2023

A Unique Method of Using Information Entropy to Evaluate the Reliability of Deep Neural Network Predictions on Intracranial Electroencephalogram

Elakkat Dharmaraj Gireesh
University of Central Florida

A UNIQUE METHOD OF USING INFORMATION ENTROPY TO EVALUATE THE RELIABILITY OF DEEP NEURAL NETWORK PREDICTIONS ON INTRACRANIAL ELECTROENCEPHALOGRAM

by

ELAKKAT DHARMARAJ GIREESH
MMST, Indian Institute of Technology, Kharagpur, 2004,
MBBS, Calicut Medical College, 2000

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical & Computer Engineering
in the College of Engineering & Computer Science
at the University of Central Florida,
Orlando, FL

Summer Term
2023

Major Professor: Varadraj P. Gurupur

ABSTRACT

Deep Neural networks (DNN) are fundamentally information processing machines, which synthesize the complex patterns in input to arrive at solutions, with applications in various fields. One major question when working with the DNN is, which features in the input lead to a specific decision by DNN. One of the common methods of addressing this question involve generation of heatmaps. Another pertinent question is how effectively DNN has captured the entire information presented in the input, which can potentially be addressed with complexity measures of the inputs.

In the case of patients with intractable epilepsy, appropriate clinical decision making depends on the interpretation of the brain signals, as recorded in the form of Electroencephalogram (EEG), which in most of the cases will be recorded through intracranial monitoring (iEEG)). In current clinical settings, the iEEG is visually inspected by the clinicians to arrive at decisions regarding the location of the epileptogenic zones which is used in the determination of surgical planning. Visual inspection and decision making is a very tedious and potentially error prone approach, given the massive amount of data that need to be evaluated in a limited amount of time. We developed a DNN model to evaluate iEEG to classify signals arising from epileptic and non-epileptic zones.

One of the challenges of incorporating the deep neural network tools in the medical decision making is the black box nature of these tools. To further analyze the underlying reasons for DNN’s decision regarding iEEG, we used heatmapping and signal processing tools to better understand the decision-making process of DNN. We were able to demonstrate that the energy rich regions, as captured by analytical signals, is identified by DNN as potentially epileptogenic, when arriving at decisions.

We explored the DNN's ability to capture the details of the signal with information theoretical approaches. We introduced a measure of confidence of DNN predictions, named certainty index, which is calculated based on the overall outputs in the penultimate layer of the network. We employed the method of Sample Entropy (SampEn) and were able to demonstrate that the DNN's prediction certainty is related to how effectively the heatmap is correlated to the SampEn of the entire signal. We explored the parameter space of the SampEn calculation and demonstrate that the relationship between SampEn and certainty of DNN predictions hold even on changing the estimation parameters.

Further we were able to demonstrate that the rate of change of relationship between the DNN output and activation map, as a function of the sequential DNN layers, is related to the SampEn of the signal. This observation suggests that the speed at which DNN captures the results is directly proportional to the information content in the signal.

To all my great teachers

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee members: Prof. Varadraj Gurupur, Prof. Kalpathy Sundaram, Prof. Vikram Kapoor, Prof. Jiann-Shiun Yuan and Prof. Christian King for all the guidance for my dissertation. Also, my sincere thanks to my committee chair, Prof. Varadraj Gurupur who helped me with every step of the PhD process. I would also like to thank my previous advisor Prof. Sharma Thankachan who initially introduced me to the idea of research in this field. I also want to acknowledge the support from the faculty and staff of Department of Electrical & Computer Engineering, for all the guidance in every step of this process.

I would like to thank the AdventHealth Epilepsy Center Orlando, for providing the data used in this study. Also, I would I like to thank all the patients, from whom the data used in this study was acquired from. The plight of the patients was also a strong motivation to move forward in this study from an application standpoint. In addition, I would like to extend my gratitude to AdventHealth Medical Group for financially supporting my educational efforts.

Finally, I would like to thank my family for all the support, understanding and encouragement. Without their support this work may not have been possible. Also, my exposure to philosophy helped me in exploring the field of information, as I have done in this study. I would like to thank all those who helped me in stepping through this sphere of knowledge.

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CNN-Convolutional Neural Network

DNN-Deep Neural Network

EEG- Electroencephalogram

Grad-CAM- Gradient-weighted Class Activation Mapping

iEEG- Intracranial Electroencephalogram

LSTM - Long Short-Term Memory

SampEn- Sample Entropy

CHAPTER 1: INTRODUCTION

The Deep neural networks (DNNs) are fundamentally information processing machines, which uses features from input to come up with predictions for various kinds of tasks including signal classification, image recognition, medical diagnosis, and natural language processing. While DNNs allow for learning from huge amounts of data, the question of whether DNN used the relevant information from the data has always been pertinent. Heatmapping approaches have been described for exploring the learning process of DNN for enhancing our understanding on why the model finalized on a specific decision. But the question of, whether the heatmaps captured all the information in the model and how it is related to the confidence of the predictions of DNN has not been explored. Using information theoretic approaches including Sample entropy (SampEn) measures, can potentially enable us, in scrutinizing the learning process of DNN, thereby having an independent measure of reliability. In this study we evaluate the learning process of DNN, using the fundamental measures of information content in a time series.

Information theory-based approaches have already been used in the field of DNN for improving the performance[1],[2] as in cross-entropy measures. Also, measures of information are commonly in training and validating DNN models. Taking a step forward, we have attempted to use the information theoretic approach to assess the heatmaps of DNN outcomes and compare the confidence of predictions. Further, we use the activation gradients to assess how the information propagates across DNN and correlate this measure with the information content in the signal.

With the advancements in the computational tools, our understanding of the medical conditions has vastly improved. At the same time several challenges exist, which need to be addressed, especially in the field of neuroscience, both from a clinical and basic science standpoint. Modern engineering techniques, including deep learning methods, enhance the capabilities of analyzing tremendous amount of data, emerging in the medical field. At the same time, the science behind these modern engineering techniques needs to be thoroughly explored to ensure that optimal results are obtained, when they are applied in the field of medical science. This would be especially important from a standpoint of adoption of these techniques to different critical applications of human service, especially in evaluation and management of medical conditions.

Epilepsy as a major medical challenge with data analysis

Epilepsy is one of the common neurological disorders and accounts for 30 million disability adjusted life years. Around 7.6 per 1000 persons have epilepsy during their lifetime [3]. Approximately 1% of the US population suffers from epilepsy and in 30-35 % of these patients, seizures cannot be controlled medically and are considered to have refractory epilepsy [4]. According to World Health Organization data, 70 million people suffer from epilepsy. Worldwide, epilepsy is the fourth major cause of brain disorders [5]. The age standardized prevalence of epilepsy is 621.5 per 100,000 population[6]. Estimated total cost of care of epilepsy patients, in United States, amounts to approximately $24 billion per year[7]. This is in addition to the huge social, emotional, and personal sufferings of these patients and families.

Epilepsy, in many patients can be a lifelong disease also which intensifies the impact of this disease. The epilepsy incidence in different age groups is given the Figure 1.

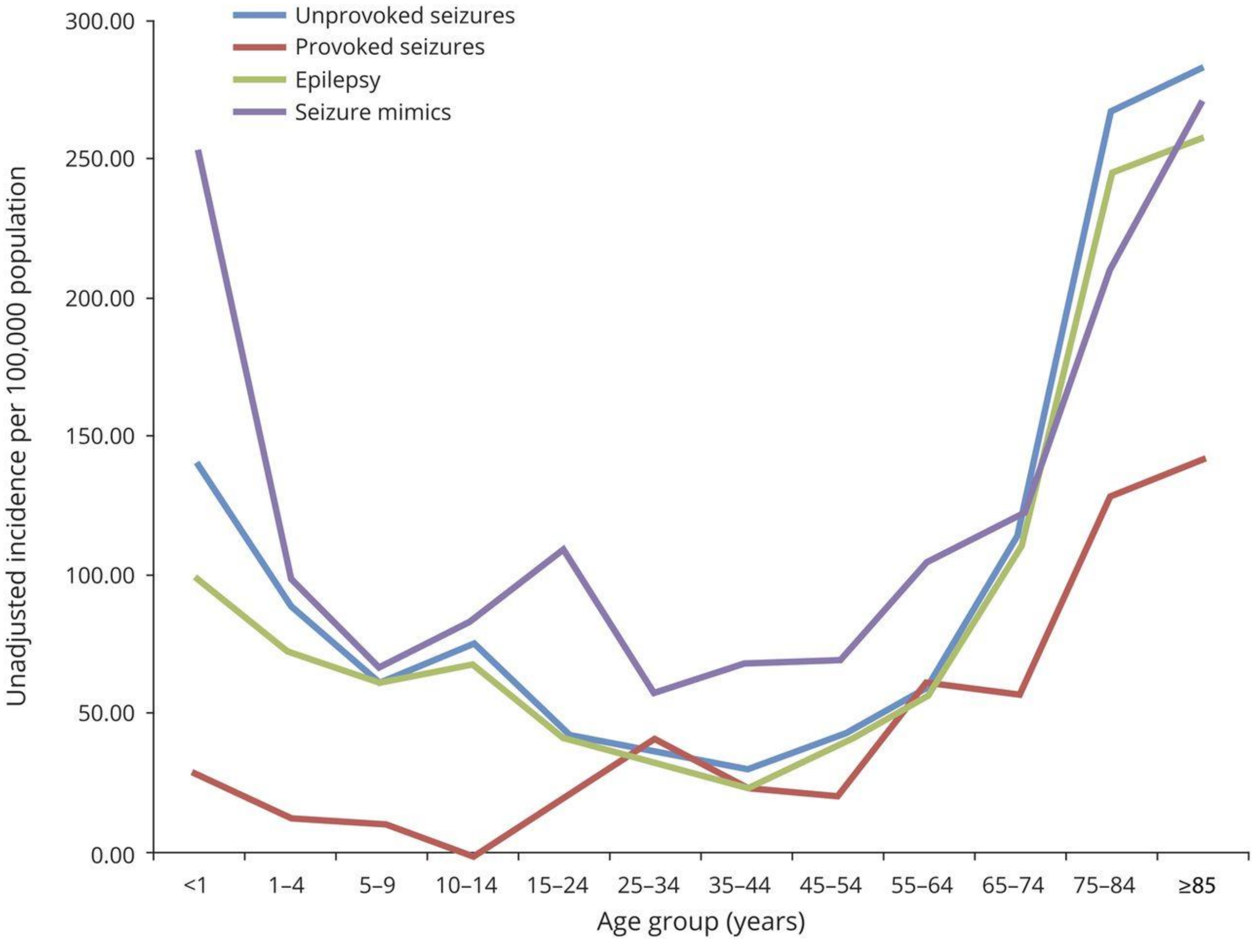


Figure 1. Graph showing the incidence of epilepsy in different age groups.
There are usually two peaks of incidence of epilepsy, 1. At early age group until the age of 4 years. 2. Late age starting between 65- 70 years of age.

Patients who develop refractory epilepsy may suffer from recurrent seizures, status epilepticus and neurocognitive impairments which all can lead to significant morbidity and mortality. If medications fail, they often need to be treated with surgical options which is planned based on intracranial EEG (leg) monitoring, which aims at accurate determination of the epileptogenic zones. Before planning for intracranial EEG monitoring these patients undergo extensive evaluation which includes following steps as illustrated in Figure 2:

1. Epilepsy monitoring unit evaluation where the typical seizures are captured to make initial impressions about the nature of seizures, clinical findings and broad localization.

2. Magnetic resonance imaging to identify the structural abnormalities of the brain

3. Magnetoencephalography (MEG) which helps in identifying the epileptiform discharges arising from the different regions of the brain due to the magnetic components in the signal.

4 Neuropsychology evaluation to determine the neurocognitive status of the patients along with evaluating the deficits arising from the recurrent seizures.

5. Positron Emission Tomography (PET) scan aimed at identifying the potential abnormal metabolic patterns in the regions of epileptic activity.

6. Single photon emission computerized tomography (SPECT) to evaluate transient changes in the amount of glucose uptake in the regions of epilepsy.

7.Wada testing which helps in establishing which side of the brain significantly contributes to the language and memory functions, which enables the clinical team in avoiding causing any additional neurocognitive deficits to the patient.

8. Functional magnetic resonance imaging which helps in identifying the brain regions controlling the language, motor and sensory functions.

After the initial evaluation, a clinical board consisting of epileptologists, neurosurgeon, neuropsychologist, social workers etc., evaluate the data and come up with a plan for surgical options. In some cases, a surgery based on the initial evaluation itself, can be made but in a considerable number of cases additional evaluation with intracranial EEG may be needed. For this purpose, patients undergo intracranial EEG monitoring after placement of the electrodes inside the brain. Two main types of electrode placements are employed in this regard:1.surface electrode placement which may be either as a single strip of electrode or a grid of electrodes 2. depth electrodes which are placed deeper into the brain tissues.

Typical intracranial electrode placement is shown in Figure 3 and signals from the brain is recorded for prolonged periods and analyzed for seizure onset zones. Identifying the features suggestive of seizure onset in these signals can sometimes be challenging, given the huge amount of data that need to be visually analyzed by clinicians Figure 4. It may be noted that even with surgery some patients may not get complete seizure freedom, suggesting failure of that modality of assessment in the localization of epileptogenic zones. Failure to accurately identify the epileptogenic zones in iEEG may be contributing to the inadequate seizure control reported even after epilepsy surgery, which ranges from 30-60% [8],[9]. This raises the critical need for tools which can accurately analyze the iEEG signals and identify the epileptogenic zones.

Motivation for our work

As mentioned in the previous section visual evaluation of iEEG data to arrive at clinical decisions is extremely challenging and can potentially lead to erroneous decisions. It is in this context that the approaches of deep learning would provide better tools for analyzing the iEEG signals to improve the detection of epileptogenic zones. DNNs have been used in multiple fields including image classification [10], signal processing[11,12] and natural language processing

[13]. In this study we evaluate the use of deep neural network for evaluating iEEG signals to accurately identify the seizure onset zones, using signals recorded during the interictal periods (when patients do not have seizures). We employed a short duration (1 minute) of signal to identify the epileptogenic zone in contrast with multiple days of iEEG monitoring used in the current medical decision-making process.

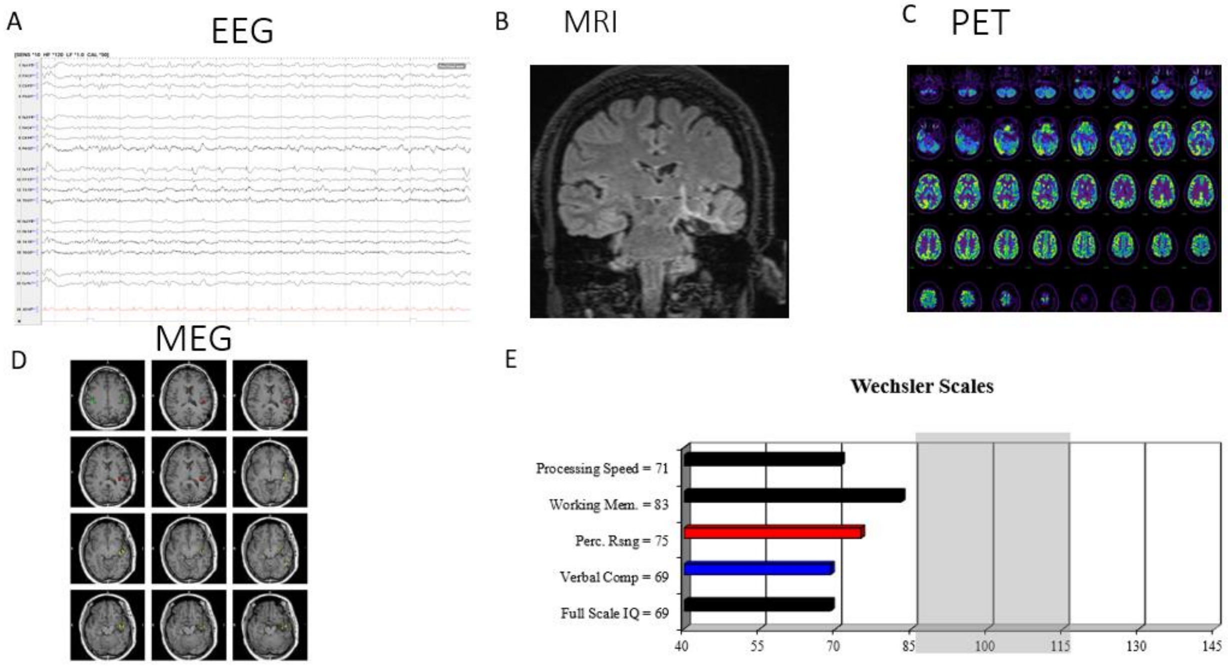


Figure 2. A broad overview of the work up for epilepsy surgery planning.
A. Recording of the EEG and capturing the typical events of the patient to correlate the clinical and electrographic patterns. B. MRI brain to identify any focal abnormalities in the different brain regions. This can be pathological changes like, encephalomalacia, focal cortical dysplasia, hemosiderin deposition etc. C. Positron emission tomography (PET) scan performed to identify any regions in the brain that is having abnormal metabolism. D. Magnetoencephalography (MEG) scan: This modality captures the magnetic signals corresponding to the EEG. This method acts as a complementary study for EEG and is especially useful in identifying deeper epileptogenic foci. This also helps in functional mapping of the different brain regions. E. Neuropsychological assessment of the patients help in identifying preexisting neurocognitive morbidities. It also helps in better identifying the seizure focus.

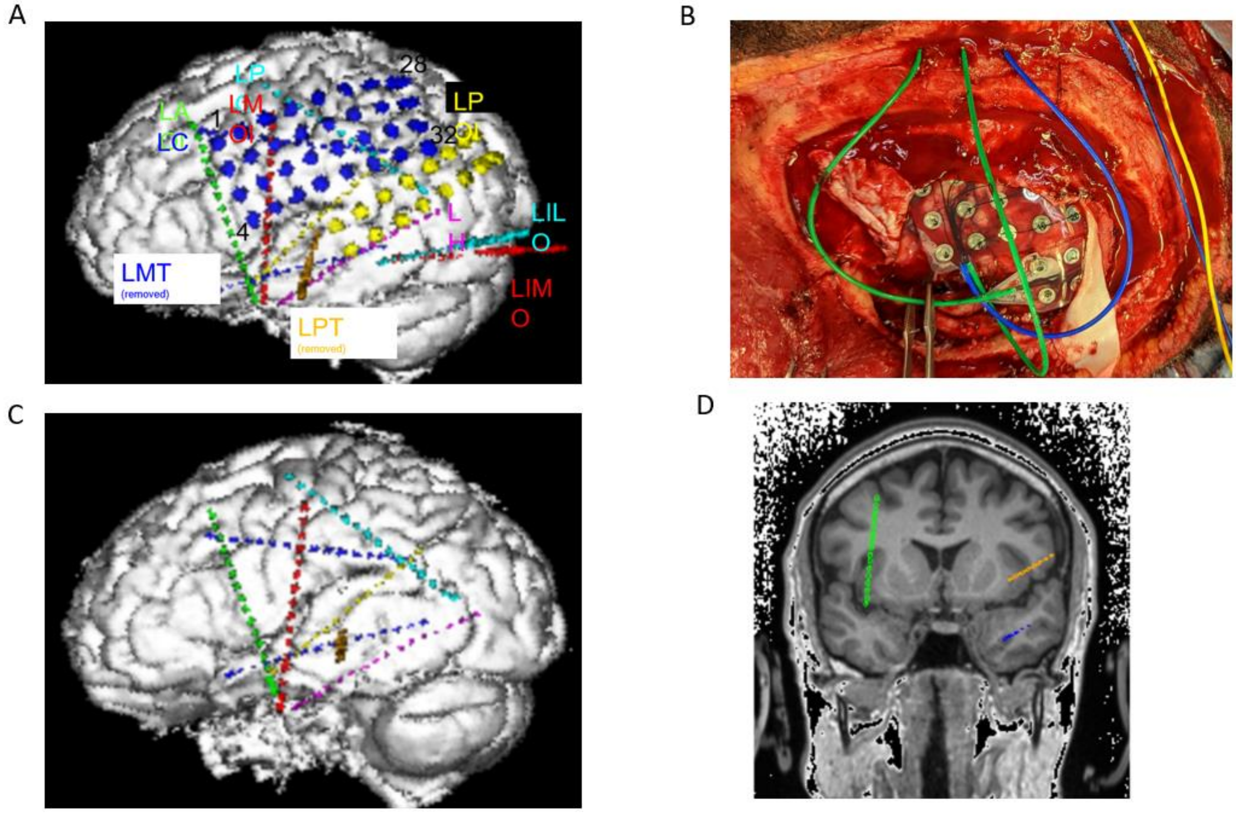


Figure 3. Different modalities of intracranial EEG monitoring
A. Shows surface electrode placement in the form of grids or strips B. Partial view of the surface electrode placement during surgery. Depth electrode placement (also known as stereo EEG) D. Coronal view of the deep electrode placement shows the deep structures of the brain being monitored.

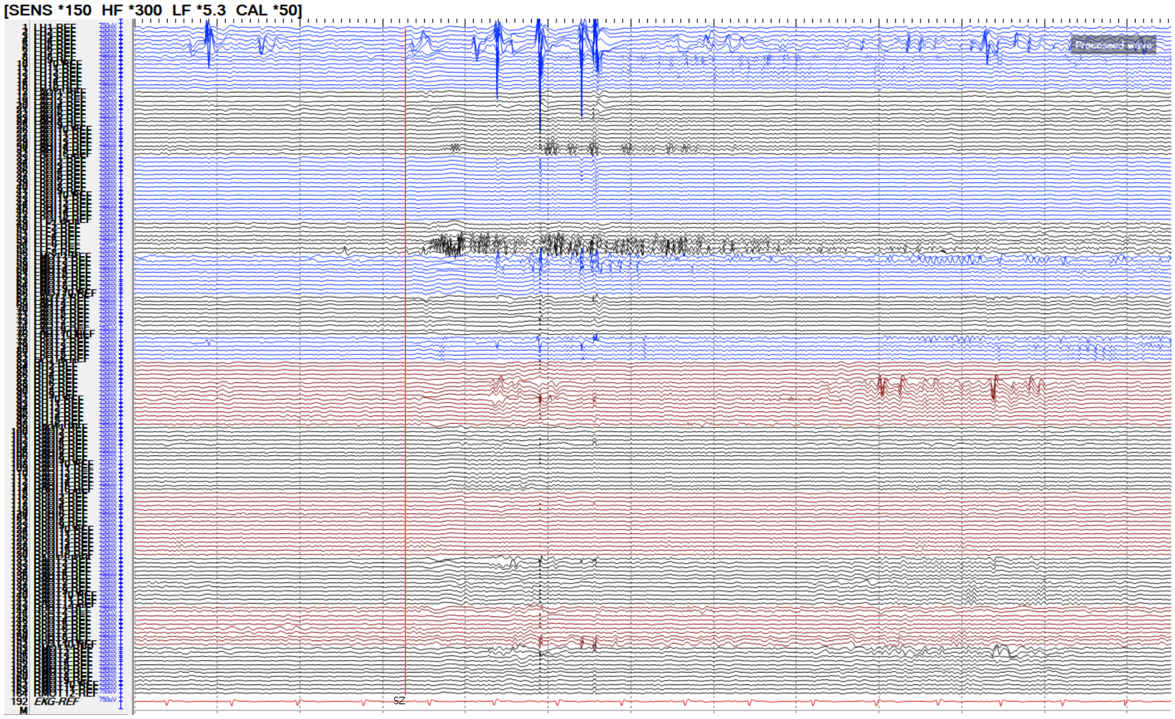


Figure 4. The intracranial EEG data acquired for around 12 seconds from the different intracranial electrodes simultaneously.
This kind of data is continuously recorded in patients for several days.

Use of deep learning methods for evaluation of intracranial EEG

With the advent of advanced computational techniques, along with enhanced computing power, different deep learning tools have been used for the evaluation of brain signals.

The patients undergoing intracranial EEG monitoring are placed with stereo-EEG or grid type of electrodes and admitted to neuro ICU. The EEG is monitored for several days for capturing interictal and ictal data. This monitoring may span several days, and the data is obtained from several electrodes usually between 150-250 in number. It is extremely difficult for the interpreting physician to go through the entire data to arrive at decisions. Also, even if they go through the entire recording of data, analyzing data from each electrode is not practical. This calls for the need for having automated methods that can help in predicting the epileptogenic zones from iEEG data.

Another concern of visual evaluation of the data is inter-individual variation between different physicians. Having quantitative measures and automated tools will address the individual bias associated with the visual inspection. Another advantage of using deep learning approach will be the ability of the network to accumulate knowledge by training with multiple patient’s data over time.

Challenges for incorporation of deep neural network tools in medical management

While being capable of processing huge amounts of data to arrive at conclusions, based on the training they received, it is often difficult to intuitively understand why the DNN reached a particular classification decision. It is difficult to identify which datapoints or features of data, lead the DNN to take a particular decision on behalf of a specific input. This difficulty, in understanding the decision-making process of deep neural networks can make it difficult to incorporate these tools to the medical decision-making process. Therefore, better understanding of the DNN’s decision making process is imperative in seamless incorporation of these advanced technologies to clinical field. In addition, understanding the signal features contributing to pathological states can further our understanding of the basic pathophysiology of disease states like epilepsy.

Certainty regarding the decision of DNN

The DNN processes the data to arrive at decisions which can be used for classification tasks. The certainty of decision on each sample is usually unclear. This will be extremely important especially from a standpoint of medical decision making, which also take into account other factors including clinical and imaging data for arriving at conclusions. Therefore, having a measure of the confidence of the DNN on its decision would be extremely useful in comparing the different possible outcomes.

Relevance of information content of the signal that is used by the DNN in arriving at the decision.

The iEEG signal contains crucial information about the system which may influence the final decision of the DNN in both positive and negative ways. It is known that the DNN uses a complex non-linear process in arriving at decisions on individual inputs. This process depends on several factors including complexity of input data, number of DNN layers, type of activation functions and the original training data. Ensuring that the DNN utilized the entirety of the information content in the signal when it arrives at a particular decision would be extremely important from a reliability standpoint. In the case of EEG signals the semi-regularity of the signals need to be considered when estimating information content. This kind of a measurement can pave the way for comparing different DNNs by exploring how effectively they have captured the information content in the signal/ images. It would also be important to explore the parameter space of the information theoretical measures used for calculation, when evaluating the relevance of each these measurements in the final decision making.

The gradients of DNN outputs with respect to the activation maps and how they are related to the confidence of DNN decisions.

While the DNN decision is made over several layers of the network, it is important to assess the contribution of the activation maps in different layers. How these relationship changes over the multiple layers for each sample input into the network, would help in better understanding of the underlying process. Also, it is possible that the information content in the input signal is a critical factor in influencing the gradients of outputs with respect to activation maps. We explored these relationships in the context of iEEG signals.

Research questions and relevance

The main research questions asked were:

- Research question (RQ1): How to use deep learning tools to analyze intracranial EEG (iEEG) data to identify epileptogenic zones?

