoised data are saved separately for further use. Dummy signals, which are flat lines that do not contain useful information, are kept in their original shape and will not be considered in model training. This is accomplished by checking the amplitude of the data – any signal with total amplitude smaller than 10^{-4} is considered a dummy signal.

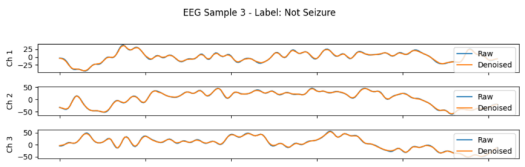
5. **Sliding Windows:** Construct a sliding window function that defines the boundary of the real-time, continuous EEG signal. As we would like to predict seizures in the next 30 seconds, our window size is half that (15 seconds). A window is labeled as a seizure if ¿30% of the following 30 seconds is labeled as seizure. The sliding window function is then applied to all the signal arrays ("x_train", "x_val" and "x_test") , which give them a shape of (N, 15, 23, 256).

6. **Data Normalization:** Reshape "x_train", "x_val" and "x_test" into 2D matrices for scaling. Standardization is then performed using StandardScaler from scikit-learn, which transforms the features to have a mean of 0 and a standard deviation of 1. The function .fit_transform() is first applied to the training set, then .transform() is applied to the validation and test set. This ensures consistency between datasets and avoids data leakage. The modified datasets are then reshaped back to their original shape.

Once processed, the data are saved in a new .npz file with numpy.savez_compressed, for later use. The processed data can be seen below in Figure 1.

```
X_train shape: (7724, 15, 23, 256)
X_val shape: (966, 15, 23, 256)
y_train shape: (7724,)
y_val shape: (966,)
Train seizure %: 17.14%
Val seizure %: 17.18%
```

(a) Shape and distribution of training and validation sets



(b) Graphical representation of raw and denoised signals, denoised (red) overlays raw (blue)

Figure 1: Numerical and graphical representations of the processed data.

5 BASELINE MODEL

As a baseline model, we elected to use an SVM-based model found on Github (dragonpilee). This model was trained and tested on Bonn University's EEG dataset, with 11500 samples split into 5 classes, one of which is classified as being a seizure. This data is transformed, normalized and balanced before being trained on, achieving an F1-score in the mid-90s.

On our dataset, this model performed significantly worse, despite using the same statistical tools for feature extraction. All accuracy metrics scored 60%, much lower than the original source. This is likely due to the different measurements present in our EEG dataset. Nonetheless, we still believe our chosen dataset to be suitable for this project due to its size and documentation.

All features are included

	precision	recall	f1-score	support
0.0	0.60	0.60	0.60	328
1.0	0.60	0.60	0.60	328
accuracy			0.60	656
macro avg	0.60	0.60	0.60	656
weighted avg	0.60	0.60	0.60	656

Figure 2: Baseline model's performance on dataset

6 MODEL ARCHITECTURE

The final model is a 1D Convolutional Neural Network (CNN) followed by a unidirectional Long Short-Term Memory (LSTM) network. The model processes EEG input of shape (n, 15, 23, 256), where n is the batch size, 15 is the number of time steps per window, 23 is the number of EEG channels, and 256 is the sampling frequency (time points per channel). The input is reshaped into (n × 15, 23, 256) so that the CNN can process all time steps in parallel.

The CNN stage consists of a single Conv1D layer with 23 input channels, 32 output filters, a kernel size of 3, and "same" padding (padding = 1). This is followed by a ReLU activation for non-linearity, batch normalization to stabilize learning, and a max pooling layer of size 2 to downsample temporal features. The CNN outputs are flattened and passed to a unidirectional LSTM with a hidden size of 64 to capture temporal features across the 15 time steps.

The final hidden state of the LSTM is passed through a dropout layer (to reduce overfitting) and a fully connected layer with one output neuron for binary classification (seizure vs. non-seizure). The model is trained with BCEWithLogitsLoss, using the pos_weight parameter to address class imbalance. Optimization is performed with the AdamW optimizer (weight decay = 1×10^{-2}) to improve the generalization ability of the model.

The configuration of this architecture was chosen to combine the CNN's ability to extract spatial features from multi-channel EEG signals with the LSTM's strength in temporal dynamics, resulting in a robust performance on our classification. A visual representation of our model architecture and data pipeline can be seen below.

