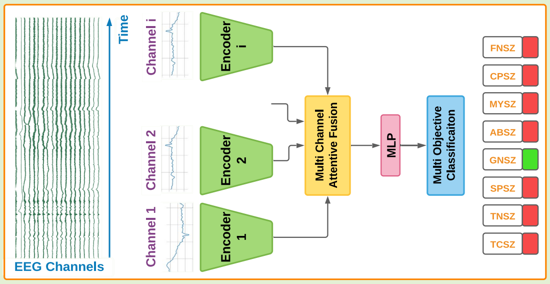
Interpretable Seizure Classification Using Unprocessed EEG With Multi-Channel Attentive Feature Fusion

Darshana Priyasad, Tharindu Fernando, *Member, IEEE*, Simon Denman, *Member, IEEE*, Sridha Sridharan, *Life Senior Member, IEEE*, and Clinton Fookes, *Senior Member, IEEE*

***Abstract*—Identification of seizure type plays a vital role during clinical diagnosis and treatment of epilepsy. However, the clinical evaluation of seizure type is highly dependent on the observed medical symptoms and the experience of the epileptologists who perform the evaluation. A key diagnostic tool is the electroencephalogram (EEG), which captures brain activity and can be used to determine the type of seizure occurring. EEG channels show non-stationary and dynamic behavior following the onset of a seizure event, and each EEG channel can display unique characteristics based on the seizure type and the epileptic foci. This paper proposes a novel deep learning architecture with attention-driven data fusion using raw scalp EEG data from a 10-20 layout, where independent shallow deep networks are trained on each channel. Unlike most state-of-the-art methods that first employ a data engineering step, we directly pass the EEG signal from each channel through a deep convolutional network consisting of SincNet and Conv1D layers, which learn robust features directly from the input signals, increasing model interpretability. However, the importance of each channel and the temporal information varies based on conditions particular to the recording, and this can adversely affect the overall recognition. We propose an approach based on the attentive fusion of channels to ensure only salient features from individual channel encoders are captured, passing the fused information to a Deep Neural Network for classification. Our proposed method has obtained an average F1-score of 0.967 on the Temple University Hospital Seizure Corpus, the largest publicly available seizure dataset.**



***Index Terms*—Attention, multi-channel fusion, seizure classification, SincNet, raw waveform.**

I. Introduction

EPILEPSY is a common, but serious neurological brain disorder, a key symptom of which is temporal disruptions of electrical activity in the brain (seizures). Epilepsy can have many possible known (symptomatic) and unknown (cryptogenic) causes [1], [2]. A seizure can originate in any area of the brain and it may or may not spread to other areas [1], [3], and seizures are treated differently based on the type. Seizure type is typically identified based on medical history and clinical signs supported by Electroencephalograms (EEG). The impacts of seizures (intellectual disability, learning difficulties, mortality risk and etc. [4]) can be mitigated through correct seizure type identification, followed by drug therapies or surgery [5]. However, inadequate and inaccurate clinical histories and overlapping symptoms can make seizure classification a challenging task, even for experts [6]. When a decision on the seizure type cannot be made through clinical diagnosis, a video-EEG is used where neurologists visually examine recordings. Since patients can be monitored for days, the visual examination can be tedious. Therefore, automated seizure classification through EEG signals is an appealing potential method to assist neurologists.

Advances in deep learning have recently demonstrated huge success in healthcare-related applications [7], [8] in areas such as the analysis of clinical imaging [9], [10], genomics [11], seizure detection [12], [13] and classification [14], [15] and, mobile healthcare devices [16], [17]. However, data quality, data volume, interpretability, and domain complexity remain challenges when using machine learning and deep learning based techniques in healthcare applications [7].

Epileptic-seizure related machine learning architectures fall into three broad categories: seizure detection, seizure prediction, and seizure classification; which are associated with

Manuscript received May 7, 2021; revised June 4, 2021; accepted June 4, 2021. Date of publication June 17, 2021; date of current version August 31, 2021. This work was supported in part by the Australian Research Council (ARC) Discovery under Grant DP140100793. The associate editor coordinating the review of this article and approving it for publication was Dr. Varun Bajaj. *(Corresponding author: Darshana Priyasad.)*

The authors are with the Signal Processing, Artificial Intelligence and Vision Technologies (SAIVT) Research Program, Queensland University of Technology, Brisbane, QLD 4000, Australia (e-mail: dp.don@qut.edu.au; t.warnakulasuriya@qut.edu.au; s.denman@qut.edu.au; s.sridharan@qut.edu.au; c.fookes@qut.edu.au).