  Relevance: iEEG data acquired is a complex signal and evaluation through visual inspection alone, of this signal, is extremely limited and can potentially lead to erroneous decisions. Having standardized tools for evaluation of these signals is extremely important in patient outcomes.
- RQ2: What are the signal features that the DNN is identifying while detecting epileptogenic zones?

  Relevance: Having a better understanding of what signal features lead to the DNN's decision to classify the signal, will be important in incorporating these methods in clinical decision making. Further, this would improve our understanding of the basic neurophysiological processes.

- RQ3: What signal features are maximally noted in heatmaps?

  Relevance: Understanding the signal features that are prominent in the heatmaps would allow for correlation of the DNN decisions with the existing neurophysiological knowledge. Also, as mentioned previously, this would help in furthering the knowledge about neurophysiological processes. It may be noted that the signal features detected in the heatmaps could correspond to both physiological and pathological signals.
- RQ4: What measures can be used as surrogate for certainty of decision of DNN?

  Relevance: When DNN is trained, it comes up with average accuracy for validation. While that is a good measure of reliability of the DNN predictions, that does not give confidence measure about individual predictions. Having a measure of certainty for individual predictions would enhance the reliability of DNN predictions.
- RQ5: What measures in signal's information content can be used to assess the quality of heatmaps used to arrive at decision making?

  Relevance: The DNN being essentially an information processing machine, quality of the DNN decisions should be measured based on how effectively the DNN decisions are related to the information content of the original samples. Utilizing the information measures that best captures the features of the signal (regularity etc), may enable in assessing the DNN decisions.
- RQ6: How does the parameter space of signal's information content affect the certainty of decisions?

Relevance: There are multiple parameters that can be varied in information content measurement. Therefore, exploring that parameter space is important in finalizing the information theoretic approaches in quality assessment of DNN decisions.

- RQ7: How does the gradient of the DNN results with respect to the activation maps vary over the consecutive layers of DNN?

  Relevance: In DNN the information is processed through multiple layers and the activation of the various layers is the contributing factor to the final decision. Having an understanding the of how fast the DNN arrives at decisions would help in optimizing its architecture. This would also help in translating the network architectures various applications.

Also, from a practical standpoint we explored how the DNN model predictions match with the clinical decisions made after multimodal evaluation along with patient's history and clinical findings. This is important as the translation of the DNN methodologies to clinical applications is expected to improve the patient care while helping in bringing down the costs of medical care.

CHAPTER 2: LITERATURE REVIEW

Deep Neural Networks have been used in multiple fields including image classification[14,15], signal processing and natural language processing [10]. DNN based tools can be effectively used in classifying the EEG signals recorded from epileptogenic zones. Failure to accurately identify the epileptogenic zones in iEEG may be contributing to the inadequate seizure control reported event after epilepsy surgery, which ranges from 40-60% [8,16]. This raises the critical need for tools which can accurately analyze the iEEG signals and identify the epileptogenic zones. We explored the existing literature from the standpoint of DNN applications in EEG/ iEEG evaluation. We also reviewed the literature for the confidence measures on DNN outputs and role of information theoretical approaches in evaluating DNN. It is known that the information theoretic approaches can be utilized to improve the robustness of DNN results. The DNNs can compress the input data to lower dimensional representations which help them to come up with predictions. Therefore, exploring the relationship of information in iEEG with the processing and results of DNN would be important.

DNN for epilepsy detection in scalp EEG

The majority of the patient's suspected to have epilepsy undergo EEG study lasting less than 1 hour. One of the aims of this procedure is to evaluate the presence of interictal epileptiform abnormalities in the form of sharp waves, spikes, spike waves, polyspikes etc., which suggest that the patient may be having epilepsy. In some cases, this study can be prolonged and seizure detection tools are being developed to evaluate prolonged recording data.

Automated detection of the interictal discharges has been previously studied with various signal processing or computational tools. With the advancement of machine learning techniques these detection technologies have improved significantly.

Classification of EEG signal to ictal (happening during seizure) or interictal (signal from epilepsy patients that happen when they do not have seizures), was performed using instantaneous amplitude and frequency. The approach used a multivariate empirical mode decomposition to decompose the EEG to multiple intrinsic scales. This extracted data from EEG was passed through neural networks[17] to achieve the classification of the signal . This approach is computationally expensive given the decomposition being attempted. Another study employed the features extracted including approximate Entropy of the wavelet sub-bands, Hilbert Envelope of the sub-bands and wavelet statistical features for training machine learning algorithm[18]. This method classified EEG into normal, interictal and ictal states. It may be noted that this study used approximate entropy (of the entire signal) as input to the machine learning algorithm.

Another study evaluated deep learning based on CNN, on inputs given after signal transformations. The study was based on two public datasets including ictal and interictal EEG and their Fourier, wavelet and empirical mode decompositions, achieving 99-99.5 % accuracy in classifying non-seizure vs seizure recordings [19]. The study used a CNN model, and the data was fed as a 2D matrix. The study achieved an accuracy above 95% and a better performance was achieved in the case of seizure data with Fourier Transform, which was implicated to be due to the big difference in the energy distribution in the various frequencies in ictal recordings. On the other hand, raw data or empirical mode decomposition data gave better results in identifying focal vs non-focal signal arising from the epileptogenic regions. This study did not address the question of the signal features that the neural network learned.

The advantage of the above studies is that they are using non-invasive data which can be easily obtained from patients (although acquiring it continuously may have practical difficulties). The intention of the above studies was primarily identifying seizures from existing EEG. Those approaches may not be applicable in online seizure prediction. Also, given the limited spatial resolution, use of these approaches in epilepsy localization for surgical purpose is overall limited.

DNN for intracranial EEG analysis

Patients with intractable epilepsy are evaluated for possible surgery as noted in the previous section. These patients usually undergo intracranial EEG evaluation. This data is recorded form several electrodes simultaneously and in high spatial and temporal resolution. Because of that reason visually assessing this signal can be challenging and various automated approaches for this purpose have been described. The intracranial EEG based deep learning studies have variously explored the possibility of detection of interictal discharges, seizure detection, seizure prediction or seizure onset detection (in which case the challenge is to detect the onset of seizure from data being recorded live). Some of the analysis is done on already recorded data while some applications involve pathological changes from data acquired live from ongoing recording. All these methods represent various potential applications of the deep learning technology to address the multifarious challenges in this field.

Interictal discharge detection

One of the previous studies demonstrated that automatic feature generation based on deep learning was a useful tool for interictal epileptiform discharge(IEDs) detection [17]. Specifically, the meaningful features representing IEDs were automatically learned with CNN. Given the varied nature of the interictal signals, CNN based filters were used for IED feature extraction. It may be noted that the signal arising from the different regions of the brain may have different electrophysiological origins, may be undergoing different levels of filtering, which may be contributing to the different nature of the IED. Interictal discharges visually identified by experts were used in this study. A method of kernel convolution in one dimension was used in identifying the useful representations of the signal. The learning process of CNN was also explored by correlating the learning weights of convolutional layers with averaged IED. While it exemplifies the potential of evaluating and exploring IEDs, this study did not address the evaluation of DNN for identifying the electrodes located at epileptogenic zones.

Intracranial EEG transformed into spectrogram was used for training CNN, in a study for identifying intracranial interictal epileptiform discharges (iED)[20]. This strategy used 1000 intracranial EEG epochs randomly chosen from 307 subjects and annotated independently by two experts. The intracranial EEG was converted into a spectrogram. The model used a method of template matching algorithm and residual neural network architecture. The detector reported sensitivities between 91-100% with a mean accuracy of 0.94. This method was able to show significant improvement compared to template matching algorithm alone. At the same time the accuracy with external test set was noted to be relatively small (0.71) suggesting the need for training with larger datasets.

DNN strategies for seizure onset detection

DNN based methods have been proposed for seizure presence detection using the American Epilepsy society Seizure prediction challenge dataset[21]. This included intracranial EEG signals (iEEG) from five (5) dogs and two patients, with a total of forty-eight (48) seizures and a total duration of 627 hours of monitoring. In this study methods based on fusion of three CNNs and fusion of four CNNs gave 95 % accuracy value.

A one dimensional convolutional neural network combined with a random selection and data augmentation strategy has been described for seizure onset detection in long term EEG and intracranial EEG data[22]. Two different parallel 1D-CNNs are used simultaneously to learn high level representations. The classification results for each patient were evaluated at segment-based level and at event-based level. In the stacked CNN model, the EEG segments are sent to both blocks simultaneously and the proposed method achieved an accuracy of 99.54 % for stereoEEG (which involves placement of multiple depth electrodes) dataset. One of the advantages of this work was use of 1D-CNN(1-dimensional CNN) which avoided use of any additional preprocessing of the EEG signal. Also given the use of two parallel 1D-CNN blocks the network was able to learn different high-level representations at the same time.

CNN based methods have been used for analyze human EEG data to get better understanding on how brain behaves prior to seizures[23]. The iEEG data is converted to an image like format before processing. A multiscale CNN architecture was used for this purpose to learn the different representations of iEEG data. Short time Fourier transform was used to convert the iEEG to a two-dimensional representation, which displays the changing power spectra as a function of time and frequency. CNN model was trained by optimizing “binary cross-entropy” cost function with “Adam” parameter update. This method was noted to have Area under the curve (AUC) score of 0.84.

DNN for detection of epileptogenic zone in iEEG

A time-frequency hybrid network has been described for identifying focal or non-foal iEEG signal[24]. In this approach, short-time Fourier transform (STFT) and 1d convolution layers are performed on the input iEEG in parallel. Study was performed with Bern-Barcelona iEEG dataset. The dataset contains 3750 focal iEEG signal pairs and 3750 non-focal iEEG signal pairs. This allows for extracting features of time-frequency domain and activation maps. This method was achieved an accuracy of 94.3 %.

Another approach describes feature extraction based on entropies evaluated at different frequency bands, thereby creating a 2D feature map. Further analysis is performed using a CNN and the network is trained with binary classification. This study used two datasets: 1. Bern-Barcelona Dataset2. Juntendo Dataset. The Juntendo dataset was recorded from patients suffering from temporal lobe epilepsy caused by focal cortical dysplasia. The entropies calculated used Shannon entropy, Renyi entropy, generalized entropy, Phase entropy (two types), Approximate entropy, Sample entropy, and Permutation entropy. A loss function of “categorical_crossentropy” was used. A classification accuracy of 99.5% was obtained with this approach.

A multi-branch fusion model which identifies epileptic and non-epileptic signal has been described, considering the wave features and higher order features of the signal[25]. The two branches employed were bi-directional long short-term memory attention machine (Bi-LSTM-

AM) and 1D-CNN. The study employed 12 time-domain features and 6 frequency domain features. The study employed a Bi-LSTM with an attention machine, which was expected to learn specific features of individual patient's signals. The two LSTM's incorporated were arranged in such a way that they process the data in opposing directions. This study was able to achieve epileptogenic signal identification with high accuracy (97.6%). The study did not explore the decision-making features of the iEEG.

As discussed above various approaches have been employed to address the question of identifying epileptogenic regions from iEEG, with variable efficacy. These methods are generally limited by the duration for which the signals acquired or because of the lack of convincing data after surgery that the specific brain region was epileptogenic. Therefore, we have RQ1 to better evaluate this challenging question. In addition we have RQ2 to better understand how the DNN address the question of identifying epileptogenic regions. We are exploring the solutions for this question by evaluating the heatmaps of DNN outcomes.

Seizure onset detection in iEEG

A study employing responsive neural stimulator data (RNS), a device that is implanted to control the seizures, addressed the question of iEEG seizure onset detection with deep learning strategies[26].The study used 5226 ictal events collected from 22 patients implanted with RNS. The CNN was developed with an aim of providing personalized annotation from RNS data for the occurrence of seizures. This network used 23 convolutional layers. The inputs are time series of intracranial voltage measurements along with patient identifier. The DNN output include 1. Probability that the recording contains an ictal pattern and 2. onset of ictal pattern in seconds. Accuracy was evaluated based on concordance with expert opinion and an agreement in the range of 99.8 % was noted.

This study mainly questions the online detection of the seizures especially given the specific question being addressed by the RNS device. But it may be noted that identifying the abnormal pattern which is later emerging as seizure is important in this strategy as well. The question that we are asking on RQ1 is relevant for these kinds of applications as well since the identification of seizure patterns is an important component of this research question.

Measures of confidence in DNN predictions

Some of the major concerns of using DNN in safety-critical fields like brain signal processing include 1. the lack of transparency and expressiveness of the model [27] 2. The lack of measures to estimate the certainty of each prediction. The models are usually limited by two types of uncertainty which several studies have attempted to address in various ways. 1. Epistemic uncertainty which is the limitation of the model due to the lack of adequate knowledge. This primarily arise from the lack of adequate data for the training leading to a poor determination of the model parameters. 2. Alleatoric uncertainty, which arises from the stochasticity of the input data. In this case the best prediction possible will be a high entropy prediction.

Usually, neural networks do not give an estimate of certainty in individual predictions. Given any trained DNN model, the prediction of any new data presented will be associated with some level of uncertainty. This can be uncertainty caused by model or that contributed by data itself [28]. A method "distributive uncertainty" was described, which parameterize the prior distribution over predictive distributions. This method helps in distinguishing data and distributional uncertainty [29].

An approach using simple statistics from softmax distributions has been described to identify misclassification of data including computer vision, natural language processing, and speech recognition tasks [30]. This study addressed the question of error and success of prediction, on whether it is possible to predict if a classifier will predict a test example correctly or not. The second part of the study addressed if it is possible to predict if a test example is from within or out of distribution of the training data. This study showed that the prediction probability of incorrect and out of distribution examples tended to be lower than that for the correct examples. This study demonstrated the potential use of softmax prediction probability as a method for error and out-of-distribution detection for various types of data.

Mutual information and softmax variance has been described as a tool for estimating measures of uncertainty for adversarial example detection[31]. The study examined various measures of uncertainty for detection of adversarial examples and demonstrated that softmax variance can be seen as an approximation of mutual information. In addition, the measures of uncertainty were compared by projecting to lower dimensional spaces. This study also demonstrated that the dropout is not sufficient to capture the full Bayesian uncertainty, and therefore proposed an extension to the dropout schemes.

Another metric of certainty called attribution based confidence (ABC) has been described, which helps in deciding whether an output of DNN on a certain input can be trusted[32]. The theoretical motivation for this approach was from axios of Shapley values. The computation of ABC metric involved importance sampling in the neighborhood of high dimensional input using relative feature attributions. This method did not require access to training data or additional calibration. Further, the evaluated the method on MNIST and ImageNet data sets using out-of-distribution data, Adversarial inputs and physically realizable adversarial patches. This method uses a deliberative top-down approach adding a causal deliberative system. ABC employs feature attributions for dimensionality reduction and uses the importance sampling in the reduced-dimensional neighborhood of the input to estimate DNN model's conformance. The features identified are the locally relevant ones for a given input.

The uncertainty approaches in general try to give an overall estimate of the uncertainty of predictions. While they are helpful in estimating the confidence one can have, in the outcomes of the model, there is no specific score on the reliability of an individual prediction, for example the classification of an EEG signal. Therefore, we have RQ4 were we try to evaluate the DNN outputs in the penultimate layer to estimate the confidence of predictions. We introduce a score for the reliability of an individual prediction called certainty index. It is an index how certain the given model is about a specific classification output. The certainty of the prediction is assessed through the differences in the outputs prior to softmax layer.

Heatmapping for identifying the relevant regions that DNN is focusing for decisions.

The exact features of a signal or the image that the DNN uses to arrive at a decision has been a major question in the field of deep leaning [33] [34]. This will be a very relevant question especially when a new technology is being adopted in the medical field. The rationale behind the

DNN predictions may have to be interpretable and relatable for the user employing this tool to arrive at a particular decision.

A common method for assessing the relevance of input features to the decision has been saliency maps or heatmap. These maps identify the most relevant input features that caused maximum response in the DNN to arrive at a particular decision. The most common visualization methods of heatmapping have been classified into two categories: 1. Based on backpropagation method, which uses gradient signal passed from output to input 2. Perturbation based methods which employs selective removal or alteration of input features and estimate outputs based on the new features, thus enabling estimation of the relevance of each input feature.

One of the backpropagation methods described involves global average pooling of the activation maps formed in the final convolutional layer [35] The activations from the convolutional layers is multiplied with the weights from the fully connected layer, generating class activation maps (CAM). Finally, the map generated in this way is up-sampled to match the size of the image or signal. In the case of images (in which the original study was performed), CAM allows for visualization of the predicted class scores on a given sample image, highlighting the discriminative regions that the CNN detected.

Heatmapping approaches and Grad-CAM Heatmapping methods are used for evaluating the contribution of different regions of the data to the decision-making process. In this method activation maps are calculated based on the gradients of the network output with respect to the last convolutional network. Various other methods of heatmapping have been described like sensitivity analysis of neural network using partial derivatives [36], deconvolution methods [37] and layer-wise relevance propagation (LRP) algorithm [38].

Explainable deep learning models

Given the concerns of black box nature of DNNs more recent studies have focused on explaining the decision making of the DNN using various approaches. The various heatmapping approaches have been used for this purpose.

In one of the deep learning based studies for evaluation of cardiac arrhythmia, a method of Local Interpretable Model-agnostic Explanation(LIME) was used[39]. The study used a hybrid 1D-CNN model that combined 1DpCNN and Gated recurrent Unit (GRU). The heatmapping approach, LIME represents a model that uses a perturbation technique to generate a new dataset by manipulating the instance features. Then the machine learning model is used to make new predictions based on the new onset. Subsequently LIME trains an interpretable liner model on new dataset to generate explanation. In this method the LIME was able to identify the regions of Electrocardiogram (EKG) including QRS complex, P-wave and T-wave, which are known to be essential in interpretation of arrhythmia. The maximal hotness regions were noted to be around QRS complexes which is known to be associated with maximum types of arrhythmias. This study indicated the essential benefit and potential of using heatmapping strategies to explain the DNN.

Another approach of explainable EKG interpretation is reported using Grad-Cam algorithm for 3-lead EKG classification[40]. A lead-wise Grad-CAM approach was used in explaining the predictions of this model. Modified convolutional layers were used to capture longer patterns of EKG signals. The standard convolutional layers were replaced with Depth-wise Separable Convolutional layers. This was done to reduce the number of parameters in the model. An attention module was implemented to more effectively ensemble the features together. Given the attention module this model has an additional importance parameter. Therefore, the attention parameter was included in the Grad-CAM algorithm in addition to the pre-existing Grad-Cam generated. The results indicate which leads specifically contribute maximally to the classification of signal to a particular category. This is similar to identifying the localization of epileptogenic zone in the iEEG signals.

Estimating the information content of the signal

Brain represents one of the most sophisticated devices processing information. EEG signal is a marker of brain's information processing, as recorded on the surface of the brain or intracranially as iEEG. In the case of a series of data, the information content of the individual elements is estimated through the inverse of the probability of occurrence of each element. This idea was introduced by Shannon as information entropy and since then, this measure has been employed in estimating the complexity of the signals[41]. Later an entropy measure to assess the entropy of dynamical systems was introduced, described as Kolmogorov-Sinai entropy[42]. While this method is well applied to the real dynamic systems, presence of noise can derange these calculations easily with values going to infinity. Approximate entropy was introduced later which was capable of capturing the changing complexity of signals[43] which also quantify the regularity in the data employing the idea of pseudo-phase. A larger value of the approximate entropy corresponds to higher complexity of the signal. A low Approximate Entropy value is representative of the fact that the system is very persistent, repetitive and predictive and the patterns having repetitions all through the signa[44]. While Approximate entropy helps in evaluating the nature of the generating system, it does depend on the length of the records and can underestimate for shorter signal.

Some of the disadvantages of Approximate entropy includes lack of relative consistency and dependence on the length of data series. Another concern of the Approximate Entropy is that it leads to results which suggest more regularity than that in reality. This is because of the algorithm, which allows each vector to count itself. Sample entropy algorithm removes this self-counting issue and therefore avoids the false results from that approach. SampEn measure does not depend on the signal length (or is less dependent) and has higher relative consistency [45], [46].

Approximate entropy (ApEn) and sample entropy (SampEn) has been used in Neural respiratory signal processing[47] .This study also explored the role of embedding dimension and thresholds of calculation in the estimation of both measures. They also compared the two approaches with both simulation and experimental data. This study suggested that Sample entropy showed more consistent results compared to approximate entropy. Calculations of Approximate Entropy was done for three types of data (1. Phrenic nerve discharge from in-vitro arterially perfused adult rate, 2. Same type of data from in vivo alpha-chloralose anesthetized rat and 3. Simulation data). The analysis revealed the critical importance of the threshold of approximate entropy estimations. The ApEn values were very close to each other when the threshold chosen was less than 0.1. When the threshold was chosen more than 0.1 there was clear separation of the three classes of data. A similar impact on the estimations of ApEn and SampEn was noted for variations in the sampling rate and embedding dimensions, highlighting the importance of exploring these parameters spaces in the estimation of ApEn and SampEn.

Another study with RR interval in Electrocardiogram recordings (RR interval is the time interval between two R waves in the Electrocardiogram data), explored the role of embedding dimensions, and thresholds in estimating the measures of SampEn and Fuzzy Entropy (FuzzyMEn). They explored the statistical significance between normal sinus rhythm and congestive heart failure group and demonstrated that choosing the parameters including embedding dimension, tolerance threshold and time series length plays critical role in the outputs of SampEn and FuzzyMEn. That study also demonstrated that the FuzzyMEn demonstrated better relative consistency for distinguishing the two groups of data.

An approach of using ApEn for automatically distinguishing seizure EEG from normal signal has been described [48]. This study used a combination of ApEn values and recurrence quantification analysis (RQA) as inputs to Convolutional Neural network (CNN) for automatic detection of EEG, demonstrating the relevance and utility of entropy measures in the deep learning approaches. EEG data was taken from the Bonn database, with each set containing 100 clinical intracranial EEG recordings. ApEn was calculated based on Chebyshev distance (details on the calculation of ApEn is noted in the subsequent sections) and threshold r= 0.15 *SD. The RQA method uses a non-linear index quantification method, which obtains quantitative geometric measures. The CNN implemented used a loss function of “binary_crossentropy”, activation function of “relu” and optimizer “adam”. The ApEn measures indicated significant differences between normal and epileptic EEG. A high accuracy was achieved in this approach for classifying seizures (99.26%).

Another approach of using ApEn in Elman and probabilistic neural networks is described for automatic detection of epilepsy[49] from EEG data . The method used the hypothesis that the ApEn drops sharply during epileptic seizures. This approach was able to achieve a high accuracy with a low computational burden.

While the various potential applications of information measures to boost the quality of predictions of DNN have been described, exploration of the role of information content of the signal in the decision-making process of the DNN has been limited. It is in this context we have RQ5 which uses sample entropy measures to assess the quality of the heatmaps. Heatmaps being the representation of the regions where the DNN is focusing, we aim to identify how the information content and heatmap is correlated. Further we explore the question of how the information measurement vary in the parameter space of SampEn calculation (RQ6), to establish the robustness of the results noted on answering RQ5.

Gradients of DNN and relation with certainty

Gradient descent is one of the most widely used algorithm for optimization for neural networks[50]. While various other optimizing methods exist, stochastic gradient descent is commonly applied in the case of various DNN implementations. A direct evaluation of how the gradients in various layers are related to the information content of the input signal and the certainty of the decisions of the DNN has not been reported.

Gradients generated from loss function are driving the training of the network. Therefore, evaluation of how the gradients are changing across the network, and how those changes are related to the certainty of the decisions of the network, in individual cases will be important.

Evaluation of the gradients of cross entropy loss has been explored in the past [51]. This paper explored the properties of cross-entropy gradients and evaluated the impact of approximation of gradients. They used a approximations which were noise free and maintained a fixed length to avoid vanishing gradient problem. The study demonstrated the geometric properties of cross-entropy loss function.

A method of training deep neural network using a Mutual Information (MI)-driven, decaying Learning Rate (LR), Stochastic Gradient Descent (SGD) algorithm has been reported[52]. MI between DNN outputs and inputs are estimated and used to set the learning rate. A layerwise learning rate is set using the mutual information through the training cycle. This approach highlighted the advantages of information measures in DNN learning.

Another strategy reported, calculation of independent activation function for each neuron as a piecewise linear functions[53]. The parameters for these functions are calculated through gradient descent. The study demonstrated the potential of diverse activation functions, highlighting the importance of activation gradients in the overall performance. They did not specifically address if these functions change over layers of DNN in any specific way or if the changes are related to the information content in the input signal.

In this study we have used Grad-CAM algorithm for evaluation of activation gradients. The Grad-Cam approach has been further improvised with introduction of Grad-CAM++, with an intention of providing better localization [54]. This approach was developed to overcome the concerns that gradients at each feature map may not provide a good localization of the regions of interest, if the signal of interest appear in various places in the sample. A solution was proposed with incorporation of weighted average of the activation gradients. While that approach may give a finer grain map of the activation gradients, it is not aimed at assessing how the activation gradients change across the layers and how those changes are influenced by information content of the signal.

These approaches suggest the possibilities of using activation gradients derived from GRAD-CAM method in better ascertaining the learning process of DNN. We explore how the activation gradients change across the various layers of DNN and evaluate its relationship with the confidence of the DNN in individual results. The rate of change of activation gradients across layers is an important factor in the training of deep neural networks. By understanding the factors that affect the rate of change of activation gradients, it is possible to train deep neural networks that are more effective and more stable. Hence, we have RQ7 to evaluate the changes in the activation gradients across layers. We assess how the activation gradients change across the layers of DNN and how this change is related to the confidence of the DNN decisions. Further, we explore these changes in relation to the information content of the input signal.

CHAPTER 3: MODEL FOR PREDICTION OF EPILEPTOGENIC ZONES

[This work was published in the article “Deep neural networks and gradient-weighted class activation mapping to classify and analyze EEG”, DOI: 10.3233/IDT-228040, Intelligent Decision Technologies, vol. 17, no. 1, pp. 43-53, 2023 [55]]

The initial part of the study involved design of a deep neural network for prediction of the epileptogenic zones in the iEEG recorded from the brains of the patient’s undergoing evaluation for epilepsy surgery.

iEEG data for analysis.

The brain signals (intracranial EEG) from 10 patients undergoing epilepsy surgery evaluation, recorded using intracranial electrodes, continuously for several days using Nihon Kohden (NK) software, was used in this analysis. The electrode placement is performed in the operating room after which patients are admitted to Neuro-Intensive care Unit (NeuroICU) for prolonged iEEG monitoring. The electrode placement for a typical patient is depicted in a 3D-image in Figure 3. The placement of electrode in individual patient differed, which was decided based on pre-surgical evaluation. As noted in the figure some patients had grid type electrodes implanted while depth type electrodes (also called steroEEG) were implanted in some other patients. These patients had epileptic activity arising from various regions of the brain, including temporal cortex, hippocampus, insula or temporal-occipital cortex and were noted to have significant improvement in seizure control after surgery in these regions, proving that these locations were critical in epileptogenesis. The signals for the purpose of analysis were selected from interictal periods (when the patients did not have any identified seizures). For consistency the data was collected immediately after midnight on the second day after surgery. Collecting the data at midnight minimized the chances capturing evoked potentials. Also, this approach minimized the chances of having noise in the signal.