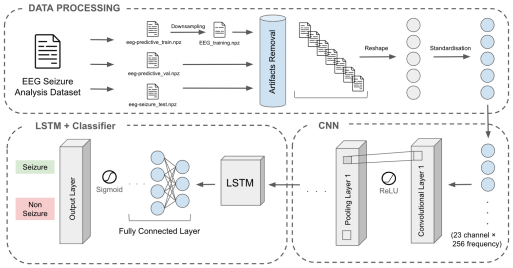


Figure 3: Neural Network Architecture

7 RESULTS

7.1 QUANTITATIVE RESULTS

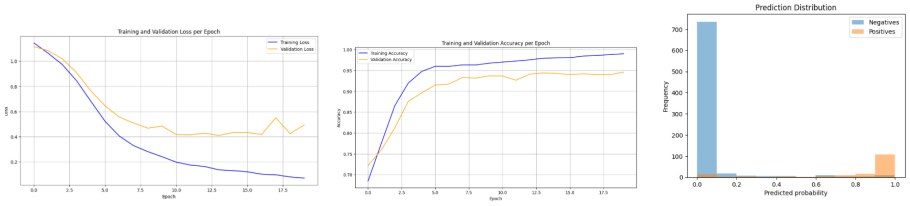
We used the following quantitative metrics to evaluate our models' performance on validation data:
Accuracy measures the proportion of all predictions that are correct, which can be misleading in imbalanced datasets like ours, where the non-seizure class dominates.
The **AUROC** (Area Under the Receiver Operating Characteristic curve) score evaluates the model's ability to discriminate between classes across the possible decision thresholds; values closer to 1.0 indicate stronger separation between seizure and non-seizure EEG data segments.
The **F1-score** is the harmonic mean of precision (positive predictive value) and recall (sensitivity), making it a balanced measure when minimizing both false positives and false negatives in seizure detection.
Recall demonstrates the proportion of true seizure events correctly detected (low recall means more missed seizures, which is undesirable in clinical situations).
The **False Positive Rate (FPR)** is the proportion of non-seizure segments incorrectly classified as seizures; a low FPR means less false alarms clinically.
From these recorded metrics, the CNN1DLSTM model achieved the best overall balance, with high AUROC and F1 scores, a strong recall and a relatively low FPR, indicating that this architecture captures most seizure events while keeping a reasonable specificity.
The quantitative metrics of each architecture our team has tested are summarized in the following table.

Model	Accuracy	AUROC	F1	Recall	FPR
1D CNN+LSTM	0.944 ± 0.001	0.961 ± 0.002	0.835 ± 0.002	0.855 ± 0.005	0.135 ± 0.005
1D CNN+BLSTM	0.931 ± 0.004	0.947 ± 0.006	0.797 ± 0.009	0.80 ± 0.02	0.21 ± 0.02
2D CNN+LSTM	0.942 ± 0.003	0.964 ± 0.003	0.827 ± 0.009	0.803 ± 0.02	0.197 ± 0.02
2D CNN+BLSTM	0.933 ± 0.001	0.962 ± 0.003	0.844 ± 0.004	0.894 ± 0.008	0.146 ± 0.008

Table 1: Validation results of models averaged over three training sessions

7.2 QUALITATIVE RESULTS

To further illustrate how the CNN1DLSTM architecture performs, we present three representative plots generated from the training process.



(a) Training and Validation Loss over Epochs

(b) Training and Validation Accuracy over Epochs

(c) Prediction Probability Distribution

Figure 4: Training and prediction graphs of our final model.

From these graphs, we can see the following insights:

1. Training Loss vs. Epoch – Shows a consistent decrease in loss over training, indicating stable optimization without severe overfitting.
2. Validation Accuracy vs. Epoch – Demonstrates a steady rise before plateauing, suggesting substantial generalization on unseen data.

3. Prediction Probability Distribution – Shows that the model gives high confidence to correctly classified seizure segments (probabilities near 1.0) and non-seizure segments (probabilities near 0.0), though some overlap remains for borderline cases.

These outputs collectively demonstrate both the model's learning dynamics and its decision confidence. The loss vs. accuracy curves show that CNN1DLSTM fits effectively upon training, while the probability histogram provides insight into the model's high confidence when classifying EEG segments.
Overall, this model performed best on signals with clear, high-amplitude rhythmic spike patterns, matching expert seizure annotations. It struggled on noisy, low-amplitude signals, and sometimes misclassified artifact-heavy segments as seizures, reflecting the trends seen in the probability overlap on the histogram. These features contextualize the quantitative results in Table 1, showing that while CNN1DLSTM is the most robust model tested, further refinement and optimization is needed to improve performance on subtle or noisy events.

7.3 Test Results on New Data

To assess the model's ability to generalize unseen data, we evaluated the final CNN1DLSTM architecture on a separate set of EEG samples that had not been used during training, validation, or hyperparameter tuning. This new dataset included recordings from patients not present in the training set, with a spectrum of seizure patterns and noise levels reflecting real-world variability. The model's performance can be seen below, and closely matched its validation metrics shown previously, achieving an overall accuracy of 94.9%, macro-averaged F1 score of 0.909, and strong recall for both seizure and non-seizure classes (0.830 and 0.974, respectively). The recall values indicate that the model is able to fairly reliably detect seizure events, but a 17% false negative rate is not ideal for clinical applications. Precision remained strong for the non-seizure class (0.965), though slightly lower for seizure events (0.867), suggesting that false positives may still be present due to artifact-heavy signals. Overall, the results demonstrate that the model meets expectations on entirely unseen data and is robust across varying samples and recording conditions, supporting its potential for application beyond the training distribution.