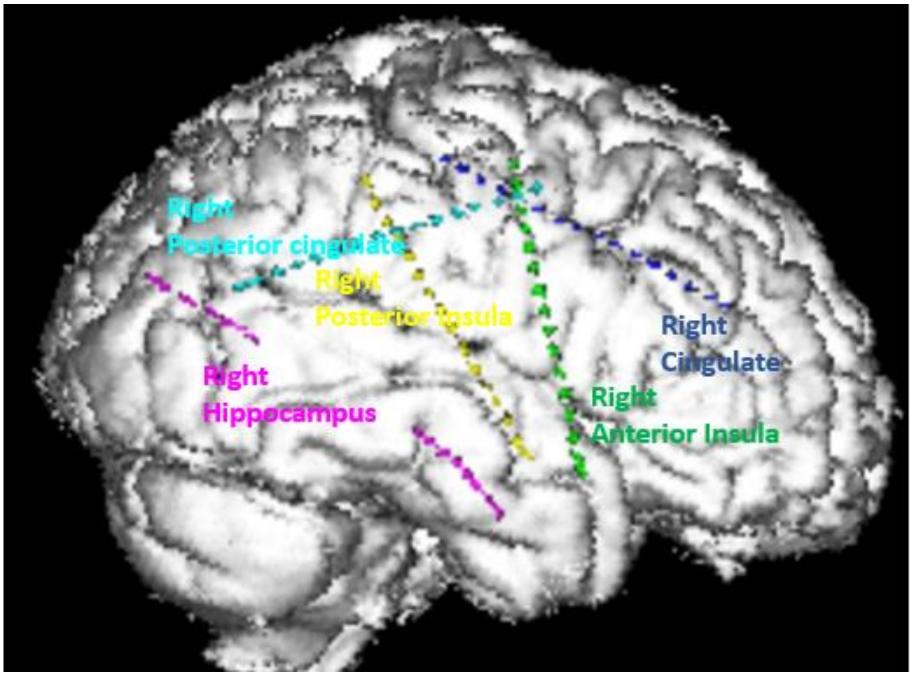
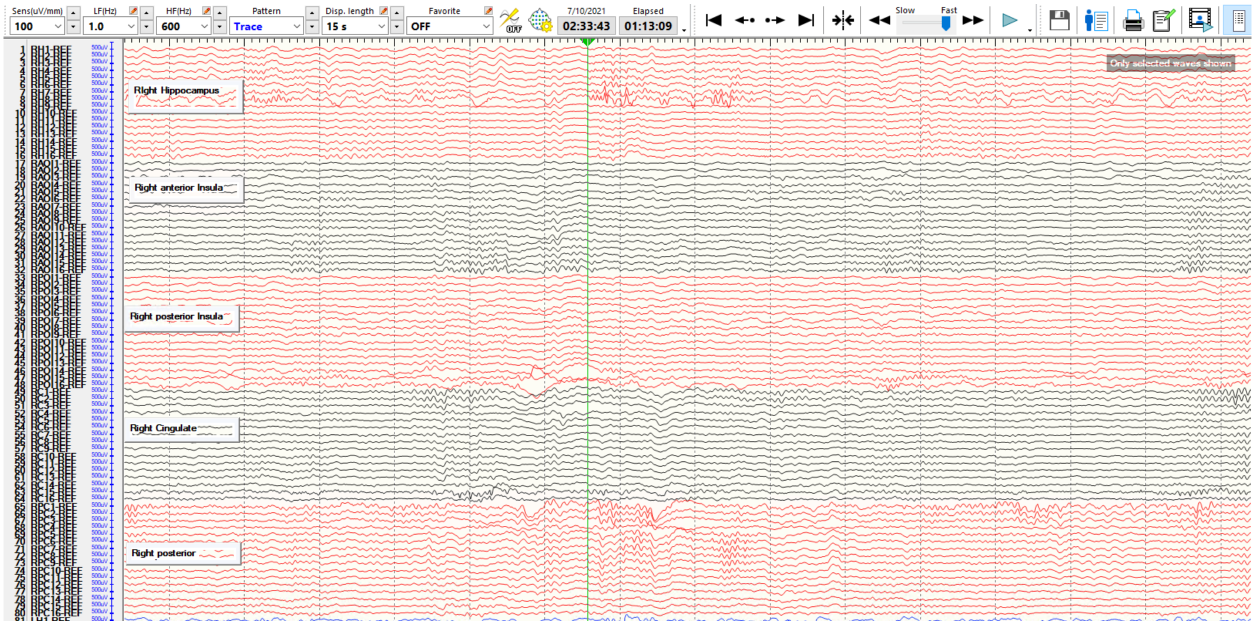


Figure 5. The electrode locations.
The positions of the electrodes inserted for iEEG monitoring. This image is generated by co-registration of CT scan images after electrode placement with 3-D reconstruction of the previously acquired MRI images. The individual colors represent the electrodes in one shaft of electrodes with the corresponding color-coded name.

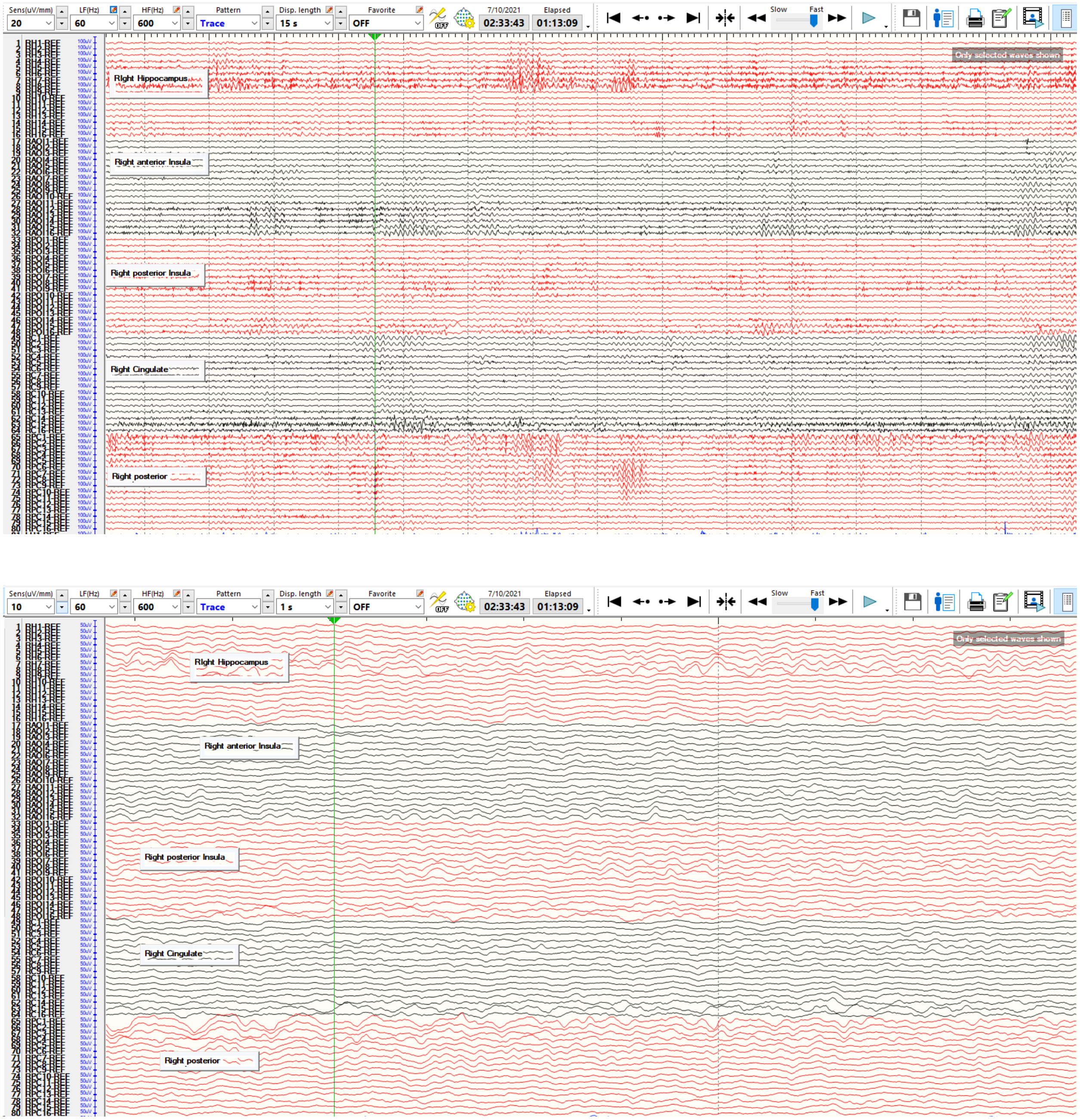


Sens(uV/mm) 100
LF(Hz) 1.0
HF(Hz) 600
Pattern Trace
Disp. length 15 s
Favorite OFF
7/10/2021 02:33:43
Elapsed 01:13:09

Figure 6. Collection of raw iEEG data.
The upper panel shows the placement of the electrodes. The labels correspond to the same-colored electrode. The lower panel shows the data collected from all the electrodes shown. It may be noted that this signal contains the broad range of frequency.

Preprocessing of data

The signals were evaluated in bipolar montage and exported to the European data format (EDF) format from NK software. The bipolar montage used allows for minimizing any contribution from noise. It may be noted that the intracranial EEG data usually does not have significant noise concerns (or minimal compared to the scalp EEG), but at times it can be contaminated by common noises (e.g. machine artifact) that affects all the channels. One minute data was collected from each patient, one day after surgery for placement of the electrodes. Only high frequency components of the signal (60-600 Hz) were evaluated in this analysis. There were two reasons for selecting the high frequency component of the signal, which included 1. minimizing the contribution of any potential artifacts and 2. the current knowledge that the higher frequency components tend to be more associated with the seizure onset zones. Signals of 1 minute duration was collected from each electrode. This data was parsed into one second duration (2000 samples) along with capturing the information about which category the particular electrode belongs to (epileptic vs non-epileptic). As noted previously, the information on whether a particular electrode location is epileptogenic is determined by whether the patient who undergo surgery in that particular location became seizure free after surgery. Individual samples of one second duration along with category information was pooled together from all the 10 patients and shuffled to avoid bias towards any individual patient data. A total of 10000 samples were randomly chosen from this pool (with equal representation of epileptic and non-epileptic data).



Sens(uV/mm) 20 | LF(Hz) 60 | HF(Hz) 600 | Pattern Trace | Disp. length 15 s | Favorite OFF | 7/10/2021 02:33:43 | Elapsed 01:13:09

Sens(uV/mm) 10 | LF(Hz) 60 | HF(Hz) 600 | Pattern Trace | Disp. length 1 s | Favorite OFF | 7/10/2021 02:33:43 | Elapsed 01:13:09

Figure 7. High frequency filtered version of the original signal shown in different scales. The upper panel shows 15 seconds of data, and the lower panels shows 1 second of data.

For evaluating the real-world outcomes, the 1-minute data was passed through the model after the same preprocessing as noted above and the electrodes which get categorized as maximum number of times as epileptogenic, was identified. These electrodes were compared with the actual electrode locations where surgery was done as noted in Figure 18.

There have been efforts to understand what DNN has learned in the process of training. This included methods to estimate what individual neuron or a layer of neural network in a DNN has learned. These methods significantly depend on the implementation of the network, starting from the type of neuron and layers incorporated in the network.

Deep neural network implementation

In the initial exploration of the use of deep neural networks for classification of iEEG two forms of deep neural networks were implemented for training with the data. A dense network was implemented as noted in Figure 9, with the parameters for individual layers noted.

The neurons in the case of general DNN tasks are implemented using the following formula, for input layer represented by $x= [x_1, x_2 ... x_i]$.

$$z_{ji} = x_i w_{ij}; \quad z_j = \Sigma_i z_{ij} + b_j; \;\; y_j = g(z_j) \qquad (1)$$

Where z_{ji} represents the contribution of each node along with the connecting weight (w_{ij}) to the next layer, and g(.) stands for the mapping function. A scheme of implementation of this equation is depicted below.

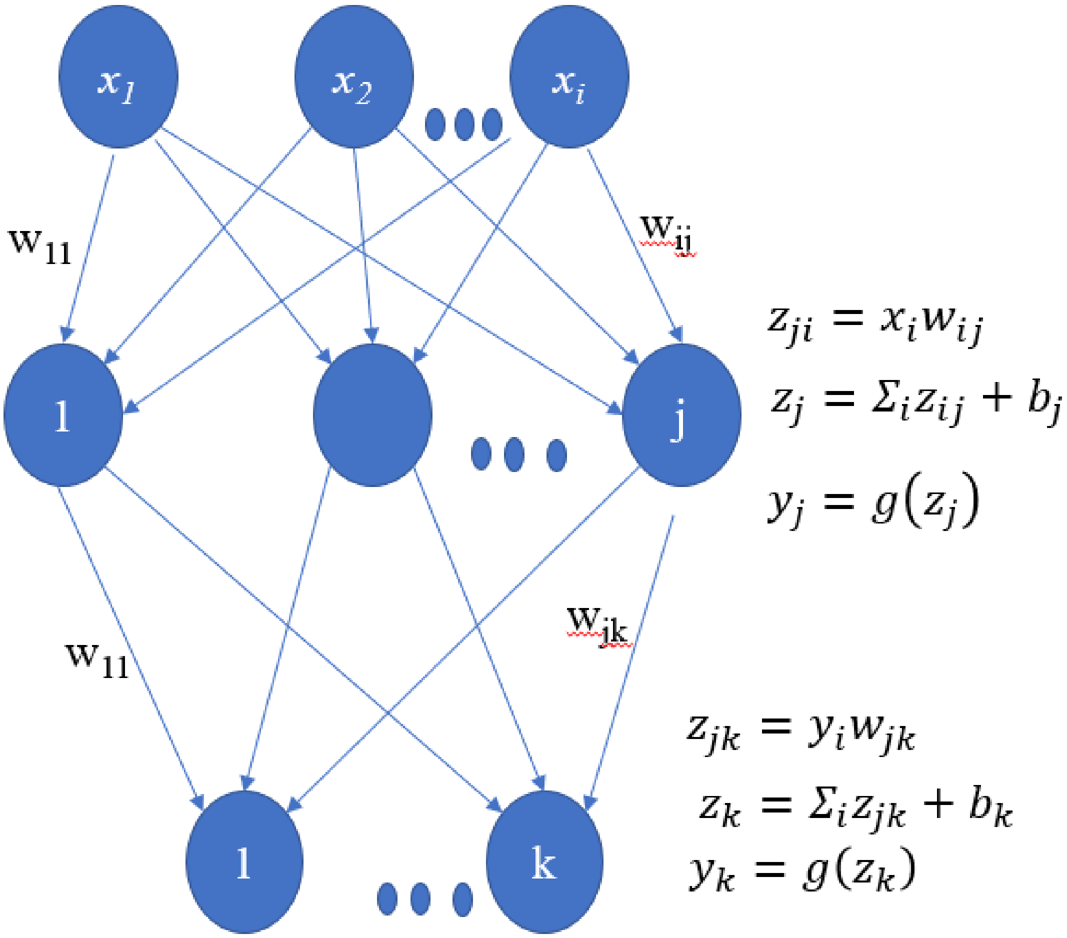


Figure 8. Schematic representation of the implementation of basic neural network with the weights and biases.
The calculations involved in each step is shown on the right side.

The input layer had 2000 nodes which correspond to the total data elements in the individual training sample. A total of 9 layers were included in the design, with binary cross entropy used for loss function and “Relu” as activation function. The final layer had two nodes which corresponded to the two classes.

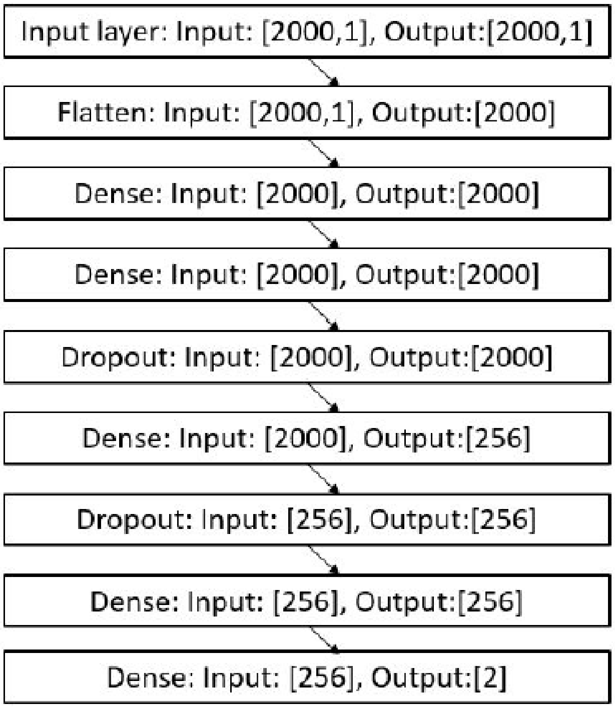

Input layer: Input: [2000,1], Output:[2000,1]
Flatten: Input: [2000,1], Output:[2000]
Dense: Input: [2000], Output:[2000]
Dense: Input: [2000], Output:[2000]
Dropout: Input: [2000], Output:[2000]
Dense: Input: [2000], Output:[256]
Dropout: Input: [256], Output:[256]
Dense: Input: [256], Output:[256]
Dense: Input: [256], Output:[2]

Figure 9. The dense network model implemented for evaluation of the intracranial EEG.

Also, a 1D-convolutional neural network is implemented with the neurons represented by the formula (1),with structure as noted in Figure 10. The initial layer corresponds to the number of elements in the signal (2000). One additional convolutional layer was present in the model along with subsequent drop out layers. A drop out % of 0.33 was used in after the first two convolutional layers. One dimensional CNN offers faster training speed compared to higher dimensional CNNs. In general, the first layers in 1D CNN would act as a local signal detector More higher order features are detected in the subsequent convolutional layers. While it is unclear, what components of the signals contribute to the activations of subsequent layers, it is possible that the higher frequency components of the iEEG or burst of epileptiform activity may be contributing to the activations in the subsequent layers. It can be assumed that the deeper layers may capture more information about a specific class of signal. The convolutional layers were followed by a dense layer, max-pooling layer, flattening layer and additional two dense layers.

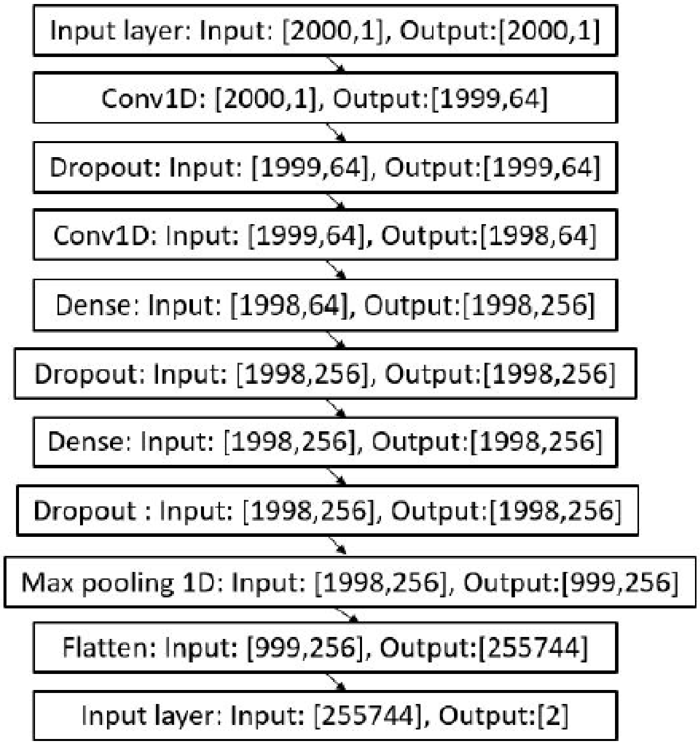
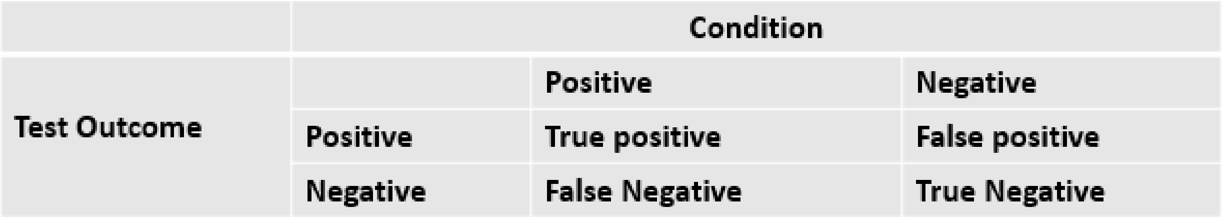

Input layer: Input: [2000,1], Output:[2000,1]
Conv1D: [2000,1], Output:[1999,64]
Dropout: Input: [1999,64], Output:[1999,64]
Conv1D: Input: [1999,64], Output:[1998,64]
Dense: Input: [1998,64], Output:[1998,256]
Dropout: Input: [1998,256], Output:[1998,256]
Dense: Input: [1998,256], Output:[1998,256]
Dropout : Input: [1998,256], Output:[1998,256]
Max pooling 1D: Input: [1998,256], Output:[999,256]
Flatten: Input: [999,256], Output:[255744]
Input layer: Input: [255744], Output:[2]

Figure 10. Schematic of Convolutional neural network implementation.

The neural networks were trained with 90 % of the data and validated with 10 % of the same data. Total number of training epochs was 100. A confusion matrix was generated to show the different components of the results as noted in Figure 11:



		Condition	
Test Outcome		Positive	Negative
	Positive	True positive	False positive
	Negative	False Negative	True Negative

Figure 11.Confusion matrix schematic.

Generating heatmaps of DNN predictions

One of the challenges of incorporating the deep learning-based tools in the medical field is the black box nature of the DNN. We explored the reasons for the DNN predictions using heatmaps and later correlating heatmaps with the signal features. The depiction of the input along with contribution of different regions of the signal in this decision-making process, is referred as heatmap and can be estimated for each layers using various algorithms.

Heatmaps can be generated using [36]a sensitivity analysis of neural network using partial derivatives [37] deconvolution method [56] the layer-wise relevance propagation (LRP) algorithm [38] and gradient-weighted class activation maps (Grad-CAM)[57]. Grad-CAM uses the gradient information flowing to the convolutional layers, thereby assigning the relevance values to individual neurons. This can be particularly useful in analyzing the classification of iEEG, given the relevance of transients in the classification process. Neurons in the convolutional layers look for class specific information (parts of signals, transients etc.). Heatmap of the convolutional neural network was generated from 1500 samples with grad-CAM algorithm as described above. Correlation coefficient between the heatmap signal and analytical signal was estimated to assess the similarity between signals.

Class-discriminator map defined as Grad-CAM is evaluated for any class c, as follows [57]. The gradient of the score for class c as given by y^c , with respect to the feature map activation Ak of a convolutional layer is given by

$$\alpha_k^C = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial y^C}{\partial A_i^k} \tag{2}$$

y^C is the DNN output for a particular class *c (before softmax)*.

α_k^C indicates the importance weight of *k*-the filter for class *c*.

A_i^k is the *i*-th element in *k*-th activation map.

N is the number of elements in feature map.

With the gradient weighted class activation map (grad-CAM) for a layer obtained as

$$L_{Grad-CAM}^C = Re\,L\,U(\Sigma_k \alpha_K^C A^k) \tag{3}$$

A typical heatmap generated using Grad-CAM algorithm for a typical iEEG segment is shown below in Figure 12.

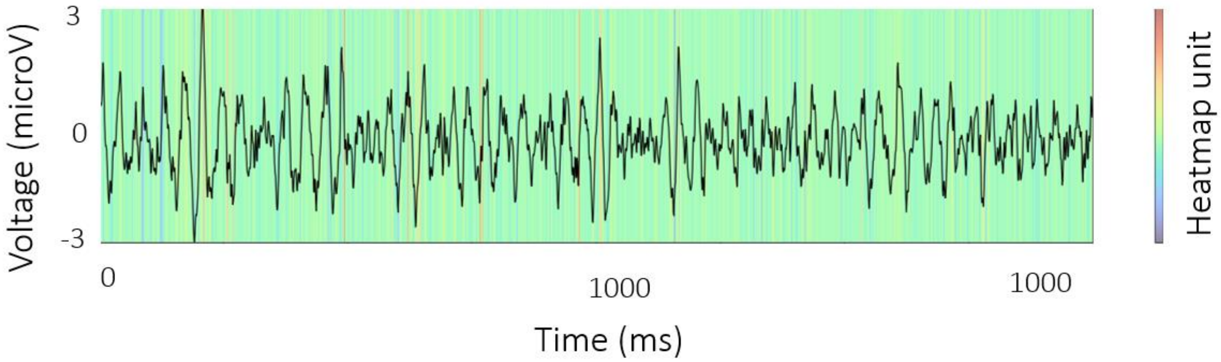


Figure 12. Heatmap of typical iEEG.
The heatmap is shown in colormap with colors close to red indicating regions of the iEEG signal that contributed maximally to a decision on DNN classification.

Calculation of analytical signal

Conventional evaluation of the iEEG signals involve visual inspection. This employs clinical expertise and signals are evaluated for transients, high frequency components etc. Having an understanding of, which visually discernible local features are captured by the CNN will help in incorporating the results of the model in clinical decision making. Various signal processing tools can be used to analyze signals. One of such tools is Hilbert transform, which allows for calculation of analytical signal. It functions as a broadband phase shifter which provides rotation of the initial phases of all frequency components of signal by an angle of $\pi/2$.[58]

The instantaneous envelope of the signal was estimated as analytical signal with the help of Hilbert transform of the original iEEG signal as follows.

$$z(t) = z_r(t) + jz_i(t) = x(t) + jHT\big(x(t)\big) \quad (4)$$

Here *x(t)*—real valued signal

z(t)-analytical signal

z_r the real value component and

z_i the imaginary component of the signal.

Where HT of x(t) is defined as (x(t)) = 1

$$HT\big(x(t)\big) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(k)}{t-k} dK \tag{5}$$

This is a signal with no negative frequency and in continuous time, every analytical signal z (t) can be represented as [24]

$$z(t) = \frac{1}{2\pi} \int_{0}^{\infty} z_{\omega} e^{j\omega t} d\omega \tag{6}$$

The analytical signal generated is representative of the instantaneous envelope of the signal. To evaluate the relationship of heatmap with analytical signal and original signal cross correlation was calculated between those pair. A high correlation between the heatmap generated using the Grad- CAM and analytical signal would suggest that the heat map is capturing this envelope of the signal as represented through analytical signal in classification process. This cross correlation is estimated as follows(*7*).

$$z(k) = \sum_{l=0}^{\|x\|-1} x_l * y_{l-k+N-1}^{*} \tag{7}$$

Where, Z(k)- cross correlation series

x- Heatmap as time series

y- Either original signal or analytical signal as calculated above.

N = max(x,y)

K-fold cross validation

Given the variance in the data, it is important to evaluate how the results change when the training and test data changed. A ten-fold cross-validation was employed for assessing the consistency of accuracy of the model [59]. Initially, the iEEG data were randomly divided into K (in this case ten) equal portions. Nine out of ten portions of iEEG signals were used to train the DNN and the remaining one-tenth of the iEEG signals was used to test the model. The above strategy is repeated ten times by shifting the test and training dataset. The average accuracy along with the standard error was reported.

Schematic of the data analysis from generation of model to heatmap based analysis

Given the multiple steps involved in this analysis a schematic of the analysis was developed to represent the types of analysis performed as depicted below.

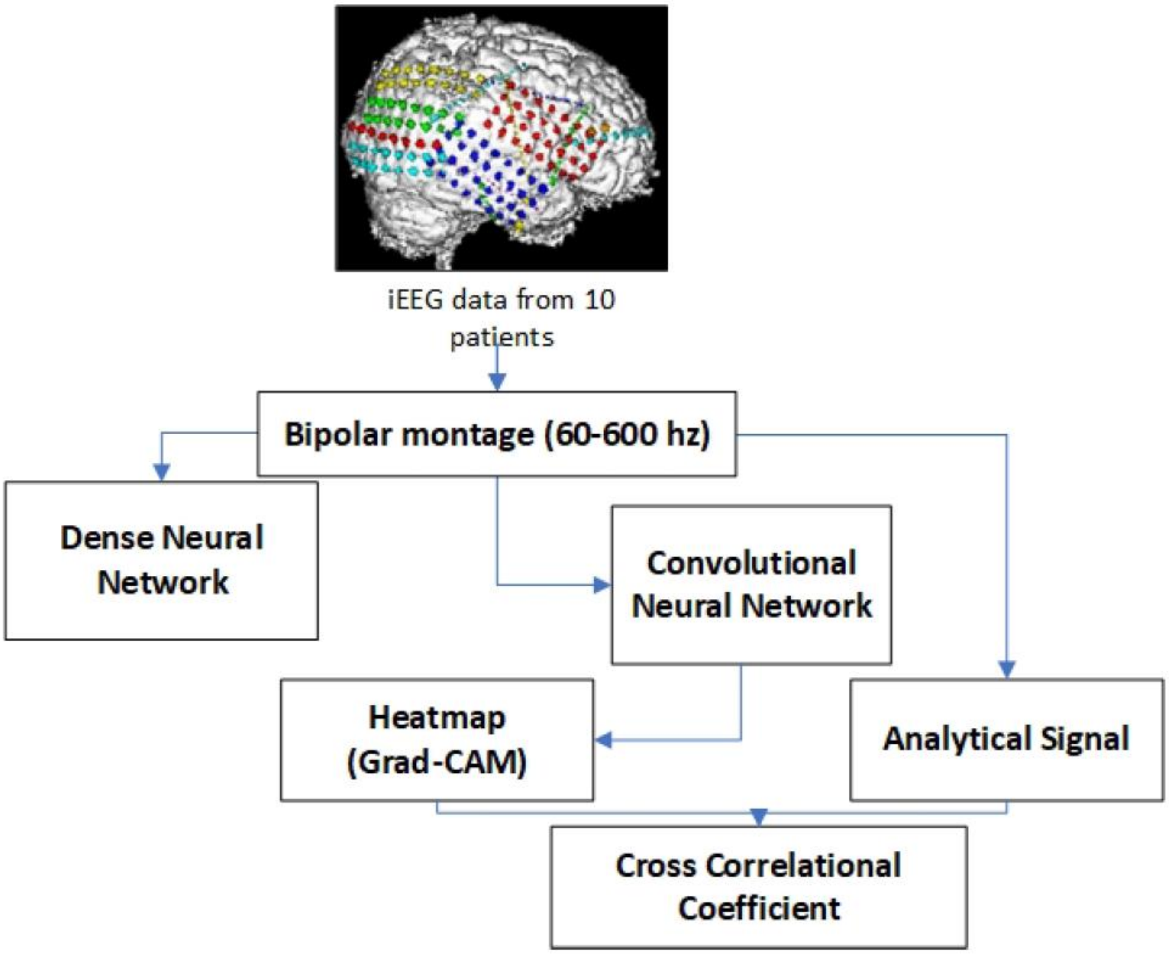

iEEG data from 10 patients
Bipolar montage (60-600 hz)
Dense Neural Network
Convolutional Neural Network
Heatmap (Grad-CAM)
Analytical Signal
Cross Correlational Coefficient

Figure 13. The scheme of data analysis.
The schema of research used in developing a model for classification of iEEG signals to epileptic and nonepileptic signals.

Data cleared from 10 patients for 1 minute during interictal periods processed through bipolar montage and filtered to 60-600 Hz frequency is used for the analysis. Two types of deep neural networks were used: A. Dense neural networks B. Convolution neural network which is a deep neural network with convolutional layers. The output from convolutional neural network used for calculating heatmap with Grad-CAM algorithm. The original signal is used to calculate the analytical signal using Hilbert transform. The heatmap generated with the use of Grad-CAM is cross-correlated with the analytical signal to estimate cross-correlation coefficients.

Evaluation of real-world data

To evaluate the application in the real-world data iEEG signals acquired from 15 patients (of which 10 patients were included in the initial modelling, but data was recorded from different time points), was processed with the model. One minute data from each patient was acquired leading to a total number of Number of electrodes x 60 samples the number of times each electrode was identified to be epileptogenic by the model was calculated and the electrodes were sorted according to that number. The top 10 electrodes thus identified was compared with the electrodes clinically identified to be epileptogenic by the clinical team.

Results

Given the initial motivation of the study was performed in two sections. Initial part of the study was targeted at developing a model to predict the epileptogenic regions, in patients undergoing iEEG monitoring for epilepsy evaluation. Once this was achieved, we compared the model results with real world data which established strong concordance with clinical results. This led to the necessity of developing understanding the backbox character of the DNN, which was explored using heatmaps and we established the relationship between heatmaps and signal characteristics.

Model for prediction of the epileptogenic zones in iEEG

The data was obtained from 10 patients during interictal periods, for a duration of 60 seconds, filtered between 60-600 Hz. The data from all 10 patients were pooled together, shuffled keeping the individual electrode labels. Two types of deep neural networks were designed as shown in the figure 2. The network is trained with 90% of the iEEG data for 100 epochs and the weights and biases are saved as model file and validated with 10 % data. The trained network had an F1 score of 0.99 in case of dense neural network and 0.87 in case of convolutional neural network. The accuracy, loss and confusion matrix are shown in Figure 15. .

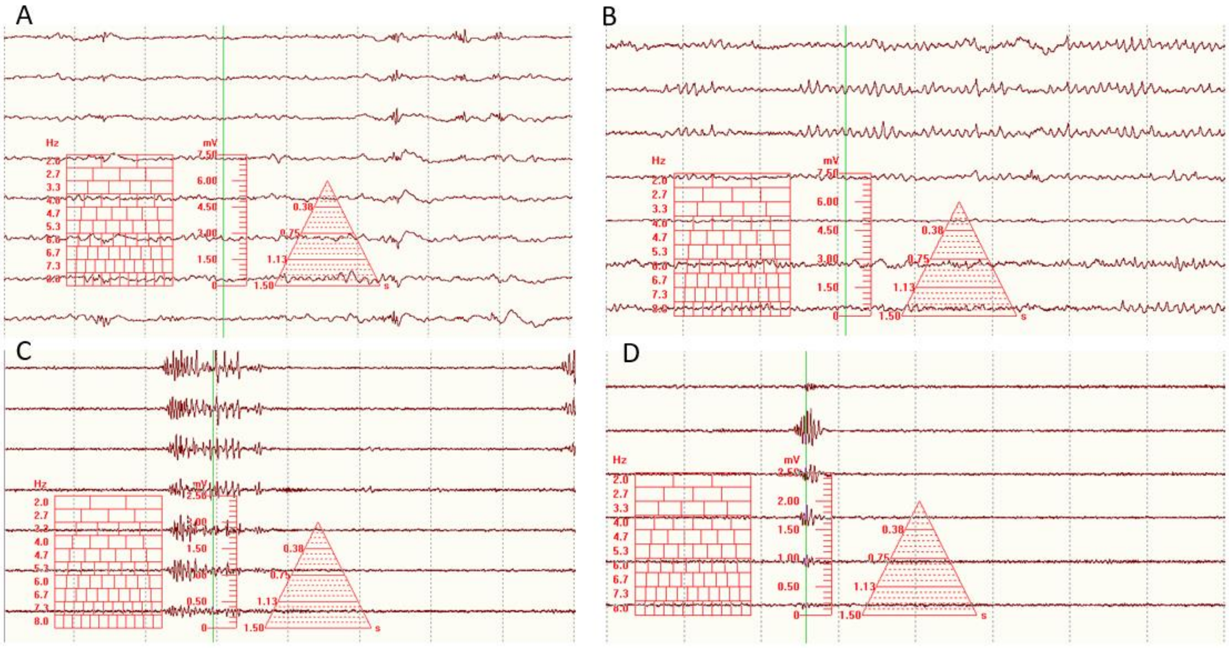


Figure 14. Samples of signal used for anlysis showing the amplitude and frequency content in the signals along with the various signal features.
The scales are also included in each case. A& B: Broad band signals. C&D: High frequency signals.

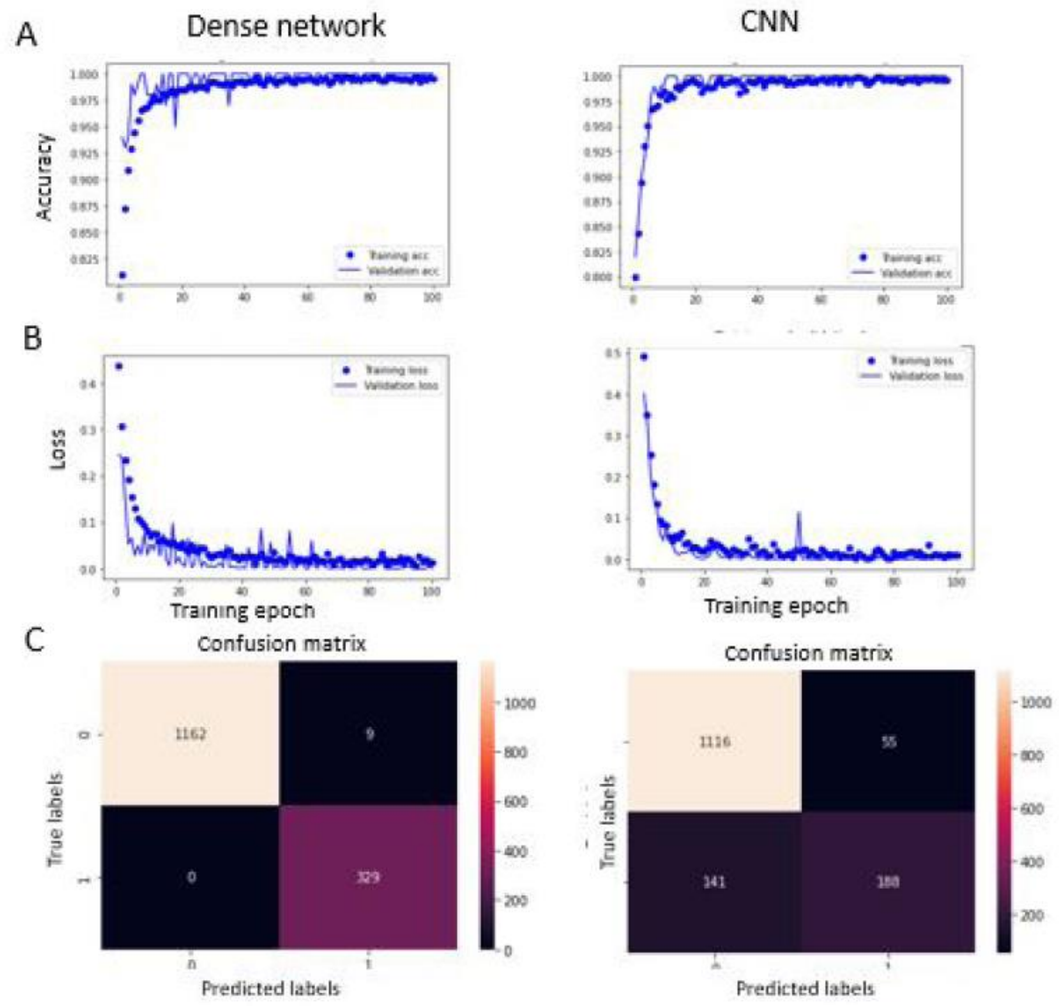


Figure 15. DNN prediction results for dense network and CNN.
A. The change in accuracy (training and validation) during the training epochs for a total of 100 epochs. B The change in loss over the course of the training, C. The confusion matrix demonstrating four classes of results which is later used for calculating the final accuracy.

Table 1. DNN predictions with the scores of accuracies for dense neural network.

Class	Precision	Recall	F1-score	Support
Non-epileptic	1	0.99	1	1193
Epileptic	0.98	0.99	0.99	307
Accuracy			0.99	1500
Macro avg	0.99	0.99	0.99	1500
Weighted Avg	0.00	0.99	0.99	1500

Table 2. DNN predictions with the scores of accuracies for convolutional neural network.

Class	Precision	Recall	F1-score	Support
Non-epileptic	0.89	0.96	0.92	1193
Epileptic	0.78	0.52	0.63	307
Accuracy			0.87	1500
Macro avg	0.83	0.74	0.77	1500
Weighted Avg	0.86	0.87	0.86	1500

Analytical signal and correlation with the heatmap

To evaluate the components of the signal that contribute to the DNN decision on whether an individual sample belongs to epileptic or non-epileptic category, heatmap was generated for 1000 randomly chosen signals from the pooled shuffled data from 10 patients. Also, the analytical signal for the same signals were calculated using Hilbert transform. The three signals were plotted for visual evaluation as noted in the Figure 16. The cross correlation was calculated between each heatmap-signal pair and heatmap- analytical signal pair. The maximal values of cross correlations are plotted in Figure 17.

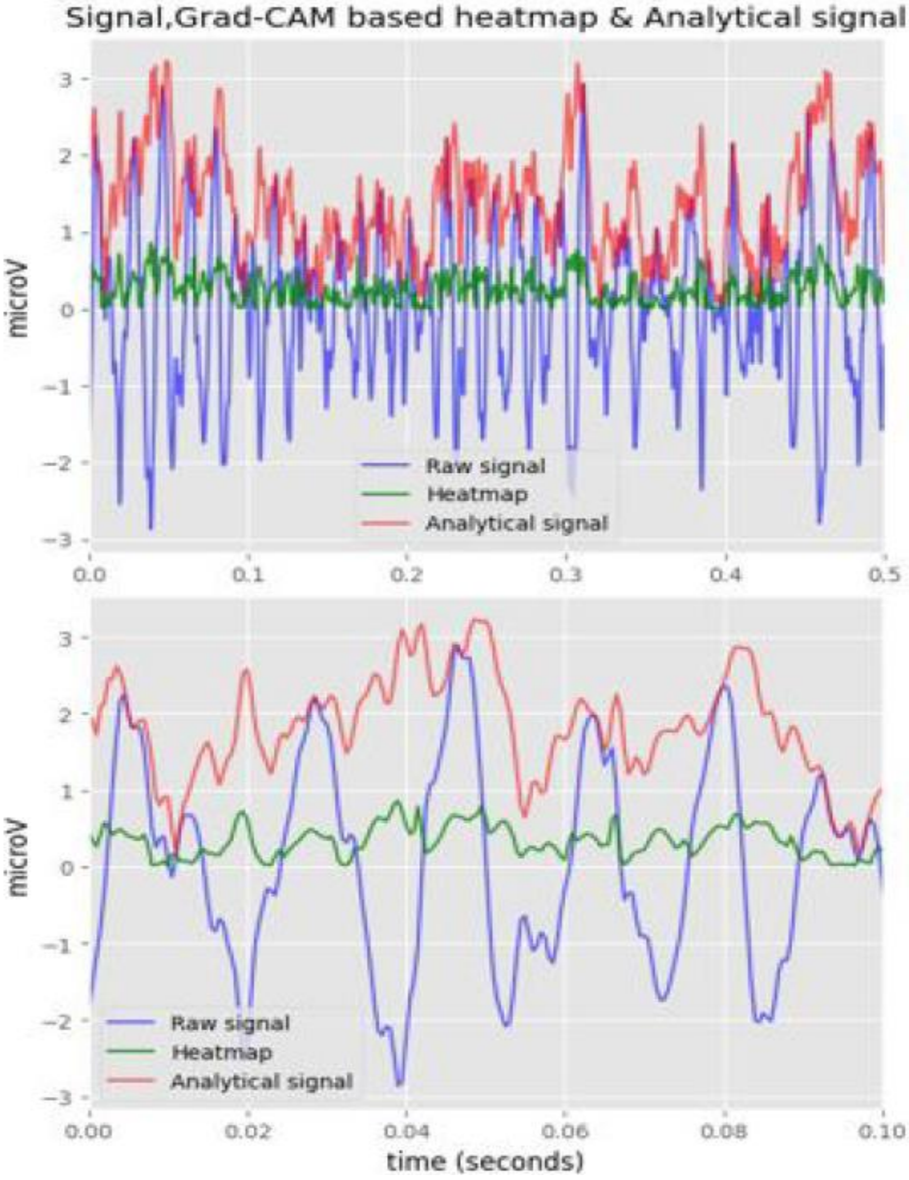


Figure 16. The cross correlation between raw signal and analytical signal.
The upper panel shows the original signal (blue), heatmap(green) and analytical signal (red).
The lower panel shows the same signals on a smaller time scale, showing the intermittent synchronization.

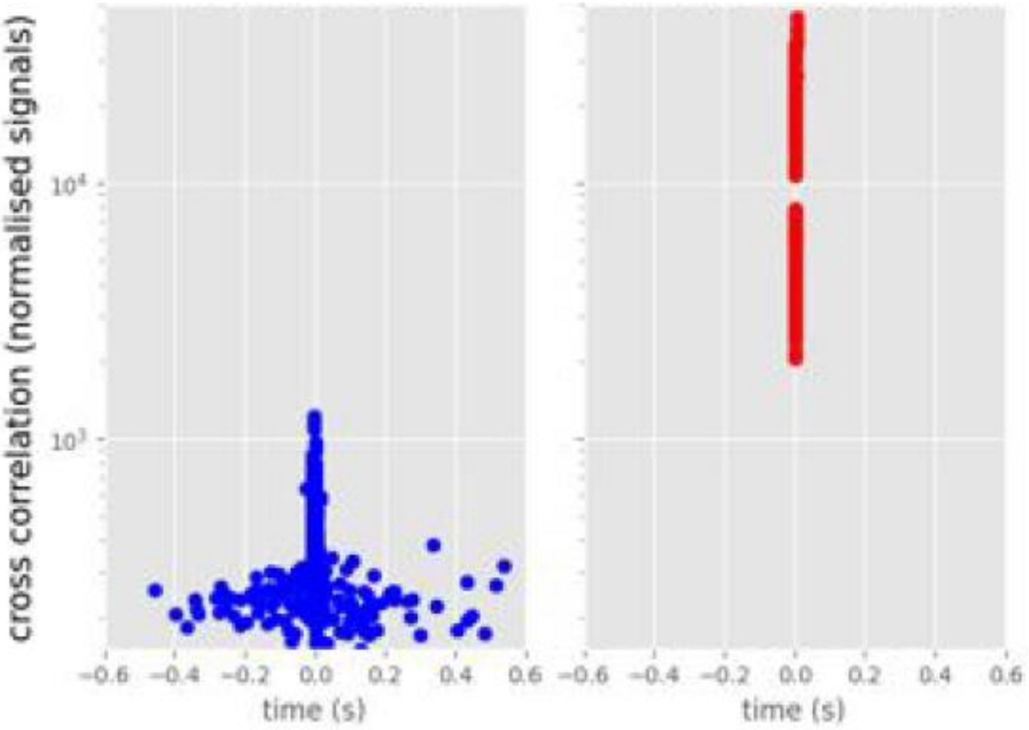


Figure 17. Correlation values of heatmap with original signal and analytical signal.

Cross correlation between analytical signal and heatmap shows significant similarity between these two signals shown in Figure 16. The absolute maximal/ minimal values of correlation were plotted against the corresponding lags, which shows significant higher correlation between the heatmap and analytical signal (Figure 17). For comparison, between two cross correlation values the average of the absolute value of maximal/ minimal correlation was estimated. This mean value for cross correlation between heatmap and original signal was (for absolute values) 365.27±144 (SD), with a mean lag of 2.6 ±78 ms and the same in the case between heatmap and Hilbert transform was 8184±6175 with a mean lag of 0.7±0.5 ms.

The distribution shown in indicates that temporal correlation between the heatmap and analytical signal was narrower compared to that with original signal. The heatmap was generated using the Grad-CAM algorithm for 1000 samples of data, for CNN which showed regions of higher and lower relevance in decision making by the model, shown in figure 5. The analytical signal was generated using Hilbert transform for the corresponding signals.

A K-fold cross validation [4] was applied on the data set with K=10, which demonstrated consistent accuracy of prediction for epileptogenic zones at 91% with a standard deviation of 1.3 % for dense network and 91.1 % with standard deviation of 0.8% for CNN.

The cross validation showed relatively lower accuracy for dense network, along with a higher standard deviation. This may be related to the fact that dense network utilizes less reliable features compared to CNN which may be employing details in local features for classifying the signals. As an example, the application of DNN for identification of the epileptogenic zones was demonstrated on one of the patient’s data is shown in figure 6. The predicted electrode zone was compared and noted to be overlapping with the epileptogenic zone in as identified in the clinical decision-making process. This patient has undergone surgery in the same region with control of seizures reported.

For identification of the abnormal epileptogenic zone in a new patient data, 1 minute data parsed into 1 second epoch was passed through DNN> This will yield n= 60* Number of electrodes predictions. A total of 15 patient’s data (10 patient’s data which was initially used for model development and additional 5 patient’s data used for this evaluation) was used for this evaluation. The electrodes with maximal number of predictions as epileptogenic zone were placed as the electrode with maximal likelihood of epileptogenicity. All the other electrodes were sorted based on the number of times they were predicted to be an epileptogenic focus. This method gave an additional layer of statistics to improve the reliability of prediction in a new patient data. Comparison of the electrode locations that was deemed to be epileptogenic through clinical evaluation and DNN predictions is given in Table 3 & Table 4.

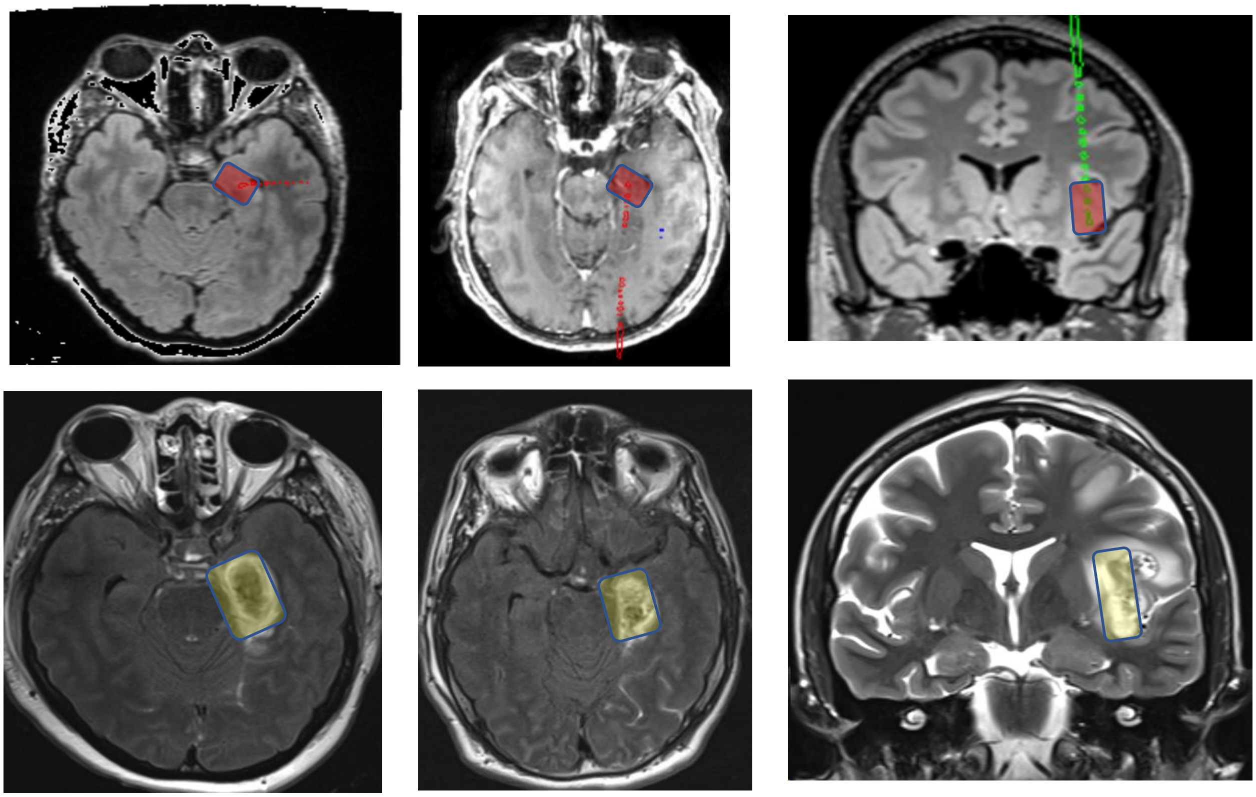


Figure 18. Comparison of the DNN model prediction with clinical data.
The upper panel shows where the DNN predicted the epileptogenic zones (shown in the red boxes). The lower panel shows the regions of the brain lesioned as part of treatment based on clinical decision making. These patients had good clinical outcome, indicating that the lesioned regions were actually epileptogenic and was the cause of the seizures in these patients.