Final Test Classification Report:

	precision	recall	f1-score	support
0.0	0.9653	0.9738	0.9695	801
1.0	0.8671	0.8303	0.8483	165
accuracy			0.9493	966
macro avg	0.9162	0.9020	0.9089	966
weighted avg	0.9486	0.9493	0.9488	966

Figure 5: Model Performance on Test Set

8 Discussion

Both the quantitative and qualitative results demonstrate our final model's ability to accurately distinguish between seizure vs. non-seizure activities. We think the model is performing well as it retained high accuracy ($> 94\%$) on both balanced and imbalanced datasets, suggesting that it is not simply memorizing patterns or relying on class probability distributions, but instead learning patterns and recognizing features. This provides encouraging evidence for the feasibility of using machine learning aided approaches to detect seizures from EEG signals, despite the natural variations in EEG signals across gender, age, and individual pediatric subjects.
We are surprised that unidirectional LSTMs outperformed bidirectional LSTMs. Initially we expected BLSTM's dual-sequencing capability to better capture temporal dependencies from both past and future signals, thereby improving prediction accuracy. However, our results contradicted that hypothesis, as for both 1D and 2D CNN inputs, BLSTMs performed worse than unidirectional LSTMs. While unexpected, this outcome actually makes clinical seizure prediction surprisingly more practical. In current reality, the most urgent need is continuous, real-time seizure prediction, where models only have access to past and current signals. Therefore, although the experimental

results opposed our initial assumption, it presents a more logical and implementable solution for a clinical diagnosis setting.

In conclusion, we learned that high-quality and carefully cleaned data, combined with thoughtful choices of machine learning models and extensive hyperparameter tuning, can yield strong results. More broadly, our findings reinforce the potential for machine learning to benefit healthcare, particularly in diagnostic applications. To maximize this impact, current AI researchers must recognize the importance of interdisciplinary knowledge and focus on using machine learning techniques to empower societal development.

REFERENCES

Adrian-Catalin Badea and Collaborator. EEG Seizure Analysis Dataset. https://www.kaggle.com/datasets/adibadea/chbmitseizuredataset, 2021. Accessed: 2025-06-12; Subdataset extracted from CHB-MIT EEGs dataset :contentReference[oaicite:1]index=1.

Joana Batista, Mauro F Pinto, Mariana Tavares, Fábio Lopes, Ana Oliveira, and César Teixeira. EEG epilepsy seizure prediction: the post-processing stage as a chronology. *Scientific Reports*, 14(1): 407, January 2024.

Mayo Clinic. Seizures - Symptoms and causes — mayoclinic.org. https://www.mayoclinic.org/diseases-conditions/seizure/symptoms-causes/syc-20365711. [Accessed 12-06-2025].

dragonpilee. Epileptic seizure detection system. GitHub. https://github.com/dragonpilee/Epileptic-Seizure-Detection-System, Accessed: 2025-06-12.

W. T. Kerr, K. N. McFarlane, and G. F. Pucci. The present and future of seizure detection, prediction, and forecasting with machine learning, including the future impact on clinical trials. *Frontiers in Neurology*, 15:1425490, 2024. doi: 10.3389/fneur.2024.1425490.

Zhejun Kuang, Simin Liu, Jian Zhao, Liu Wang, and Yunkai Li. Epilepsy EEG seizure prediction based on the combination of graph convolutional neural network combined with long- and short-term memory cell network. *Appl. Sci. (Basel)*, 14(24):11569, December 2024.

Rohit Kundu. F1 score in machine learning: Intro calculation. https://www.v7labs.com/blog/f1-score-guide, 2022. [Accessed: 2025-06-12].

E. K. S. Louis et al. *EEG in the Epilepsies*. American Epilepsy Society, 2016. Accessed: 2025-06-12.

Sheng Wong, Anj Simmons, Jessica Rivera-Villicana, Scott Barnett, Shobi Sivathamboo, Piero Perucca, Zongyuan Ge, Patrick Kwan, Levin Kuhlmann, Rajesh Vasa, Kon Mouzakis, and Terence J O'Brien. EEG datasets for seizure detection and prediction- a review. *Epilepsia Open*, 8(2): 252–267, June 2023.

R. Zhu, W.-X. Pan, J.-X. Liu, and J.-L. Shang. Epileptic seizure prediction via multidimensional transformer and recurrent neural network fusion. *Journal of Translational Medicine*, 22(1):895, 2024.