Table 3 Electrodes locations thought to be involved in the epileptogenicity in clinical analysis. This is based on clinical evaluation, EEG, iEEG (including capturing typical seizures), MEG, PET and MRII. The first row indicates patient numbers.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LAT1	LH1	RH1	RH1	LAOI3	RMTO4	LH3	LH1	LH4	LH1	LH3	LC9	LPOI4	RH1	RPOI1
LAT2	LH2	RH2	RH2	LAOI4	RMTO5	LH4	LH2	LH5	LH2	LH4	LC10	LPOI5	RH2	RPOI2
LAT3		RH3	RH3	LAOI5	RMTO6	LH5	LH3	LH6	LH3	LH5	LC11	LPOI6	RH3	RPOI3
		RH4	RH4	LAOI6	RPTO1	RH1	LH4	LH7	LH4	LH6	LC12		RH4	RPOI4
		RH5		LAOI7	RPTO2	RH2	LH5	LH8	LH5		LC13		RH5	RMOI1
		RH6			RPTO3	RH3	LH6	LAST1	LH6		LC14		RH6	RMOI2
		RH7			RPTO4	RH4	LH7	LAST2	LMST1		LC15		RH7	RMOI3
					RPTO5	RH5	RH1	LAST3	LMST2		LPC3		RH8	RMOI4
					RPTO6	RH6	RH2				LPC4			RST7
					RPTO7	RH7	RH3				LPC5			RST8
					RPTO8	RH8	RH4				LPC6			RST9
					RPTO9		RH5				LPC7			RST10
					RPTO10		RH6				LPC8			RST11
							RH7							RST12
							RH8							

Table 4. DNN based prediction (top 10 predictions) of the electrode locations.
This is based on 1 minute of interictal data (from the first 24 hours of electrode implantation). First row corresponds to patient numbers.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LAT1	LH 1	RH7	RH2	LH1	RMTO4	RH1	RH4	LPST2	LH2	LH4	LH2	LPOI6	LH1	RST6
LAT2	LH 2	RH2	RH3	LH7	RMTO6	LH5	RH6	LPST3	LH3	LH6	LH3	LPOI7	LH2	RAOI14
LAOI4	LC2	RH3	RH4	LAOI1	RMTO3	RH2	LH6	LAST5	LH1	LH5	RH8	LTIP2	RH6	RPOI14
LPOI2	LC3	RH4	RH1	LAOI2	RPTO9	LH4	RH3	LPT1	LMST2	LH7	LH1	LPOI3	RH5	RMOI10
LOP2	LC4	RH5	RC11	LAOI3	RPTO10	RH4	RH5	LPT9	LF7	LPST3	LH4	LPOI4	LH3	RMOI13
LOP3	LC5	RH1	LC14	LAOI12	RPTO7	RH5	RH7	LT1	LF20	LH8	LH13	LPOI5	RH7	RIFI9
LOP4	LC7	RH6	LC12	LPOI1	RPTO8	RH6	RH8	LPT2	LPT8	LPSt4	LC1	LTIP1	LAOI2	RPOI10
	LAOI1	RH8	RC10	LPOI4	RMTO5	RH7	LH3	LPT13	LMST1	LH9	LLT1		LAOI3	RMOI9
	LAOI2	RPOI5	LC10	LPOI8	RPTO3	LH3	LH4	LPT15	LH5	LAST1	LHWM3		LAOI4	RST5
	LAOI3	LPOI2	LAOI7	LPC10	RPTO5	LH6	RH1	LPST5	LAST4	LASt2	LHWM4		LPOI5	RPTO1

Discussion

The potential to predict epileptogenic zones with limited duration of iEEG signal has to be compared to the current clinical approach based on multiple days of iEEG recording and capturing seizure events. The current approach poses significant risks to the patients and medical system. Also, an approach as described in this study will have significant impact in reducing the hospital admission duration and will thereby reduce the medical costs. While the model offers a promising approach, one of the concerns of using the DNN models in medical field is the limitations in understanding the decision-making process of DNN. The conventional medical decision-making process already involves putting together several pieces of data. Adding a black box-based system to this process can be difficult from the standpoint of clinicians as well as patients.

It is in this context that the heatmapping approaches have to be considered especially to explain the decision making process of DNN. The use of heatmapping technique in this study with the Grad-CAM algorithm has demonstrated that the decision making of DNN can be unraveled to a significant extent, which could help the clinical teams in appropriately incorporating these kinds of models in clinical practice. A similar approach with layerwise relevance propagation was described for classifying neurocognitive tasks [60] demonstrating the potential of this approach in understanding the neurophysiological patterns. Our study shows that the heatmapping points to areas of increased instantaneous power (as described by the analytical signal) as contributory to the classification of an epileptogenic or non-epileptogenic region. A similar evaluation to identify epileptogenic regions in iEEG has been reported in previous studies with CNN [61] on short term Fourier transform of signal. This approach reported an accuracy of 91.8% in differentiating focal and nonfocal iEEG signals. That approach did not explore the features that the network was learning. Another deep learning approach with intracranial EEG data, for identifying interictal epileptic discharges (IED) has been reported[62]. This study employed CNN and reported a 70-90% classification accuracy in detecting IEDs. Compared to that study we did not specifically target identifying any specific pattern in the EEG.

Our strategy to identify the underlying features that the model is learning goes with the conventional medical approaches which tries to identify the transients in iEEG which is subsequently correlated with the regions of epileptogenicity. While being congruent with the conventional methods, this approach also helps in further unravelling the underlying epileptogenic pathophysiological processes that may be giving rise to certain forms of signals in the iEEG. A similar method can potentially be applied to classifying other brain states (sleep, drowsiness. alertness etc.) and in further understanding the neurobiological underpinnings of it. In the context of long term EEG data, several studies[22], [63] have been reported using CNN approach for detection of seizures. They reported high accuracy (90-99 %) and sensitivity and specificity between 85-95%. But it may be noted that these studies were looking to identify seizures in the long-term EEG data (and seizures are recorded usually after several days of EEG monitoring). Compared to those studies, our approach uses 60 seconds of recorded interictal data, when no seizures are recorded. The fact that we have focused on higher frequency components of the iEEG signal would have partially contributed to the ability of the DNN to identify the epileptogenic zones from a limited duration of recording. From clinical studies it is known that the fast activity patterns represented in the higher frequency components are more correlated with epileptogenic zones. Another reason for the enhanced ability could be the use of specific filters used in the CNN, which matches with the higher frequency components in the iEEG data.

*Table 5***.** *Existing DNN models on iEEG, exploring epileptogenic zones/ epileptic activity compared with the current model.*

Study	**Deep Learning Strategy**	**Input Formulation**	**Frequency Range (FR)/Sampling Rate(SR)**	**Task**	**Accuracy**
[61]	CNN(Convolutional neural network) with STFT(short term Fourier transform)	Data from 5 patients. 20 s of data	SR: 512 Hz	Differentiate focal and non-focal epileptogenic signal	91.8%
[22]	1D-CNN with data augmentation strategies	24 patients, 916 h data; & 18 patients,2565 h data.	SR:256 Hz	Seizure detection	99%
[62]	CNN	Data from 12 patients.	NA (not available)	Interictal epileptic discharge detection	79–87%
[25]	CNN + LSTM (long short-term memory attention machine)	Three data samples	SR:1–512 Hz, 1–173 Hz or 2048 Hz	Epileptogenic vs. non-epileptogenic	97.6%
[21]	1-CNN, 2-CNN, 3-CNN, 4-CNN	2016 Kaggle competition; Data from 5 dogs and 2 patients	SR: 400 Hz	Seizure classification	76–95%
[23]	CNN	2016 Kaggle competition; Data from 5 dogs and 2 patients	SR: 400 Hz	Seizure prediction	87.85% sensitivity in seizure prediction
[26]	CNN	Responsive neural stimulator data from 22 patients	SR: 250 Hz FR: 4–125 Hz	Seizure identification	84%
Curr-ent study	CNN	Data from 17 patients; 1 min data	SR: 2000 Hz; FR:60–600 Hz	Epileptogenic vs. non-epileptogenic	91–95%

Conclusions

The study demonstrated a clinically significant potential use of DNN for evaluation of iEEG data for identifying epileptogenic zones. Further, the underlying signal features that significantly contributed to the decision making of DNN was unraveled. Apart from medical application this strategy highlights the advantage of using heatmapping approaches in unravelling underlying neuroscientific details.

CHAPTER 4: ESTIMATION OF CERTAINTY OF DNN PREDICTIONS AND CORRELATION WITH INFORMATION CONTENT

[This work was published in the article “Information Entropy Measures for Evaluation of Reliability of Deep Neural Network Results”, DOI: 10.3390/e25040573, Entropy. 2023 Mar 27;25(4):573. [64]]

As noted in the previous chapter, we explored the possibility of using the DNN models to predict epileptogenic regions in the brain. When these approaches are used in real world, one of the challenges is in corelating with clinical context. Accuracy of the overall DNN gives an idea about the trustworthiness of the overall network. But it does not give an estimation of the confidence the network has in individual predictions. This can be a limitation in individual decision making, especially for example in the medical field where each decision on individual patient is a confluence of information from various body systems, medical factors, psychological and social factors. Therefore, it may be helpful to have a measure of confidence in the individual decisions of DNN.

In general, DNN come up with a score for the individual prediction which is noted as y_i ($y_i$being the i-th entry in a particular layer), in the following equation based on the inputs. With x_i (x_1 $x_2, \ldots$ x_n) as inputs, w_{ij} ($w_{i1,}$ $w_{i2,..}$ w_{in}) as weights and w_o as the bias term[65].

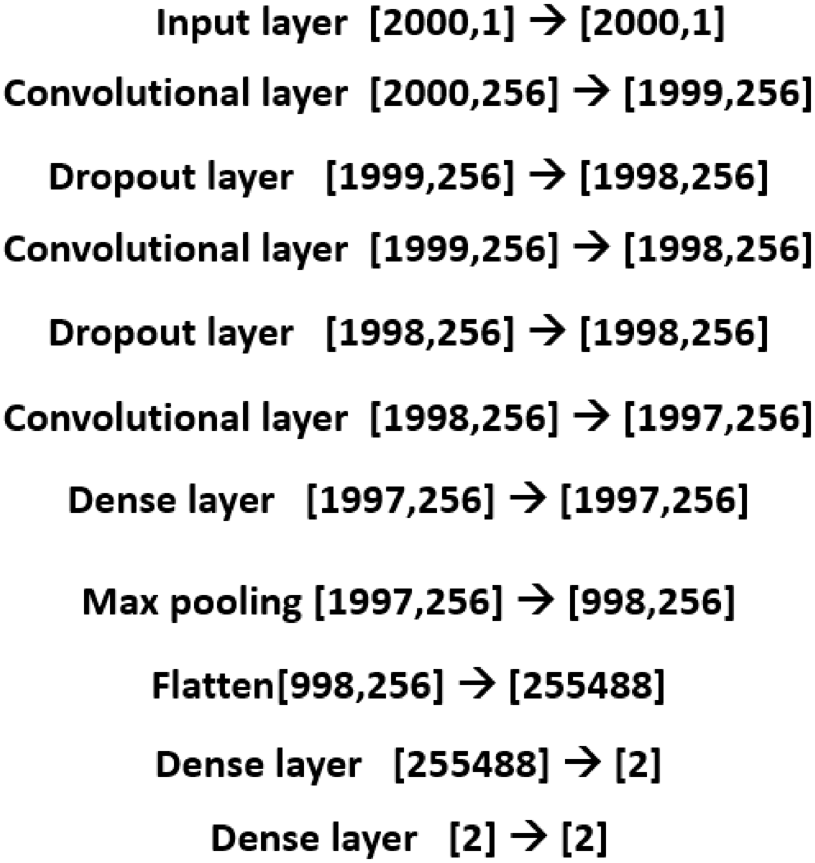
$$y_i = \sum_{j=1}^{n} x_j w_{ij} + w_o; \;\; z\,(i) = g(y_i) \tag{8}$$

The activation function is represented by *g(.)* and in the last layers DNN commonly uses a softmax function as activation to normalize the probability distribution with the following function.

$$g(y_i) \; = \frac{exp(y_i)}{\sum_{i=I}^{k} exp\;(y_i)} \tag{9}$$

We introduce the method of assessment of certainty in the neural network predictions. This is a measure of how certain DNN is about each of the individual predictions. This method was motivated by the following two ideas: 1. In the case of biological neural networks the decision of a subsequent neuron firing is partly dependent on the summation of the post-synaptic action potentials (both excitatory and inhibitory), which is like the inputs to DNN's last layer. Broadly, we can consider the "decision" of the neuron to fire and transmit the information to the next layer as a surrogate of certainty. Therefore, intuitively we can consider that the biological neurons are considering the positive and negative inputs in arriving at the decision and possibly at assessing the certainty of predictions. 2. The Grad-CAM algorithm used in the heatmap generation (that we used in this study) is based on the gradients of scores for individual classes (before the softmax layer), explained in detail in the methods section. Therefore, using a certainty measure based on the same score was considered appropriate, when evaluating the relationship between heatmap and information measures like sample entropy used in this study. This is different from accuracy which is a measure of overall correctness of predictions.

For the implementation of these methods, we used the convolutional neural networks which is a form of DNN. Apart from basic deep neural network methods, signals with time varying features can be analyzed and classified using convolutional neural network (CNN) which are better capable of capturing features using a sliding window along the signals [27]. This allows for identifying the local features of signal which are often used in identification of epileptogenic zones by clinicians (sharp waves, spikes, high frequency oscillations etc.) and are particularly suited for evaluating pathological features in iEEG signals which may contribute to the decision making of DNN. These pathological features include high frequency activity and ripples. Given this potential, CNN was employed as a strategy for signal classification in this study and heat-mapping method was implemented on CNN model. The depiction of the input along with contribution of different regions of the signal in the decision-making process, is referred as heatmap and can be estimated for individual layers of CNN using various algorithms. Essentially heatmaps help in identifying the key features that was used by the model to arrive at a particular decision. Heatmaps can be generated using various methods which were alluded to in previous chapter and the methods of Grad-CAM analysis was used for the evaluation of CNN. The implementation of CNN used in this study is depicted below.



Input layer [2000,1] → [2000,1]

Convolutional layer [2000,256] → [1999,256]

Dropout layer [1999,256] → [1998,256]

Convolutional layer [1999,256] → [1998,256]

Dropout layer [1998,256] → [1998,256]

Convolutional layer [1998,256] → [1997,256]

Dense layer [1997,256] → [1997,256]

Max pooling [1997,256] → [998,256]

Flatten[998,256] → [255488]

Dense layer [255488] → [2]

Dense layer [2] → [2]

Figure 19. The Convolutional neural network implemented with higher number of convolutional layers.

The CNN implemented consisted of 11 layers with 3 convolutional layers with kernel size of 256. The input had 2000 nodes which corresponded to the number of elements in the input samples. Binary-cross entropy was used as the loss function and Relu was used as activation function. The model was trained with 8000 samples obtained from iEEG signals which is described in the previous chapter. The validation was performed with 2000 samples. The model was trained for 100 epochs.

The certainty in the outputs of any DNN nth layer is estimated as follows:

$$C_i^n = y_i^n - \frac{1}{N} \sum_{j=I;j\neq i}^{N-1} y_j^n \tag{10}$$

where certainty index, C_i^n is the certainty that the i-th prediction in the n-th layer is correct,

y_i^n is the i-th DNN output at n-th layer,

N is the total number of nodes

For comparison between layers and different DNNs, this measureccan be normalized to standard deviation. These certainty estimates were plotted in two groups, that favored the decision of epileptogenic electrode location and non-epileptogenic location.

Heatmap estimation

Class-discriminator map defined as Grad-CAM is evaluated for any class c, as follows[57].The gradient of the score for class c as given by yc , with respect to the feature map activation Ak of a convolutional layer is given by

$$\alpha_k^C = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial y^C}{\partial A_i^k} \tag{11}$$

y^C is the DNN output for a particular class *c (before softmax)*.

α_k^C indicates the importance weight of *k*-the filter for class *c*.

A_i^k is the *i*-the element in *k*-th activation map.

N is the number of elements in feature map.

With the gradient weighted class activation map (grad-CAM) for a layer obtained as

$$g^c = \Sigma_k \alpha_k^C A^k \tag{12}$$

These gradients can be global-average pooled which gives weights (α_k^C) of neurons based on their significance in decision making for specific input.

The ReLU function was not used, given the fact that we are working on one dimensional signal and similar approach was used in acoustic signal-based studies [19]. Also, incorporating negative values in heatmaps was considered to be important when calculating correlations with information measures of the signal. This approach helps in generating the heatmap for any layer. Since the concern that the Grad-CAM maps can progressively worsen in the earlier layers, in this study, the heatmap was only calculated based on the last convolutional layer of the model.

The details of the steps for Grad-CAM calculation are showed schematically with an example below. In this case activation maps from penultimate layer of CNN kernels A^1, A^2.. A^k is used and the Grad-CAM is calculated as detailed below.

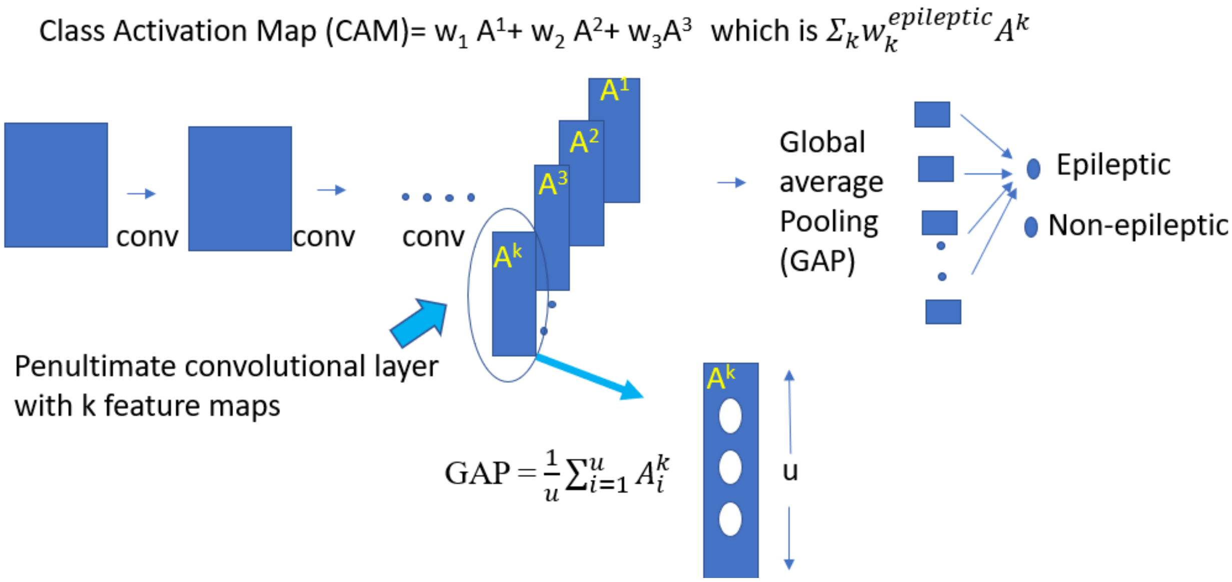

Class Activation Map (CAM)= $w_1 A^1 + w_2 A^2 + w_3 A^3$ which is $\Sigma_k w_k^{epileptic} A^k$
conv
conv
conv
A^1
A^2
A^3
A^k
Global average Pooling (GAP)
Epileptic
Non-epileptic
Penultimate convolutional layer with k feature maps
$\mathrm{GAP} = \frac{1}{u}\sum_{i=1}^{u} A_i^k$
A^k
u

Figure 20. schematic showing the Class activation map calculation.
For illustration a 1-dimensional convolutional network is shown.

Global average pooling (GAP) takes the average of all the elements.

the GAP will be calculated as

$$GAP = \frac{1}{u}\sum_{i=1}^{u} A_i^k \tag{13}$$

In this case the final score for a specific class (eg. epileptic) in final layer before softmax layer will be given by:

$$y^{epileptic} = \sum_{k=1}^{k} w_k^{epileptic} \frac{1}{\mathrm{u}} \sum_{\mathrm{i}=1}^{\mathrm{u}} \mathrm{A}_{\mathrm{i}}^{\mathrm{k}} \tag{14}$$

Calculation of the gradients for the category ‘epileptic’ is illustrated below:

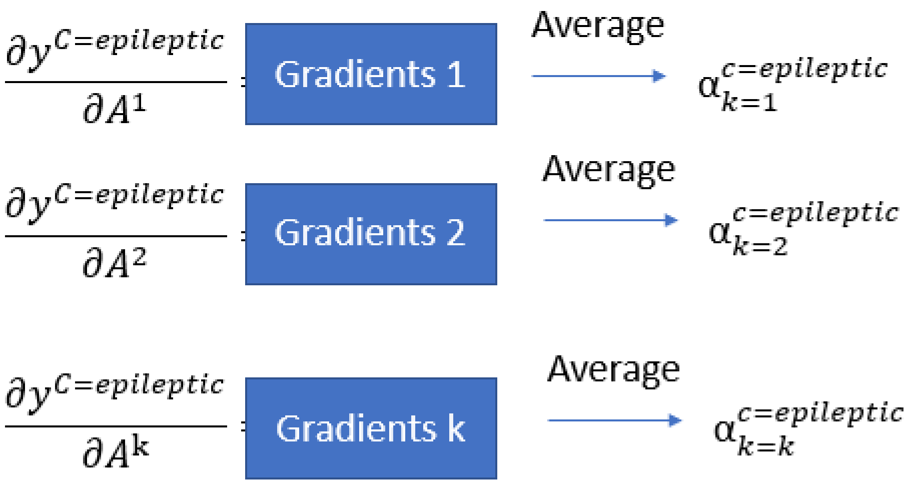

$\frac{\partial y^{C=epileptic}}{\partial A^1}$
Gradients 1
Average
$\alpha_{k=1}^{c=epileptic}$
$\frac{\partial y^{C=epileptic}}{\partial A^2}$
Gradients 2
Average
$\alpha_{k=2}^{c=epileptic}$
$\frac{\partial y^{C=epileptic}}{\partial A^k}$
Gradients k
Average
$\alpha_{k=k}^{c=epileptic}$

Figure 21. Illustration for calculation of gradients of individual sample outcomes with activation maps

The Grad-CAM for each category for that layer is calculated as a weighted combination of the feature maps as below:

$$Grad - CAM^{epileptic} = \alpha_1 A^1 + \alpha_2 A^2 + \cdots t + \alpha_k A^k \tag{15}$$

OR

$$g^c = \Sigma_k \alpha_k^C A^k \tag{16}$$

As an example, in the case of Modified National Institute of Standards and Technology database (MNIST) images. The heatmaps are calculated and plotted below the original images.

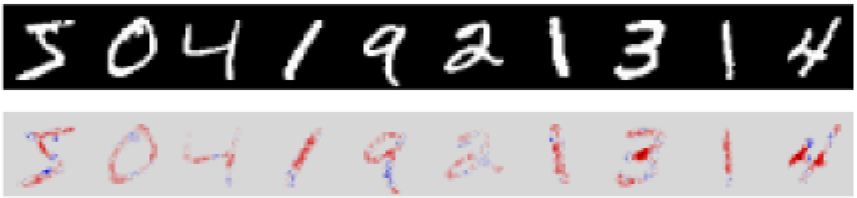


Figure 22. Illustration of the implementation of heatmapping in the case of MNIST images.

To estimate the similarity between the heatmap and the original data a cross correlation was performed as noted in the equation

$$z(k) = \sum_{l=0}^{\|g^C\|-1} g_l^C * x_{l-k+N-1}^{*} \tag{17}$$

where $\|g^C\|$ is the length of g^C,which is the heatmap for the signal X

N= max($\|g^C\|$,||X||) and x_m is 0 when m is outside range of y.

The cross correlation between heatmap and signal plotted for various samples is shown Figure 23

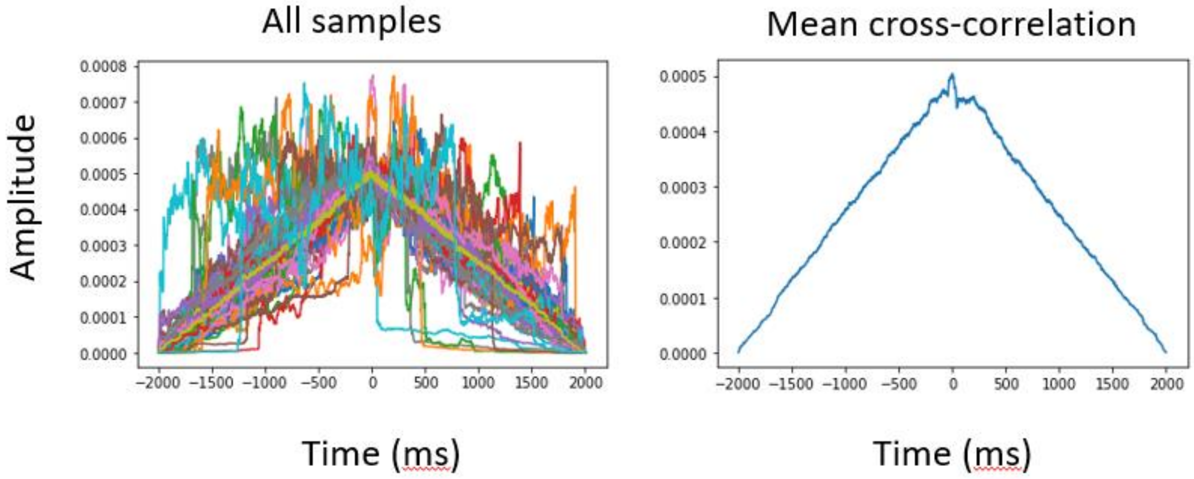


Figure 23. Correlation between heatmap and original signal.

The maximal values of this correlation were plotted against the certainty index of that data.

Sample Entropy calculation

One of the challenges of evaluating the accuracy of the DNN model is, whether the model has adequately captured the information contained in the data. To address this, we used the method of information theory-based analysis. For this purpose, sample entropy (SampEn) was estimated which is especially suited for signals where semirhythmic patterns are present. Based on Shannon's information theory[41] the information in a collection of data X can be defined as

$$H(X) = -\sum_{x \in x} p(x)\, log\, p(x) \tag{18}$$

where *X* is taking values $x_1, ..., x_n$ and p(x) is the probability associated with those values for all $x_1, ..., x_n$. But this approach does not consider the repetitive nature of signals and the information in those types of patterns. Therefore, sample entropy was used as a method to estimate the information content in the signal as noted below.

Generalized version of Shannon entropy is Renyi entropy which is described as follows:

Renyi entropy of order α is defined as

$$H_\alpha(x) = \frac{1}{1-\alpha} log\left(\sum_{i=1}^{n} P_i^\alpha\right) \tag{19}$$

with Shannon entropy being $H_{Shannon} = \lim_{\alpha\to 1} H_\alpha$

Using this approach a measure of information rate generated in a chaotic data series is described as [65]:

$$C_d(r) = \lim_{N\to\infty} \frac{1}{N^2}\left[numberofpairsof(n,m)with\left(\sum_{1=1}^{d}\left|x_{n+j}-x_{m+i}\right|^2\right)^{\frac{1}{2}} \leq r\right] \tag{20}$$

It measures with a tolerance of *r* the regularity of patterns similar to a given template of a particular length.

Using this approach, we can approximate the entropy of a time series as:

$$E_d = \frac{1}{\tau} log \frac{C_d(r)}{C_{d+1}(r)} \tag{21}$$

The τ represents the time intervals at which the system is measured.

The correlation integral allows for reconstructing the evolution of all degrees of freedom using *d* measures of a single co-ordinate.

$$C(r) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i,j}^{N} \theta\left(r - \left|\overline{\overline{\bar{X}}}_{\imath} - \bar{X}_{\jmath}\right|\right) \tag{22}$$

Where *θ(x)* is the Heaviside function.

$$H(x) \coloneqq \begin{cases} 1, x > 0 \\ 0, x \leq 0 \end{cases} \tag{23}$$

For calculation Sample Entropy of this approach was approximated and implemented as follows[43]. Given a sequence of numbers $x_1, x_2, ..., x_n$ of length N, a non-negative integer m ≤ N and a positive integer r, block *u(i)* can be defined as $x(i), x(i + 1)...,x(i + m - 1)$ and block *u(j) as* $x(j), x(j + 1)..., x(j + m - 1)$. The distance between them is defined as $d[u(i), u(j)] = max_{k=1, 2...m}(|x(i + k - 1) - x(j + k - 1)|)$.

The sample entropy, which helps in better capturing the recurring nature of data elements in a signal, is defined [44] as:

$$SampEn(m, r, N) \tag{24}$$
$$= -\,log \frac{\sum_{i=1}^{N-m} \sum_{i=1, j\neq i}^{N-m}}{\sum_{i=1}^{N-m} \sum_{i=1, j\neq i}^{N-m}} \frac{[number\ of\ times\ that\ d[|u_{m+1}(j) - u_{m+1}(i)|] < r]}{[number\ of\ times\ that\ d[|u_m(j) - u_m(i)|] < r\ ]}$$

This was performed for various sampling intervals. A cross correlation is calculated between the heatmap generated (as described previously and the sample entropy calculated). The maximal cross correlation value was plotted against the certainty values.

Illustration of sample entropy calculation

The steps in calculation of sample entropy are illustrated below. A segment from a rolling time window is chosen for the estimation of sample entropy.

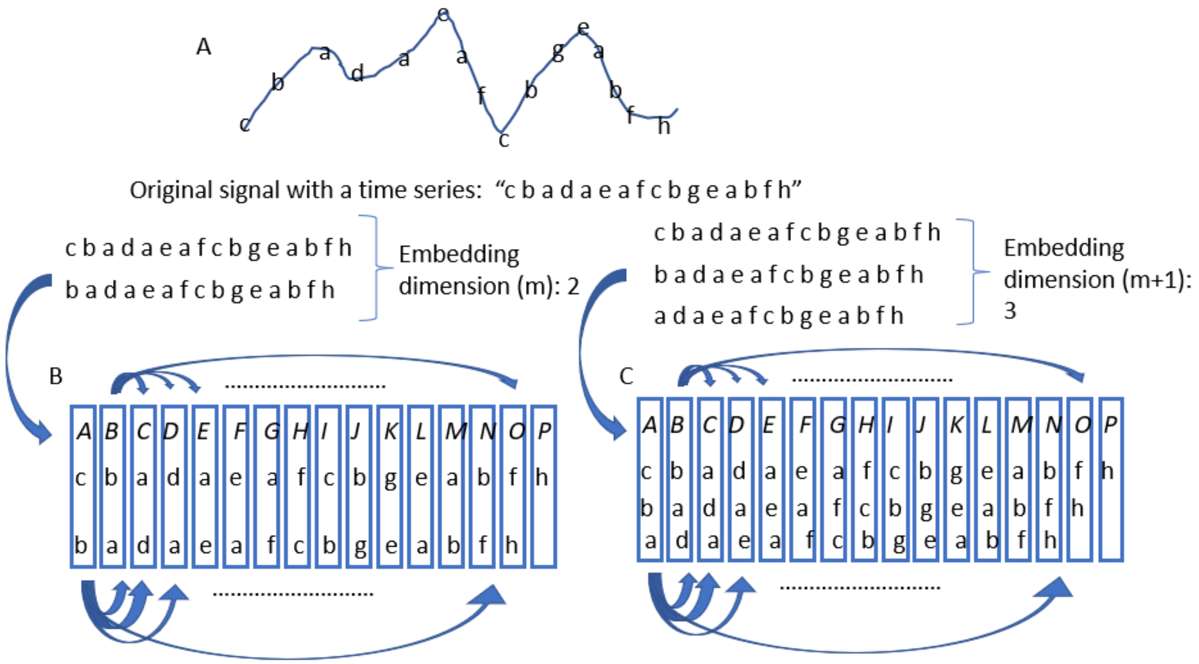
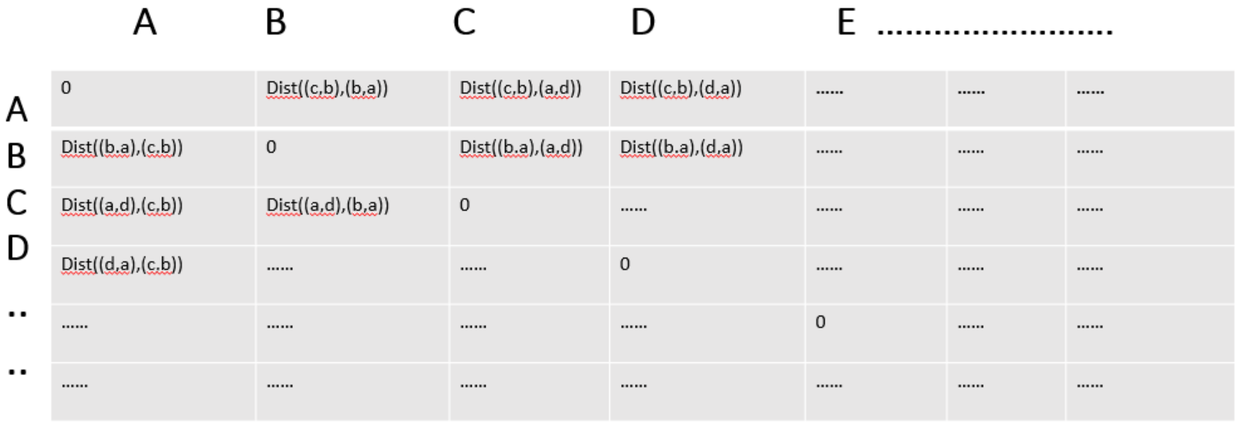

A
c b a d a e a f c b g e a b f h
Original signal with a time series: "c b a d a e a f c b g e a b f h"
c b a d a e a f c b g e a b f h
b a d a e a f c b g e a b f h
Embedding dimension (m): 2
B
A B C D E F G H I J K L M N O P
c b a d a e a f c b g e a b f h
b a d a e a f c b g e a b f h
....
c b a d a e a f c b g e a b f h
b a d a e a f c b g e a b f h
a d a e a f c b g e a b f h
Embedding dimension (m+1): 3
C
A B C D E F G H I J K L M N O P
c b a d a e a f c b g e a b f h
b a d a e a f c b g e a b f h
a d a e a f c b g e a b f h
....

Figure 24. Illustration of sample entropy calculation.
The values of time series are given by: "c b a d a e a f c b g e a b f h". The pairs (in B) and triplets (in C) formed can be represented as A, B, C…. and the distances between them is represented as the arrow between the pairs of the groupings shown.

The distance matrix in this case between all possible pairs will be represented as matrix as shown in the table below.

Table 6. Table illustrating the distance between the pairs formed in figure 24 B.



	A	B	C	D	E		
A	0	Dist((c,b),(b,a))	Dist((c,b),(a,d))	Dist((c,b),(d,a))	......	......	......
B	Dist((b,a),(c,b))	0	Dist((b,a),(a,d))	Dist((b,a),(d,a))	......	......	......
C	Dist((a,d),(c,b))	Dist((a,d),(b,a))	0	......	......	......	......
D	Dist((d,a),(c,b))	......	......	0	......	......	......
..	......	......	......	......	0	......	......
..	......	......	......	......	......	......	......

In the Table 6 the “Dist”, indicates the distance between the pairs of rows and column. The distance function can be Chebyshave distance or Eucledean distance. For examples the distance between A and B Dist((c,b),(b,a)), with Chebyshave distance being

$$D_{Chebyshave} = max(|b - c|_{,} \ |a - b|) \tag{25}$$

Similarly, the distance can be calculated in the case of triplets in as shown in Figure 24B (eg. distance between A and B: Dist((c,b,a),(b,a,d))) and a table of distance matrix similar to Table 6.

Sample entropy is calculated based on the above values in tables for m and m+1 dimensions and SampEn is calculated as follows:

$$\text{DistSum}_{m}\text{= sum(distance with embedding dimension(m) <= r * } \sigma) \tag{26}$$

$$\text{DistSum}_{m+1} \text{ = sum(distance with embedding dimension(m+1) <= r * } \sigma) \tag{27}$$

Where σ: standard deviation of signal.

Sample entropy is calculated as:

$$\mathrm{SampEn} = -\log \frac{\mathrm{DistSum_{m+1}}}{\mathrm{DistSum_m}} \tag{28}$$

The algorithm for calculation of Sample entropy used in this study is noted below. The sample entropy is calculated for successive rolling windows of length *N*, at embedding dimension of *m* and scaling parameter *r*. For evaluation of the robustness of the results the calculation is repeated with changes in one of the parameters (*m,r or N*), while keeping the other parameters same.

Algorithm 1 Sample Entropy for a time series

Sample entropy of signal *s* of length SN for embedding dimension *m*, scaling parameters r and sample entropy calculation length *N*

Input: s1, s2,…..s_{SN}

Output: SE_1,SE_2… SE_N (series of sample entropy)

```
SE ← [$0_1$, ------$0_{SN}$]
N ← length(s)
m ← embedding dimension
r ← scaling parameter
for <si in range of SN> do $sig_{si}$ ←<split s into SN segments of length N> end for
    for <$i$ in range of $N$-$m$> do $xmi$←<split $sig_i$ into segments of length m> end for
    for <i in range of $N$-$m$+$1$> do xmj← <split $sig_i$ into segments of length m> end for
    B←<total of the modulus (xmi-xmj) <r> [xmi-xmj indicates the distances between the segments]
    m←m+1
    for <i in range of N-m+1> do xmk ←<split $sig_i$ into segments of length m> end for
    A← <total of the modulus (xmk-xmk)> [xmk-xmk indicates the distances between segments]
$SE_{si}$ ← -log (A/B)
```

Results

As establishing the confidence of DNN for each prediction was important from a clinical standpoint we explored measures of certainty that the DNN has in its predictions. Further we explored the DNN's ability to capture the information content in the signals using sample entropy algorithm. The certainty of DNN predictions were explored further with its correlation with the ability of DNN to capture the information content in the signal.

A convolutional network model was implemented as described in Figure 19. The model consisted of three convolutional layers and additional dense and dropout layers. A dense layer was added before the softmax layer to get the outputs prior to the softmax function. The iEEG data lasting one second from each channel were fed into the input layer. The model was trained with 9000 samples and tested with 1000 samples. The accuracy of the model was noted to be 93%. The confusion matrix which depicts the percentage of positive and negative predictions is shown in the Figure 25.

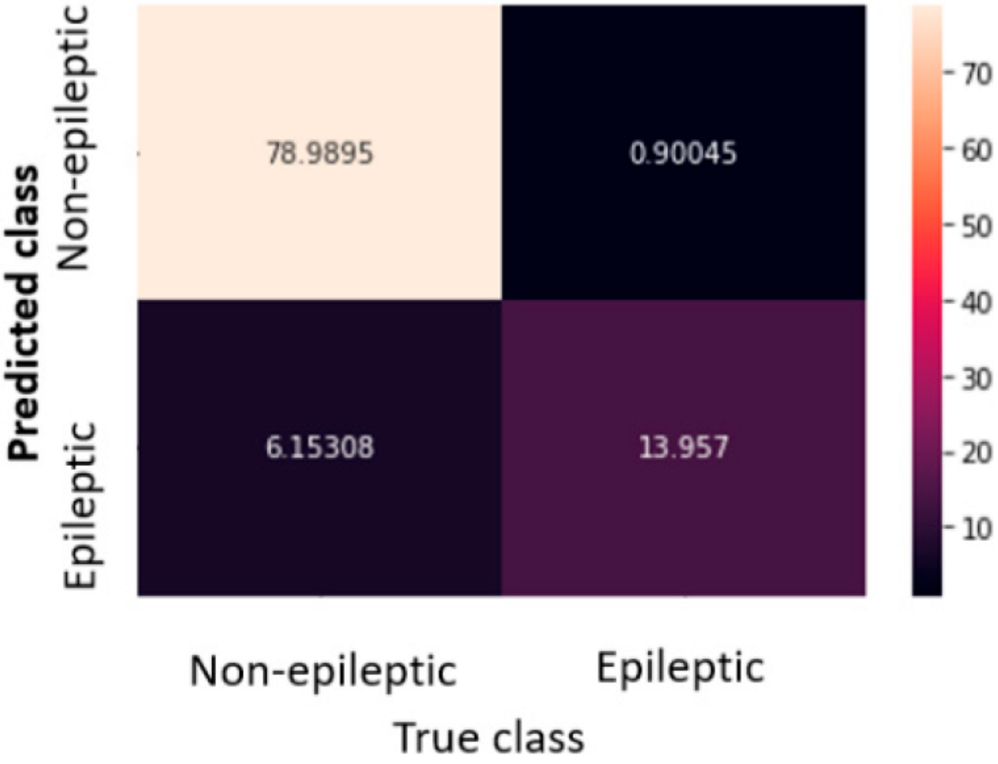


Figure 25. Confusion matrix showing the results of the test samples as percentage.

<u>Certainty in individual predictions</u>

The confidence in the prediction of each data was estimated as noted by the measure of certainty as described in the methods section. Certainty is estimated in the test data. The absolute value of certainty, for signals from epileptic electrodes ranged between 0-200 and that for the epileptic signals ranged from 0 - 60 (the more negative the value, the higher the certainty of the models that the data is from epileptic regions) Figure 26. For better comparison the data was also plotted after normalizing with the standard deviation.

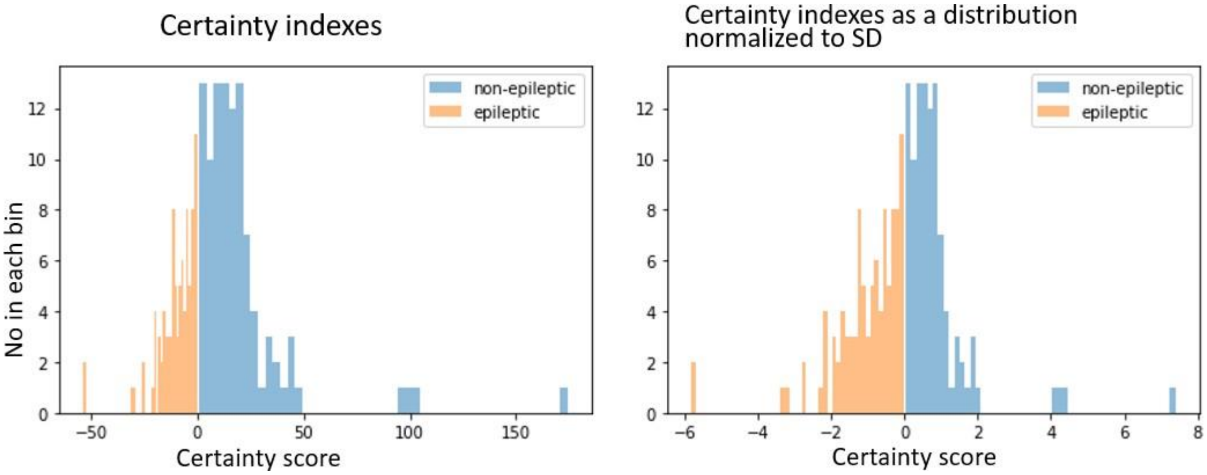


Figure 26. The certainty of the decisions on the classification of data to that arising from epileptic vs non-epileptic locations.
Please note that the negative and positive values are chosen arbitrarily for classification purpose to epileptic and non-epileptic classes and the symbols do not have other relevance.

<u>Certainty and correlation between heatmap and signal</u>

To estimate how the decision-making process of the DNN model is related to the certainty index, we calculated the correlation between the heatmap and the original signal. The maximal correlation values were plotted against the certainty index for that signal (Figure 27Figure 27. The certainty index and its relationship with the correlation between heatmap and original signal.) and the correlation coefficient was estimated as 393.9 +/- 6.7(SE). A similar range of correlation values were noted when the epileptic and non-epileptic data was evaluated separately.

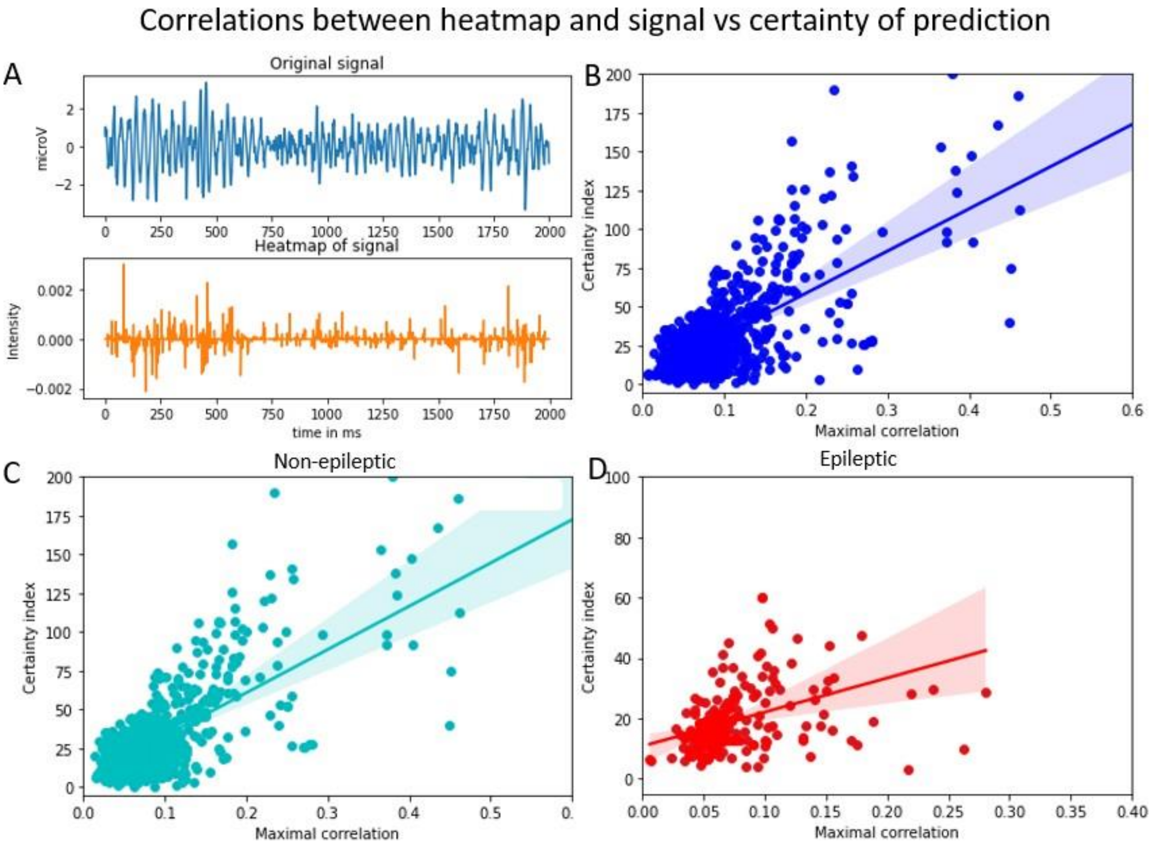


Figure 27. The certainty index and its relationship with the correlation between heatmap and original signal.

A. Original signal (upper) and the corresponding heatmap (lower) B. Scatter plot: Certainty index vs correlation between heatmap and original signal C. Cross-correlation between the heatmap and original signal plotted for non-epileptic data. D. Cross-correlation between the heatmap and original signal plotted for epileptic data.

Table 7. Regression values: certainty index vs. correlation between heatmap and original signal.

Statistical Values	Whole Data	Non-Epileptic	Epileptic
R-squared	0.76	0.77	0.69
F-statistic	3190	2714	450

Relationship between heatmap and sample entropy

The sample entropy for individual signals was calculated based on equation (*24*). For this purpose, following parameters were used: embedding dimension, m=8, scaling parameter, r= 2 x standard deviation of the signal, signal length, N= 100. The fact that the EEG signal has frequency components which range from 60-600 Hz was considered in choosing the embedding dimensions and signal length. Further analysis based on variations in these parameters are noted in sections below. A cross -correlation was calculated between the sample entropy and the heatmap and the maximal value of this cross correlation was plotted against the certainty index for individual data as shown in Figure 28. The R-squared values for the regression analyses are given in the Table 8. It may be noted that the R-squared values in the case of the cross correlations between the sample entropy and heatmap appear to be higher, compared to that between original signal and heatmap.

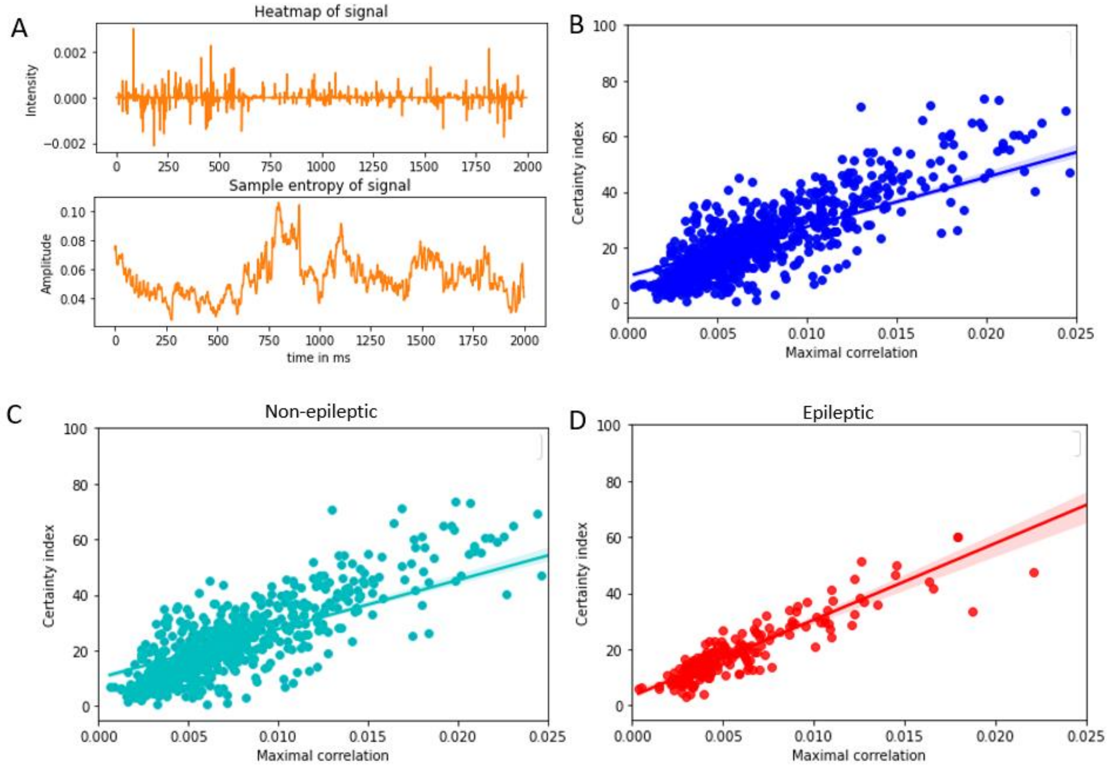


Figure 28. The certainty index and its relationship with the correlation between heatmap and sample entropy of the signal.

A. Heatmap of the signal (upper) and the sample entropy (lower) B. Scatter plot: Certainty index vs correlation between heatmap and sample entropy of signal C. Cross-correlation between the heatmap and sample entropy of signal for non-epileptic data. D. Cross-correlation between the heatmap and sample entropy of signal plotted for epileptic data.

Table 8. Regression values: certainty index vs. correlation between heatmap and sample entropy.

Statistical Values	Whole Data	Non-Epileptic	Epileptic
R-squared	0.89	0.90	0.95
F-statistic	8303	6850	4197

Relationship between heatmap and sample entropy at various embedding dimensions

The equation (*24*) shows that the sample entropy depends on the embedding dimension *m*, scaling parameter *r*, and signal length *N*. A too high value of *m* can potentially reduce the template matches performed in the algorithm. On the other hand, if *m* selected is too small there will be more template matches but the predictive information will be reduced, and the probability of forward match can be underestimated. This is especially true in the case of EEG which may have repeating patterns. To evaluate the impact of these parameters on the relationship between sample entropy, heatmap and certainty index, those parameters were varied, and the relationship was estimated. The sample entropy, m, was calculated at embedding dimensions 4,8,16 and 32 (while keeping r = 2, N =100). The certainty index vs maximal correlation between the sample entropy and heatmap at various embedding dimensions is plotted in Figure 29

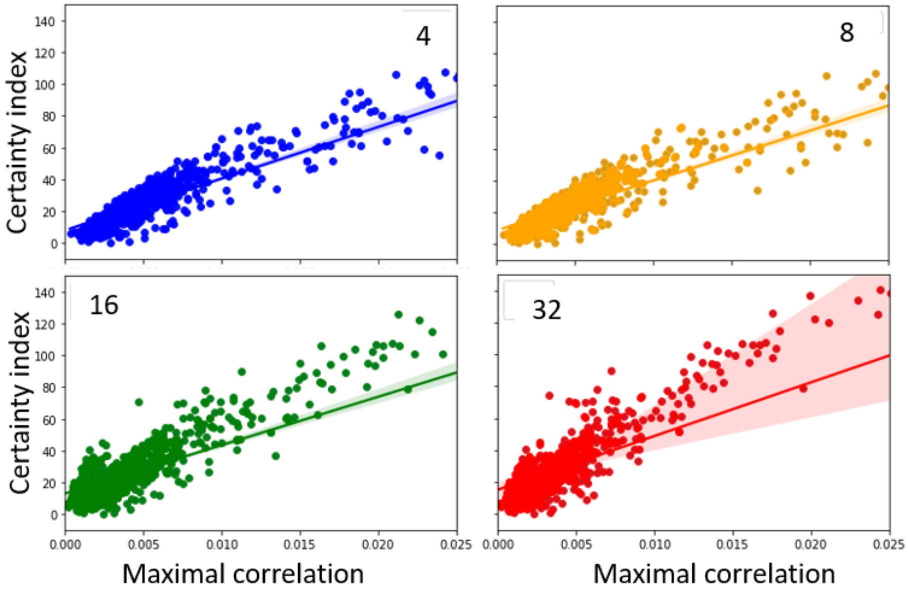


Figure 29. Sample entropy calculated for varying embedding dimensions: 4,8,16 and 32. The labels for the figures correspond to the individual embedding dimensions.

Table 9. Regression values: certainty index vs. correlation between heatmap and sample entropy at various embedding dimensions.

Statistical values	4	8	16	32
R-squared	0.91	0.91	0.83	0.63
F-statistic	11333	10402	5201	1705

<u>Sample entropy calculated at different scaling parameters.</u>

Similarly, if a high scaling parameter (*r*) value is selected, most of the templates will look like each other and they will fall below threshold and therefore the algorithm will have reduced efficiency. If *r* is too small, too many templates will fail to match. To address the effect of these variations, the sample entropy was calculated at various *r* values, while keeping an *m* =8 and *N* =100. Plots for the variations in the standard deviations (at 1.5, 2 and 2.5) is shown in Figure 30.

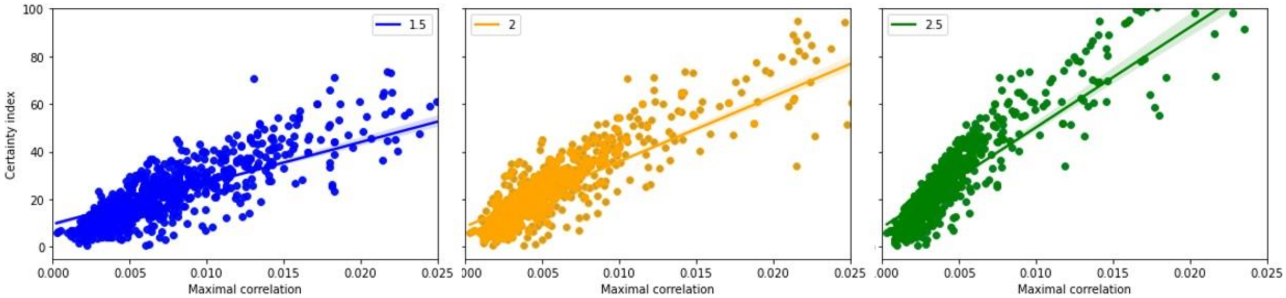


*Figure 30**.** Sample entropy calculated for different scaling parameters where (r =1.5, 2 and 2.5 x SD).*
The labels for the figures correspond to the individual scaling parameter x SD.

Table 10. Regression values: certainty index vs. correlation between heatmap and sample entropy at various scaling parameter (SD).

Statistical values	1.5	2	2.5
R-squared	0.93	0.93	0.93
F-statistic	14016	13796	14346

Sample entropy calculated at different sample lengths.

Another parameter considered in the calculation of sample entropy is length of the signal (*N*). This is the moving window of the original signal. Given the different frequency components in the EEG signal which will span different lengths of signal for each frequency, the signal length *N* was varied to calculate the sample entropy. Sample entropy was calculated at various N value, while keeping an m =8 and r =2. Plots for the variations in the standard deviations (at 1.5, 2 and 2.5) is shown in Figure 31.

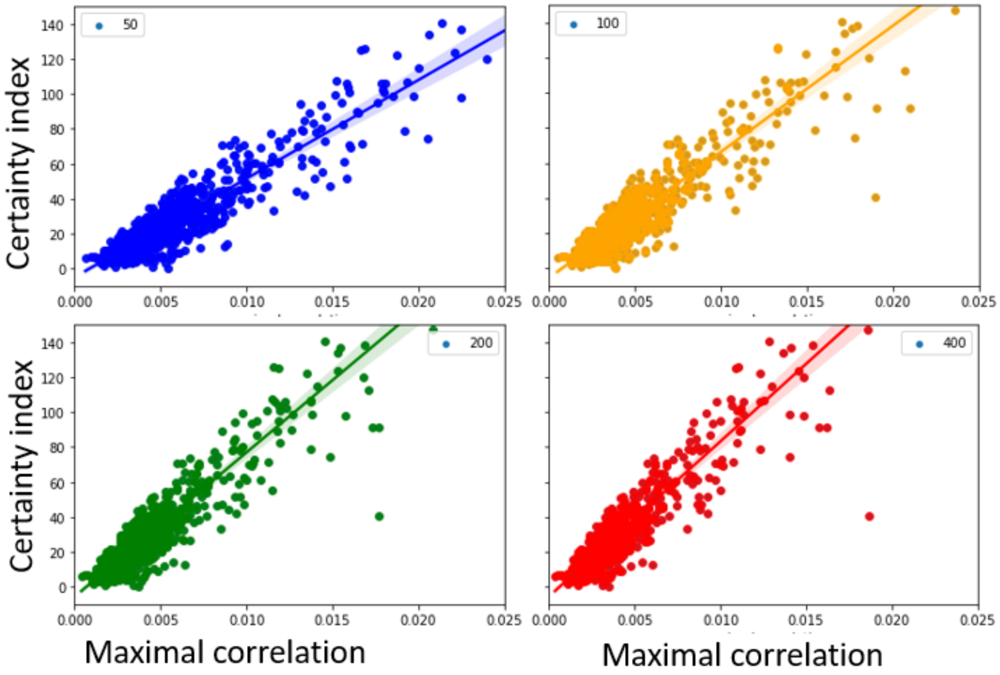


Figure 31. Sample entropy calculated for different signal lengths: 50, 100, 200 and 400. The labels for the figures correspond to the individual signal lengths.

Table 11. Regression values: certainty index vs. correlation between heatmap and sample entropy at various signal lengths.

Statistical values	50	100	200	400
R-squared	0.93	0.93	0.93	0.93
F-statistic	13509	13796	14039	13611

Discussion and comparison with existing literature

Our study shows that the heatmapping points to areas of increased instantaneous power as described by the envelope of the analytical signal) as contributory to the decision making of an epileptogenic or non-epileptogenic region.

Use of information theoretical approaches in evaluating the DNN

Our study evaluated how the ability of the DNN model to capture the information content in the signal influences the certainty about the predictions. For this purpose, we have introduced certainty index, as a measure of confidence of individual DNN predictions which is based on the outcomes prior to softmax layer. Given the fact that the heatmaps towards the final convolutional layers are representative of the highest-level abstractions of the information in the signal, that DNN is using to arrive at the decisions about sample classes, we have used a measure of the correlation between that heatmap in final convolutional layer and original signal to evaluate, how effectively the DNN has captured the information in the signal. Sample entropy measure was used as measure of information content in the signal (which was calculated independent of the model) and we were able to demonstrate that certainty index of each sample is proportional to the correlation between sample entropy and heatmap.

We would like to point out that the model that we developed for this purpose was comparable to the previous DNN models reported using intracranial EEG data (noted in Table 5). As noted previously, the main aim of this study was to explore the certainty of predictions of DNN with the help of heatmap and information content of the signal. The fact that the model we used is comparable to the reported models in accuracy, suggests a potential applicability of the approaches that we currently employed, in other models. It may be noted that the seizure detection studies usually have higher accuracy compared to classification of data to epileptogenic or non-epileptogenic from inter-ictal periods (time periods when there was no seizures recorded). The results from this study indicates that the certainty measure as described can be used as estimate of the confidence that the DNN has in the prediction of a particular data.

We are able to show that the certainty index is more strongly correlated with the correlation between the heatmap and sample entropy compared to that between heatmap and original signal. This would be expected given the fact that the sample entropy is the measure of the information and the heatmap is a depiction of the DNN's ability to capture the relevant information. Confidence measures for DNN predictions has been evaluated in the past, particularly using outputs of softmax layer. Logits from the softmax layers give a range of values that appears to give a confidence of predictions, but previous studies have also cautioned that this can be erroneous [66] especially given the discontinuous nature of input-output mappings and should be used judiciously.

Softmax confidence was further evaluated and has been considered an imperfect measure of uncertainty [67] especially for evaluating epistemic uncertainty. This study analyzed the softmax function and defined regions of softmax layers where an out of distribution input must fall to be correctly labelled as out of distribution. Statistics derived from softmax distributions was effective in determining whether a sample is misclassified or from a different distribution from the training data, suggesting its potential as a measure of certainty [68]. This study showed potential applications of this approach in diverse experimental data including computer vision and natural language processing. The method of certainty measurement that we are proposing incorporate the scores of the network (with values prior to the softmax layer), that favor a particular prediction and un-favor the other options, further enhancing the reliability of this measure. It may be noted that limitations attributed to the softmax based confidence measure may be present in this approach as well, but the improvisation incorporating the negative prediction outputs will hopefully make it more robust.

As discussed in the introduction, we preferred to use the values prior to softmax layer, for certainty index calculation, taking cues from biological neural networks. The other reason was that given the approach of Grad-CAM which used the gradients of DNN outputs prior to the softmax layer, estimation of certainty based on those values appeared more appropriate. While this may improve the reliability of this measure, further studies with different datasets may be needed for further ascertaining the wider applicability of this approach. While the measures of confidence have been addressed in various ways (Table 12), a rigorous evaluation of this measure from the standpoint of heatmaps and information content of the samples has not been reported.

Our main objective in this study was to establish how strongly this kind of confidence measures relate to the measures of information content in the signal. To the best of our knowledge, this is the first study addressing the relationship between confidence of prediction (as measured by certainty index), heatmap and information content of the sample. How the research questions are answered. The DNN model implemented in this study was able to predict the epileptogenic zones/non-epileptogenic zones with high accuracy. The heatmapping approaches along with evaluation of analytical signal showed that there is a high degree of correlation between those to signals indicating that the transients noted in the analytical signal will be significantly contributing to the decision making of DNN. This will help in unravelling how the DNN decision making can be corroborated with clinical decision making. The certainty index was introduced as a measure of confidence in the decision process of DNN.

We were able to demonstrate that there is a high correlation between sample entropy and heatmap and the value of that is correlated with the certainty in the DNN decisions in each sample. This establishes a direct correlation between information content in the signal and certainty of DNN decisions, which has not been demonstrated previously.

We evaluated how the indexes of certainty and correlations with the information measures that we used in this study compares with the existing literature. It may be noted that not many studies addressed the exact question of the relationship of certainty of decisions and information content in the signals. The table below shows of the studies that addressed potential certainty indexes, evaluated the potential utility of softmax predictions assessed the entropy of predictive distributions.

Table 12. Existing literature compared with our approach for evaluation of the certainty of the network and assessment of the certainty measure.

Publication	**Employed Method**	**Assessment of the Confidence Measures**	**Comparison to Our Approach**
Hendrycks et al. [30]	Softmax prediction probability	Correctly classified examples tend to have greater. maximum softmax probabilities	Did not assess for the relationship between information content in the samples.
Jha et al. [32]	Attribution based confidence measure	Studied effect of changing the labels of features away from the sample studied and conformance of model predictions.	Established attribution based. dimensionality reduction
Smith et al. [31]	Mutual information and softmax variance	Mutual information, expected. Kullback-Leibler Divergence and predictive variance help in computing the divergence between softmax and expected softmax.	Considered softmax variance. as a measure of mutual information
Pearce et al. [67]	Analytically studied softmax layer	Studied the effectiveness of softmax outputs as proxy for epistemic uncertainty in non-adversarial, out of distribution examples	Suggested partial capture of uncertainty. Did not explore relationship with heatmaps or information content of samples
Lakshminarayanan et al. [69]	Ensembles of neural networks	Used 1. scoring system as training criterion, 2. adversarial training	Evaluated entropy of predictive distributions to evaluate quality of uncertainty estimates; Evaluated performance compared to Bayesian networks

CHAPTER 5: GRADIENTS OF DNN AND CERTAINTY OF PREDICTION

[This work is being submitted as article “Activation gradients in deep neural networks and information content in the signal”, for IEEE-EMBS International conference on Body Sensor Networks: Sensor and Systems for Digital Health (IEEE BSN 2023)]

In the previous chapters we have described how the DNN model was developed to evaluate the iEEG, the relationship between iEEG analytical signal and heatmap and how the information content in the signal is correlated with the certainty of the decisions of DNN. As a next step we explored the flow of information through the DNN and demonstrated the relationship between activation gradients across layers and information content in the signal. We further established that in the case of iEEG signals, the activation map-DNN prediction relationship across various layers is correlated with the certainty of the DNN for specific prediction.

Relationship between DNN gradients and Certainty of decisions

The decision-making process in DNN is going through multiple layers of networks. As another way of assessing the relationship between the class discrimination and class activation maps the gradients for sample for each class can be calculated for individual input and averaged over all the inputs, which fall in that particular category. The method is illustrated below with the task being classification of signal to two categories (Figure 32):

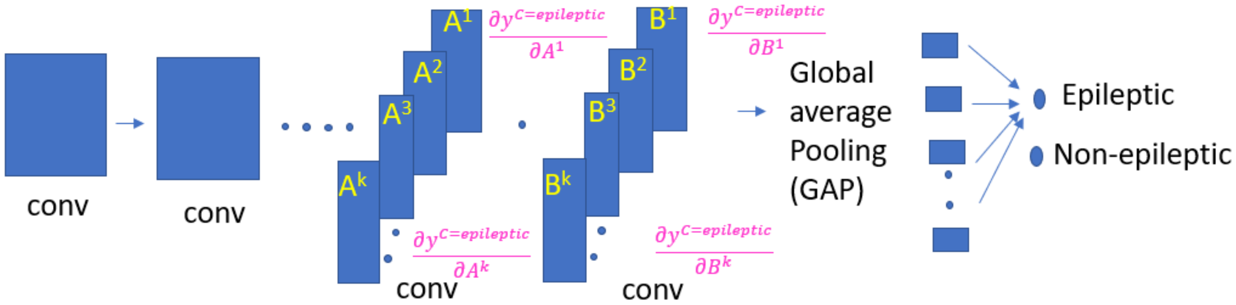


Figure 32. Schematic of the CNN layers with convolutional layers based on which the activation gradients are calculated.

The gradients for each sample's score (y^c) in relation to the activation maps will be given as a summation of all the activations gradients as:

$$Gradient_{LayerA}^{epileptic} = \sum_{n=1}^{k} \frac{\partial y^{C=epileptic}}{\partial A^{k}} \tag{29}$$

$$Gradient_{LayerB}^{epileptic} = \sum_{n=1}^{k} \frac{\partial y^{C=epileptic}}{\partial B^{k}} \tag{30}$$

It may be noted that this these gradients indicate the rate of change of the DNN output in relation to the activation maps.

Results

The activation gradients (AG) in the case of intracranial EEG data were calculated based on the convolutional model. To evaluate the relationship of these gradients to the decision-making certainty, each set of gradients was plotted against the sequence of layers from input to output (Figure 33). Dropout and pooling layers were excluded in this plot as there was no activation map in those layers. The slope for each of these layers was calculated and plotted against the certainty of the DNN out for each sample. Since AG significantly contribute to the final decision of the network, given the relationship between heatmap and certainty index, as noted in the previous chapters, we explored how the slope of the AGs across layers is related to the certainty index. The slope estimated in each case was plotted against certainty index for that sample data (Figure 34).

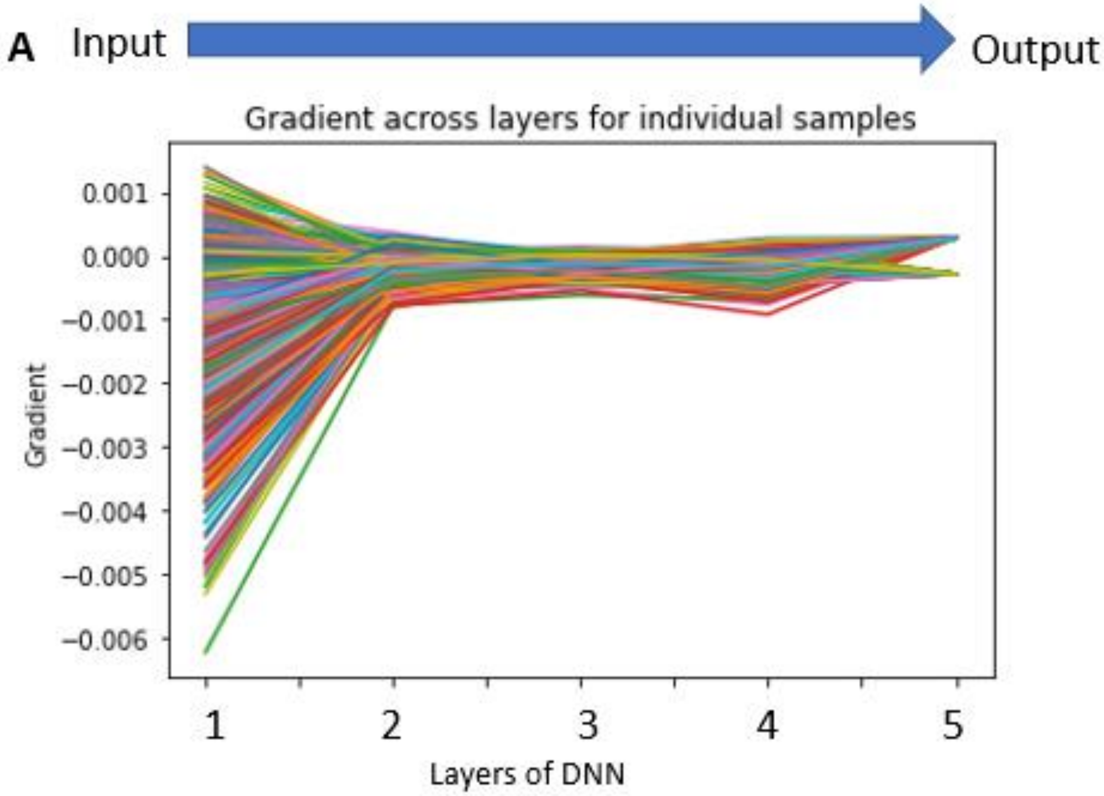


Figure 33. Change of activation gradients across DNN layers.

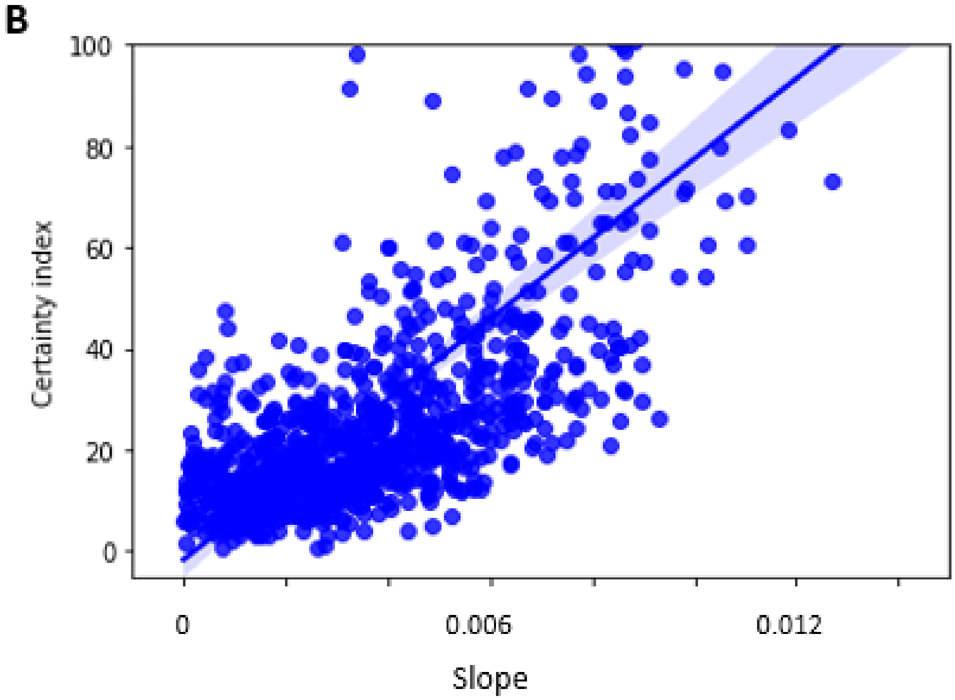


Figure 34. Relationship between the slope of activation gradients and certainty index.

The relationship between certainty index and slope of the activation gradients was explored further in the case of epileptic and non-epileptic data separately, which again showed a similar trend. This demonstrated a higher correlation in the case of non-epileptic data (Figure 35).

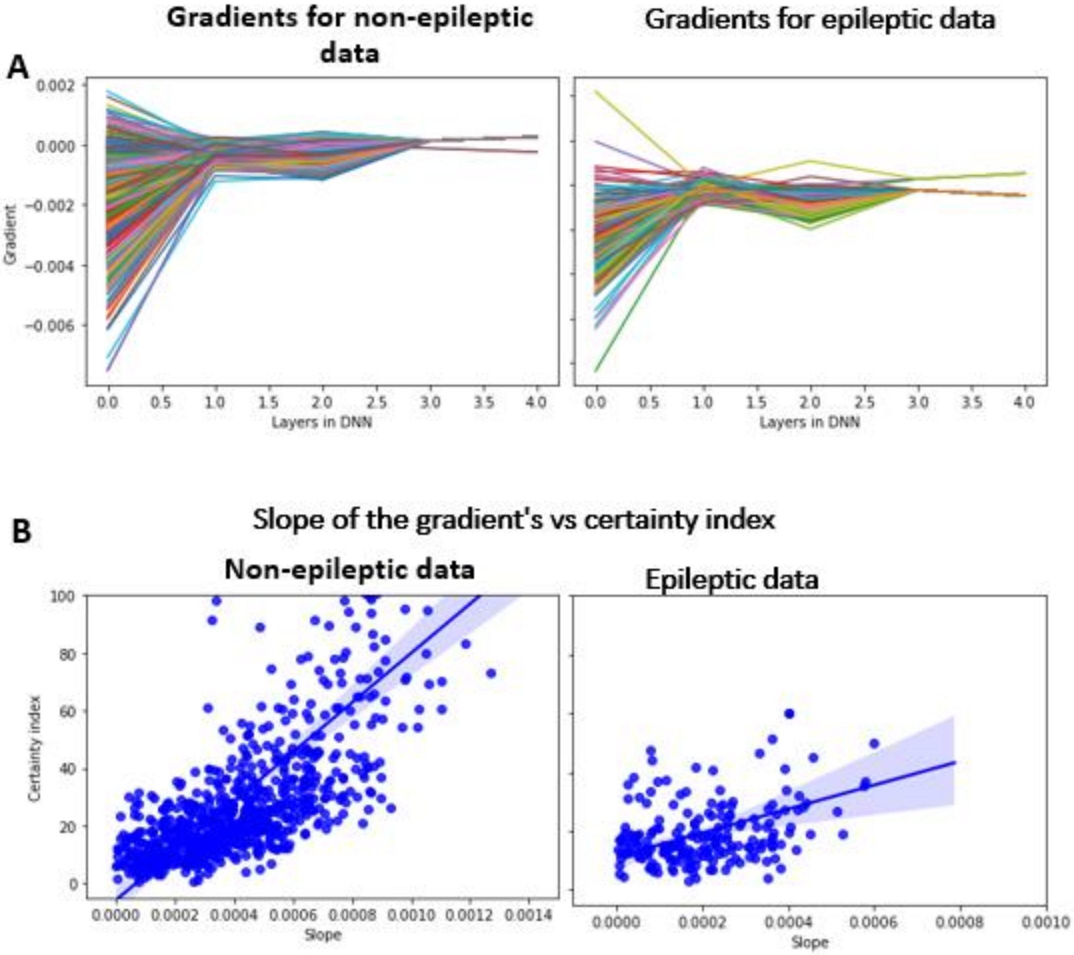


Figure 35. Slope of the gradients in the case of epileptic and non-epileptic data.

<u>Relationship between the gradients of activation with information content of the signal</u>

As detailed in the previous chapter we have shown the relationship between the certainty index and the information content in the signal. In the previous section we were able to show that a relationship exists between the certainty of predictions and the slope of activation gradients.

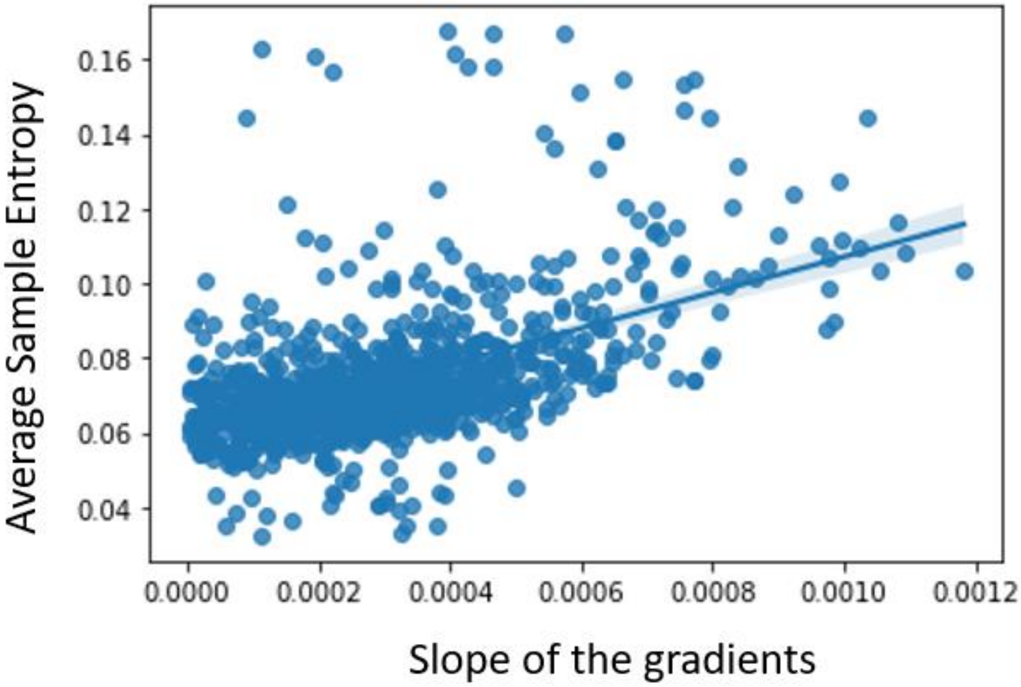


Figure 36. Relationship between the slope of the gradients across the network and the average sample entropy of the signal

The R-squared value for correlation between slope of gradients and average sample entropy was 0.64 with an F-statistic of 1764.

These results indicate that the rate of change of the learning as denoted by the change of output in relation to activation maps in each layer is directly proportional to Sample Entropy of the signal.

$$\frac{\partial}{\partial L}\left(\sum_{n=1}^{k}\frac{\partial y^{C}}{\partial L^{k}}\right)_{Signal(i)} \propto SampEn(Signal(i)) \tag{31}$$

Where L indicate the different convolutional layers of DNN (e.g., A and B in Figure 32) from the beginning to the last convolutional layer,

y^C is the DNN output for a particular class *c (before softmax)*.

$L^k$is the k-th activation map for signal in the layer L.

Signal(i)—represents the individual signals being analyzed and c represents the specific class to which the DNN has assigned Signal(i).

Discussion

This analysis establishes a correlation between the gradients of DNN outputs with respect to activation maps and the confidence of the DNN predictions. The fact that the trend remains the same on both epileptic and non-epileptic data when analyzed separately, suggest the underlying nature of the DNN gradients and the relation between activation maps and the results. A method of DNN pruning with gradient flow based saliency in DNN's have been described [70]. This method used Taylor expansion for global pruning of the network. Gradient based optimization of the architecture of the network has been suggested, so that the network parameters and architecture can be changed simultaneously, based on activation gradients [71]. To the best of our knowledge no study has explored the evaluation of the changes of activation map-based gradients over the different layers. This opens an avenue of better analysis of how the gradients are changing and how it is related to the DNN outcomes.

CHAPTER 6: SUMMARY

Taking motivation from the challenges in the iEEG analysis, we demonstrated the potential of deep learning methods in evaluation of intracranial EEG acquired from human brain, while they were undergoing monitoring for epilepsy surgery evaluation. The interictal iEEG data was used in generating a models of dense layers and convolutional layers based deep neural networks. The findings from the model generated in this study suggest that the epileptogenic zones can be identified using the DNN with great accuracy. We report an accuracy of 99% in dense network models and 87 % with CNN. CNN provides the advantage of evaluating the local features with heatmapping thereby enabling better correlation for clinical experts. The fact that the DNN is able to predict the epileptogenic zone with significant accuracy from interictal data lasting one minute, is highly promising and can have significant benefit in epilepsy surgery planning.

The potential to predict epileptogenic zones with limited duration of iEEG signal has to be compared to the current clinical approach based on multiple days of iEEG recording and capturing seizure events. One of the concerns of using the DNN models in medical field is the limitations in understanding the decision-making process of DNN. The conventional medical decision-making process already involves putting together several pieces of data. Adding a black box -based system to this process can be difficult from the standpoint of clinicians as well as patients. The use of heatmapping technique in this study with the Grad-CAM algorithm has demonstrated that the decision making of DNN can be unraveled to a significant extent, which could help the clinical teams in appropriately incorporating these tools to the decision-making process. A similar approach with layerwise relevance propagation was described for classifying neurocognitive tasks [60] demonstrating the potential of this approach in understanding the neurophysiological patterns.

We employed analytical signals to assess the features of the signal that contribute to the decision making as identified by heatmaps. This approach not only helps with identifying signal features, which can be clinically correlated, but also helps in paving way for better understanding of neural signals and thereby helping in unraveling the neural processes. A similar approach of using heatmapping strategies to better understand system functioning was described in the case of electric motors where Grad-Cam was utilized in identifying the torque sensitive regions[72]. That study showed that design optimization could be performed using this approach. Potential medical applications in these directions could include assessing treatment effects in EEG/ iEEG by evaluating the heatmaps of those signals for specific effects.

In the second part of the study, we focused on the evaluation of the results of DNN using certainty index. To arrive at this index we use both the positive prediction value for the true class and the prediction values for other classes in the possible predictions. This approach gives a score for the confidence of individual predictions which may be specifically useful when utilizing DNN approaches for individual clinical decisions. This approach can be applicable in any DNN models and can be a useful tool in assessing individual predictions based on models.

Information content, being one of the fundamental features of the signals, we used the information theoretical approaches to further analyze the workings of DNN. We used the method of sample entropy (SampEn) for this analysis, which was chosen, given the regularity in the iEEG signals. We were able to demonstrate that the correlation between the heatmap and information content of the signal as captured by the sample entropy showed better relationship with the certainty index compared to the correlation between heatmap and original signal. This points to the possibility that the DNN can focus on the information rich regions of the original signal. To the best of our knowledge this the first demonstration of the superiority of information rich regions in the accurate DNN predictions. Intuitively, this appears to be feasible and the most likely natural process as DNN is essentially an information machine and is expected to capture, maximum relevant information in the signal. This relationship between information measures and heatmaps were noted to hold even on re-estimating these measures after changing the parameters of measurement.

The third part of the study focused on how the outputs of DNN are related to the activation maps of individual layers and how it changes across the DNN layers. We were able to demonstrate that the slope of the DNN activation gradients across layers (This is the gradient of DNN outputs with respect to activation maps) is related to certainty index in a positive way. This demonstrates the fact that the earlier layers themselves are able to suggest the possible outcome in the case of samples, where the outcome is certain. Evaluation in this manner would help in the assessment of network parameters of DNN (eg. Number of layers, rate of learning etc). Further these gradients were also related to sample entropy in a positive way showing the relationship between the activation gradients and information content of the signal.

Schematic of the analysis of DNN data processing and assessment of DNN with activation gradients and information entropy

Following figure summarizes the analysis of the model that was implemented in the second and third part of the study. It may be noted that as noted in the figure sample entropy signal used in the analysis is generated independent of the DNN model.

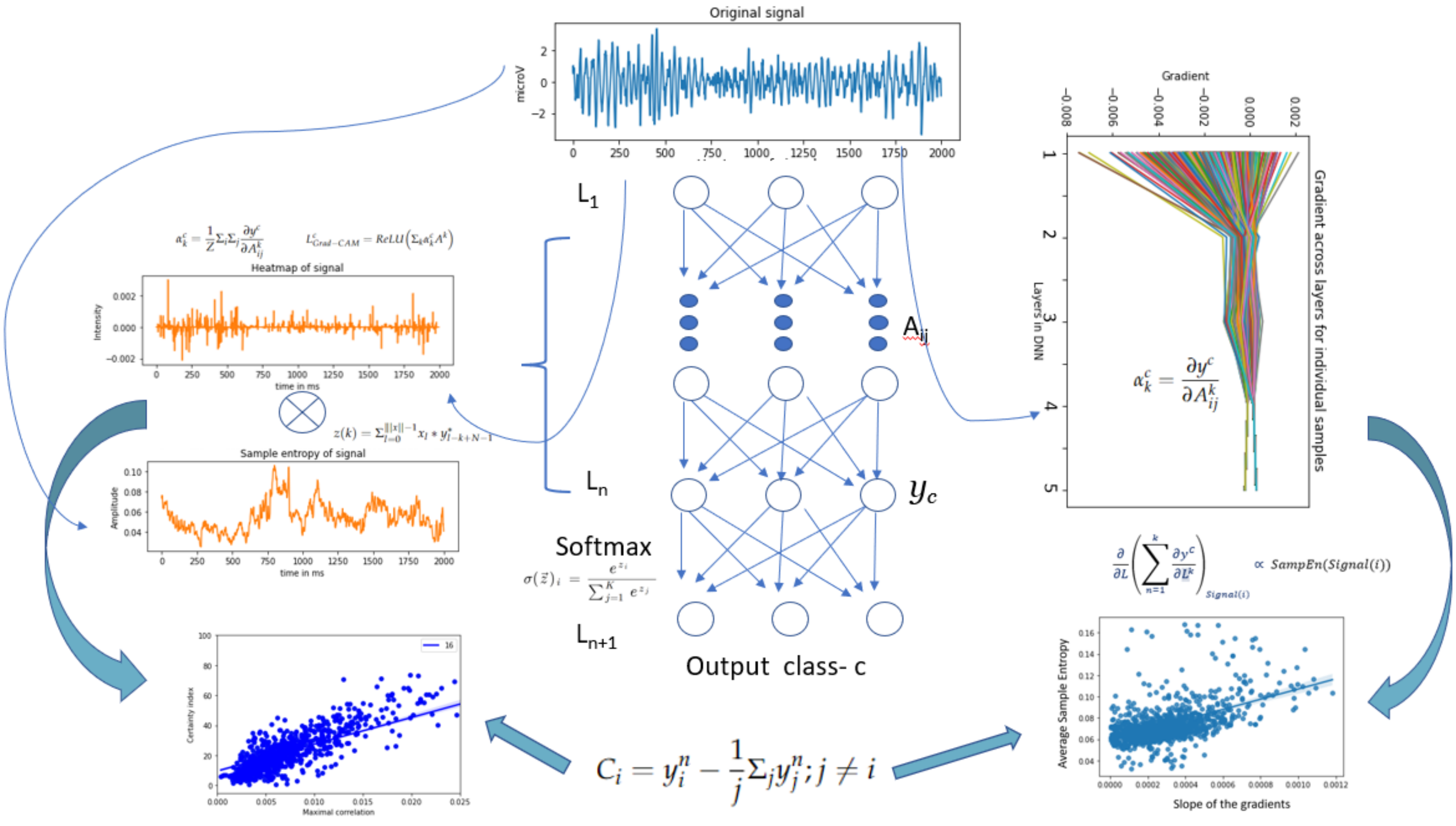

Original signal
microV
2
0
−2
0
250
500
750
1000
1250
1500
1750
2000
L_1
L_n
L_{n+1}
A_{ij}
y_c
Softmax
$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$
Output class- c
$\alpha_k^c = \frac{1}{Z}\Sigma_i\Sigma_j\frac{\partial y^c}{\partial A_{ij}^k}$
$L_{Grad-CAM}^c = ReLU\left(\Sigma_k \alpha_k^c A^k\right)$
Heatmap of signal
Intensity
0.002
0.000
−0.002
time in ms
$z(k) = \Sigma_{l=0}^{||x||-1} x_l * y_{l-k+N-1}^*$
Sample entropy of signal
Amplitude
0.10
0.08
0.06
0.04
time in ms
Certainty index
Maximal correlation
16
0.000
0.005
0.010
0.015
0.020
0.025
100
80
60
40
20
0
$C_i = y_i^n - \frac{1}{j}\Sigma_j y_j^n; j \neq i$
Gradient
−0.008
−0.006
−0.004
−0.002
0.000
0.002
Layers in DNN
1
2
3
4
5
Gradient across layers for individual samples
$\alpha_k^c = \frac{\partial y^c}{\partial A_{ij}^k}$
$\frac{\partial}{\partial L}\left(\sum_{n=1}^{k}\frac{\partial y^c}{\partial L^k}\right)_{Signal(i)} \propto SampEn(Signal(i))$
Average Sample Entropy
0.16
0.14
0.12
0.10
0.08
0.06
0.04
Slope of the gradients
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012

Figure 37. Summary of the approaches in the second and third parts of the study.
The middle panel shows the DNN process with the information flow from input to outrput. The left panel shows how the heatmap generated from the DNN process is crosss correlated with the SampEn calculated directly from the original signal. The right side panel shows the changes in the activation gradients across layers. Further it shows cross correlation between the slope of activation gradients and SampEn.

Core contributions to Computer Engineering

Using real world data acquired from human brain we were able to advance the deep learning methodologies as described below.

1. This study introduced a method to measure reliability of DNN predictions which can be used for assessing individual samples. This approach can be generalized to any neural network and can be used for comparison of DNN predictions.
   By incorporating both the negative and positive scores of penultimate layer of DNN, to evaluation of certainty, this approach presents and unique method to assess the confidence in each prediction. One of the advantages of our method is that it does not incur any significant additional computational costs to arrive at these estimates of confidence.
2. Further we established that a higher correlation between heatmap and sample entropy is associated with higher degree of certainty in the decision making of DNN. This approach establishes a direct relationship between information capturing capacity of DNN and the reliability of DNN predictions. We were able to demonstrate the robustness of this relationship at various parameter spaces. This information measure can act a s benchmark for evaluating DNN as capturing the information content will be essential for any meaningful decision making.
3. Demonstrated that the sample entropy signal (the time series representing the SampEn of original signal), featuring the information content in the original signal has stronger relationship with heatmap and certainty of decision, that that of original signal. This finding again points to the DNN's capability to focus on

the information rich regions of the original signal. This could suggest that the sample entropy calculation could act as a pre-filter to present the information rich regions of the signals which would reduce the computational costs of DNN training and predictions.

4. Evaluating the gradients in DNN layers, we demonstrated that the slope of the activation gradients is related to the certainty of decision of DNN.

   Given the complex nature of the DNN gradients, evaluating how their pattern is related to the outcome of DNN predictions would be important in designing and training DNN. These approaches will help in deciding the optimal parameters of DNN (e.g. no of nodes, layers, leaning rate etc.) along with improving the speed and reliability of predictions.

5. Further we were able to demonstrate that the DNN output-heatmap relationship is proportional to the information content in the signal.

   Information being the fundamental substrate based on which DNN is arriving at decisions, establishing the direct relationship between activation gradients and information content will pave the way for better understanding DNN learning process.

6. We also demonstrated that the analytical signal calculated through Hibert transform better correlates with the heatmaps, indicating that the DNN is focusing on the energy rich regions of the signal.

   Utilizing these strategies would enable us in better understanding the DNN learning processes and would help in correlating the DNN results with the

> preexisting signal processing-based conclusions. Our method demonstrates the potential roles of signal processing tools in better understanding the functioning of DNN.

Having better insight into the workings of DNN may help in better adoption these technologies in medical field and other fields of demand. Apart from helping as tools in the specified fields DNN may act as a method for improving the scientific understanding of the corresponding fields.

Relevance in medical applications

The possibility that the epileptogenic zone cane be identified with one minute of data has tremendous potential to change the way patients with intractable epilepsy are treated. This could reduce the need of admitting the patients to intensive care units (ICU) after the implantation of electrodes. It would reduce the cost of overall care by limiting the number of ICU days and avoiding the need of using the operating room for the second time. In addition, this kind of an advancement can reduce the medical complications and associated morbidity and mortality related to the seizures during the monitoring.

REFERENCES

1. Solla, S.; Levin, E.; Fleisher, M. Accelerated Learning in Layered Neural Networks. *Complex Systems* **1988**, *2*.
2. LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; Huang, F.J. A Tutorial on Energy-Based Learning.
3. Control, C. for D.; Prevention (CDC; others Epilepsy in Adults and Access to Care–United States, 2010. *MMWR. Morbidity and mortality weekly report* **2012**, *61*, 909–913.
4. Del Felice, A.; Beghi, E.; Boero, G.; La Neve, A.; Bogliun, G.; De Palo, A.; Specchio, L.M. Early versus Late Remission in a Cohort of Patients with Newly Diagnosed Epilepsy. *Epilepsia* **2010**, *51*, 37–42.
5. England, M.J.; Liverman, C.T.; Schultz, A.M.; Strawbridge, L.M. Epilepsy across the Spectrum: Promoting Health and Understanding.: A Summary of the Institute of Medicine Report. *Epilepsy & Behavior* **2012**, *25*, 266–276, doi:10.1016/j.yebeh.2012.06.016.
6. Beghi, E.; Giussani, G.; Nichols, E.; Abd-Allah, F.; Abdela, J.; Abdelalim, A.; Abraha, H.N.; Adib, M.G.; Agrawal, S.; Alahdab, F.; et al. Global, Regional, and National Burden of Epilepsy, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology* **2019**, *18*, 357–375, doi:10.1016/S1474-4422(18)30454-X.
7. Examining the Economic Impact and Implications of Epilepsy. **2020**.
8. Ansari, S.F.; Tubbs, R.S.; Terry, C.L.; Cohen-Gadol, A.A. Surgery for Extratemporal Nonlesional Epilepsy in Adults: An Outcome Meta-Analysis. *Acta neurochirurgica* **2010**, *152*, 1299–1305.

9. González-Martínez, J.A.; Srikijvilaikul, T.; Nair, D.; Bingaman, W.E. Long-Term Seizure Outcome in Reoperation after Failure of Epilepsy Surgery. *Neurosurgery* **2007**, *60*, 873–880; discussion 873-880, doi:10.1227/01.NEU.0000255438.13871.FA.

10. Dutta, S.; Manideep, B.; Rai, S.; Vijayarajan, V.; Sturm, I.; Lapuschkin, S.; Samek, W.; Müller, K.-R.; Swietojanski, P.; Ghoshal, A.; et al. Convolutional Neural Networks for Distant Speech Recognition. In Proceedings of the IOP Conference Series: Materials Science and Engineering; IEEE, 2014; Vol. 21, pp. 1120–1124.

11. Liu, J.; Wu, G.; Luo, Y.; Qiu, S.; Yang, S.; Li, W.; Bi, Y. EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder. *Frontiers in Systems Neuroscience* **2020**, *14*.

12. Al-Nafjan, A.; Hosny, M.; Al-Wabil, A.; Al-Ohali, Y. Classification of Human Emotions from Electroencephalogram (EEG) Signal Using Deep Neural Network. *ijacsa* **2017**, *8*, doi:10.14569/IJACSA.2017.080955.

13. Swietojanski, P.; Ghoshal, A.; Renals, S. Convolutional Neural Networks for Distant Speech Recognition. *IEEE Signal Processing Letters* **2014**, *21*, 1120–1124, doi:10.1109/LSP.2014.2325781.

14. Tiwari, V.; Pandey, C.; Dwivedi, A.; Yadav, V. Image Classification Using Deep Neural Network. In Proceedings of the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN); December 2020; pp. 730–733.

15. Kumar Mallick, P.; Ryu, S.H.; Satapathy, S.K.; Mishra, S.; Nguyen, G.N.; Tiwari, P. Brain MRI Image Classification for Cancer Detection Using Deep Wavelet Autoencoder-Based

Deep Neural Network. *IEEE Access* **2019**, *7*, 46278–46287, doi:10.1109/ACCESS.2019.2902252.

16. Jobst, B.C.; Cascino, G.D. Resective Epilepsy Surgery for Drug-Resistant Focal Epilepsy: A Review. *Jama* **2015**, *313*, 285–293.
17. Zahra, A.; Kanwal, N.; ur Rehman, N.; Ehsan, S.; McDonald-Maier, K.D. Seizure Detection from EEG Signals Using Multivariate Empirical Mode Decomposition. *Computers in Biology and Medicine* **2017**, *88*, 132–141, doi:10.1016/j.compbiomed.2017.07.010.
18. Anugraha, A.; Vinotha, E.; Anusha, R.; Giridhar, S.; Narasimhan, K. A Machine Learning Application for Epileptic Seizure Detection. In Proceedings of the 2017 International Conference on Computational Intelligence in Data Science(ICCIDS); June 2017; pp. 1–4.
19. San-Segundo, R.; Gil-Martín, M.; D'Haro-Enríquez, L.F.; Pardo, J.M. Classification of Epileptic EEG Recordings Using Signal Transforms and Convolutional Neural Networks. *Computers in biology and medicine* **2019**, *109*, 148–158.
20. Quon, R.J.; Meisenhelter, S.; Camp, E.J.; Testorf, M.E.; Song, Y.; Song, Q.; Culler, G.W.; Moein, P.; Jobst, B.C. AiED: Artificial Intelligence for the Detection of Intracranial Interictal Epileptiform Discharges. *Clin Neurophysiol* **2022**, *133*, 1–8, doi:10.1016/j.clinph.2021.09.018.
21. Ouichka, O.; Echtioui, A.; Hamam, H. Deep Learning Models for Predicting Epileptic Seizures Using IEEG Signals. *Electronics* **2022**, *11*, 605, doi:10.3390/electronics11040605.
22. Wang, X.; Wang, X.; Liu, W.; Chang, Z.; Kärkkäinen, T.; Cong, F. One Dimensional Convolutional Neural Networks for Seizure Onset Detection Using Long-Term Scalp and Intracranial EEG. *Neurocomputing* **2021**, *459*, 212–222.

23. Hussein, R.; Ahmed, M.O.; Ward, R.; Wang, Z.J.; Kuhlmann, L.; Guo, Y. Human Intracranial EEG Quantitative Analysis and Automatic Feature Learning for Epileptic Seizure Prediction 2019.

24. Sui, L.; Zhao, X.; Zhao, Q.; Tanaka, T.; Cao, J. Hybrid Convolutional Neural Network for Localization of Epileptic Focus Based on IEEG. *Neural Plasticity* **2021**, *2021*, e6644365, doi:10.1155/2021/6644365.

25. Wang, Y.; Dai, Y.; Liu, Z.; Guo, J.; Cao, G.; Ouyang, M.; Liu, D.; Shan, Y.; Kang, G.; Zhao, G. Computer-Aided Intracranial EEG Signal Identification Method Based on a Multi-Branch Deep Learning Fusion Model and Clinical Validation. *Brain Sciences* **2021**, *11*, 615, doi:10.3390/brainsci11050615.

26. Constantino, A.C.; Sisterson, N.D.; Zaher, N.; Urban, A.; Richardson, R.M.; Kokkinos, V. Expert-Level Intracranial Electroencephalogram Ictal Pattern Detection by a Deep Learning Neural Network. *Frontiers in Neurology* **2021**, *12*.

27. Roy, A.G.; Conjeti, S.; Navab, N.; Wachinger, C.; Initiative, A.D.N.; others Bayesian QuickNAT: Model Uncertainty in Deep Whole-Brain Segmentation for Structure-Wise Quality Control. *NeuroImage* **2019**, *195*, 11–22.

28. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. A Survey of Uncertainty in Deep Neural Networks. *arXiv preprint arXiv:2107.03342* **2021**.

29. Malinin, A.; Gales, M. Predictive Uncertainty Estimation via Prior Networks. *Advances in neural information processing systems* **2018**, *31*.

30. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks 2018.

31. Smith, L.; Gal, Y. Understanding Measures of Uncertainty for Adversarial Example Detection. *arXiv preprint arXiv:1803.08533* **2018**.

32. Jha, S.; Raj, S.; Fernandes, S.; Jha, S.K.; Jha, S.; Jalaian, B.; Verma, G.; Swami, A. Attribution-Based Confidence Metric for Deep Neural Networks. *Advances in Neural Information Processing Systems* **2019**, *32*.

33. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.-R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE transactions on neural networks and learning systems* **2016**, *28*, 2660–2673.

34. Ras, G.; Xie, N.; Gerven, M. van; Doran, D. Explainable Deep Learning: A Field Guide for the Uninitiated. *Journal of Artificial Intelligence Research* **2022**, *73*, 329–396, doi:10.1613/jair.1.13200.

35. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, June 2016; pp. 2921–2929.

36. Zurada, J.M.; Malinowski, A.; Cloete, I. Sensitivity Analysis for Minimization of Input Data Dimension for Feedforward Neural Network. In Proceedings of the Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS'94; IEEE, 1994; Vol. 6, pp. 447–450.

37. Bazen, S.; Joutard, X. The Taylor Decomposition: A Unified Generalization of the Oaxaca Method to Nonlinear Models. **2013**.

38. Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K.-R. Layer-Wise Relevance Propagation: An Overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R., Eds.; Springer International Publishing, 2019.

39. Abdullah, T.A.A.; Zahid, M.S.B.M.; Tang, T.B.; Ali, W.; Nasser, M. Explainable Deep Learning Model for Cardiac Arrhythmia Classification. In Proceedings of the 2022 International Conference on Future Trends in Smart Communities (ICFTSC); December 2022; pp. 87–92.

40. Le, K.H.; Pham, H.H.; Nguyen, T.B.T.; Nguyen, T.A.; Thanh, T.N.; Do, C.D. LightX3ECG: A Lightweight and EXplainable Deep Learning System for 3-Lead Electrocardiogram Classification. *Biomedical Signal Processing and Control* **2023**, *85*, 104963, doi:10.1016/j.bspc.2023.104963.

41. Shannon, C.E. A Mathematical Theory of Communication. *The Bell system technical journal* **1948**, *27*, 379–423.

42. Sinai, Y.G.; Sinai, Ya.G. On the Notion of Entropy of a Dynamical System.; Springer New York: New York, NY, 2010; pp. 3–10.

43. Pincus, S.M. Approximate Entropy as a Measure of System Complexity. *Proceedings of the National Academy of Sciences* **1991**, *88*, 2297–2301.

44. Delgado-Bonal, A.; Marshak, A. Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. *Entropy* **2019**, *21*, 541.

45. Richman, J.S.; Lake, D.E.; Moorman, J.R. Sample Entropy. In *Methods in enzymology*; Elsevier, 2004; Vol. 384, pp. 172–184.

46. Borowska, M. Entropy-Based Algorithms in the Analysis of Biomedical Signals. *Studies in Logic, Grammar and Rhetoric* **2015**, *43*, 21–32.

47. Chen, X.; Solomon, I.C.; Chon, K.H. Comparison of the Use of Approximate Entropy and Sample Entropy: Applications to Neural Respiratory Signal. In Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference; 2005; pp. 4212–4215.

48. Gao, X.; Yan, X.; Gao, P.; Gao, X.; Zhang, S. Automatic Detection of Epileptic Seizure Based on Approximate Entropy, Recurrence Quantification Analysis and Convolutional Neural Networks. *Artificial Intelligence in Medicine* **2020**, *102*, 101711, doi:10.1016/j.artmed.2019.101711.

49. Srinivasan, V.; Eswaran, C.; Sriraam, N. Approximate Entropy-Based Epileptic EEG Detection Using Artificial Neural Networks. *IEEE Transactions on Information Technology in Biomedicine* **2007**, *11*, 288–295, doi:10.1109/TITB.2006.884369.

50. Ruder, S. An Overview of Gradient Descent Optimization Algorithms 2017.

51. Li, L.; Doroslovački, M.; Loew, M.H. Approximating the Gradient of Cross-Entropy Loss Function. *IEEE Access* **2020**, *8*, 111626–111635, doi:10.1109/ACCESS.2020.3001531.

52. Vasudevan, S. Mutual Information Based Learning Rate Decay for Stochastic Gradient Descent Training of Deep Neural Networks. *Entropy* **2020**, *22*, 560, doi:10.3390/e22050560.

53. Agostinelli, F.; Hoffman, M.; Sadowski, P.; Baldi, P. Learning Activation Functions to Improve Deep Neural Networks 2015.

54. Lerma, M.; Lucas, M. Grad-CAM++ Is Equivalent to Grad-CAM With Positive Gradients 2022.

55. Gireesh, E.D.; Skinner, H.; Seo, J.; Ching, P.; Hyeong, L.K.; Baumgartner, J.; Gurupur, V. Deep Neural Networks and Gradient-Weighted Class Activation Mapping to Classify and Analyze EEG. *Intelligent Decision Technologies* **2023**, *17*, 43–53, doi:10.3233/IDT-228040.

56. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the European conference on computer vision; Springer, 2014; pp. 818–833.

57. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the Proceedings of the IEEE international conference on computer vision; 2017; pp. 618–626.

58. Batrakov, D.O.; Golovin, D.V.; Simachev, A.A.; Batrakova, A.G. Hilbert Transform Application to the Impulse Signal Processing. In Proceedings of the 2010 5th International Confernce on Ultrawideband and Ultrashort Impulse Signals; September 2010; pp. 113–115.

59. Duda, R.O.; Hart, P.E.; Stork, D.G. Pattern Classification (2nd Ed.).

60. Sturm, I.; Bach, S.; Samek, W.; Müller, K.-R. Interpretable Deep Neural Networks for Single-Trial EEG Classification 2016.

61. Sui, L.; Zhao, X.; Zhao, Q.; Tanaka, T.; Cao, J. Localization of Epileptic Foci by Using Convolutional Neural Network Based on IEEG. In Proceedings of the Artificial Intelligence Applications and Innovations; MacIntyre, J., Maglogiannis, I., Iliadis, L., Pimenidis, E., Eds.; Springer International Publishing: Cham, 2019; pp. 331–339.

62. Antoniades, A.; Spyrou, L.; Took, C.C.; Sanei, S. Deep Learning for Epileptic Intracranial EEG Data. In Proceedings of the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP); 2016; pp. 1–6.

63. Hossain, M.S.; Amin, S.U.; Alsulaiman, M.; Muhammad, G. Applying Deep Learning for Epilepsy Seizure Detection and Brain Mapping Visualization. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **2019**, *15*, 1–17.

64. Gireesh, E.D.; Gurupur, V.P. Information Entropy Measures for Evaluation of Reliability of Deep Neural Network Results. *Entropy* **2023**, *25*, 573, doi:10.3390/e25040573.

65. Yadav, N.; Yadav, A.; Kumar, M. *An Introduction to Neural Network Methods for Differential Equations*; SpringerBriefs in Applied Sciences and Technology; Springer Netherlands: Dordrecht, 2015; ISBN 978-94-017-9815-0.

66. Grassberger, P.; Procaccia, I. Estimation of the Kolmogorov Entropy from a Chaotic Signal. *Phys. Rev. A* **1983**, *28*, 2591–2593, doi:10.1103/PhysRevA.28.2591.

67. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. *arXiv preprint arXiv:1312.6199* **2013**.

68. Pearce, T.; Brintrup, A.; Zhu, J. Understanding Softmax Confidence and Uncertainty. *arXiv preprint arXiv:2106.04972* **2021**.

69. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv preprint arXiv:1610.02136* **2016**.

70. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in neural information processing systems* **2017**, *30*.

71. Liu, X.; Li, B.; Chen, Z.; Yuan, Y. Exploring Gradient Flow Based Saliency for DNN Model Compression. In Proceedings of the Proceedings of the 29th ACM International Conference on Multimedia; Association for Computing Machinery: New York, NY, USA, October 17 2021; pp. 3238–3246.

72. Grathwohl, W.; Creager, E.; Ghasemipour, S.K.S.; Zemel, R. GRADIENT-BASED OPTIMIZATION OF NEURAL NETWORK ARCHITECTURE. **2018**.

73. Sasaki, H.; Hidaka, Y.; Igarashi, H. Explainable Deep Neural Network for Design of Electric Motors. *IEEE Transactions on Magnetics* **2021**, *57*, 1–4, doi:10.1109/TMAG.2021.3063141.