Digital Object Identifier 10.1109/JSEN.2021.3090062

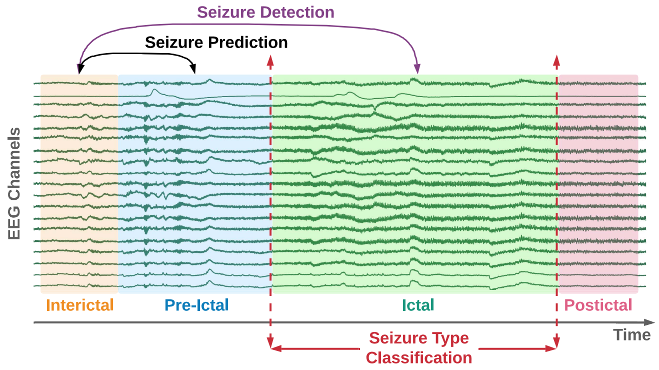


Fig. 1. The four main brain states during a seizure event: interictal, pre-ictal, ictal, and postictal; along with three common machine learning tasks associated with epileptic seizure analysis: seizure detection, seizure prediction, and **Seizure Classification**. Seizure prediction focuses on predicting the brain state prior to the seizure onset (forecast a seizure in advance) while seizure detection focuses on separating seizure and non-seizure segments of an EEG [18]. In this study, we focus on **Seizure Classification**, where given a segmented EEG with a seizure event, we identify the type of seizure..

different brain states during an epileptic seizure event as illustrated in Figure 1. Interictal, pre-ictal, ictal, and postictal refer to the normal brain state of a patient, brain state before a seizure event, brain state with a seizure event, and brain state after the seizure event respectively [18]. In *seizure prediction*, the objective of the classifier is set to predict a future seizure event by identifying the pre-ictal brain state of a subject [19]. In *seizure detection*, a classifier is designed to distinguish between EEG segments that correspond to seizure and non-seizure segments (refer Figure 1) [13]. In this study, we are focusing on *seizure classification*, where an EEG segment with a seizure event is classified into a specific seizure type (refer Section III for the considered seizure types). A brief description of the current advances in automated seizure classification is given below.

A. Related Work

Early works on automated seizure classification predominantly focused on hand-crafted (pre-deep learning) machine learning-based architectures. Li *et al.* proposed a feature extraction technique for EEG using Multi-scale Radial Basis Functions and Particle Swarm Optimization followed by Principal Component Analysis for dimensionality reduction, and a Support Vector Machine (SVM) with Radial Bias Function for classification [20]. Roy *et al.* evaluated the importance of machine learning methods (Stochastic Gradient Descent (SGD), k-Nearest Neighbours (k-NN), and XGBoost) and general deep learning architectures (AlexNet [21], VGG [22], and ResNet [23]) for automated seizure classification [24]. The results show that some hand-crafted machine learning-based techniques outperform deep learning-based methods with transfer learning. The main reason behind this is the significant domain difference when applying transfer learning to adapt networks originally trained on images to 1D signals. However, more recent deep learning architectures have been able to surpass hand-crafted machine learning-based approaches [14], [15].

Asif *et al.* proposed a deep learning framework for seizure classification based on multi-spectral feature embeddings. It is argued that the features learned through the network (SeizureNet) are capable of achieving higher accuracies on smaller networks through knowledge distillation [14], [25]. First, they transform the EEG data into saliency encoded spectrograms sampled at different frequencies and spatial resolutions, followed by using automated feature learning through separate deep nets, combining the features before final classification. However, this feature ensemble approach results in a high parameter count and increased inference time. Furthermore, the generation of spectrograms as the input to the network can result in information loss, and using 2D convolutions over features generated from 1D signals limits the interpretability of the network [26]. Zeng *et al.* has proposed a Hierarchy Graph Convolution Network based seizure classification model using time and frequency domain features extracted from EEG signals [27]. The topological relationship between electrodes is utilized (the electrodes near the epileptic foci show fluctuating and inconsistent voltages [27]) by the proposed network.

Liu *et al.* proposed a hybrid deep model (a combination of both Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)) for seizure classification using a Short Time Fourier Transform (STFT) of EEG signals [28]. Due to the use of both a CNN and RNN separately, it is argued that the network is capable of learning both spatial and temporal features from the data and the fusion using bi-linear pooling has resulted in increased accuracy over a single network, highlighting the importance of both spatial and temporal features for seizure classification. Ahmedt *et al.* proposed a Long-Short Term Memory (LSTM) and CNN based architecture with Neural Memory Plasticity (NMP) [15]. Their results suggest that the long-term mapping of relationships obtained through memory modules is capable of learning a clear separation of seizure classes. However, the class-wise recognition rates on the TUH-TUSZ dataset [29] suggest that the model has focused more on seizure types with a high number of samples (Focal Non-Specific Seizure (FNSZ), Generalized Non-Specific Seizure (GNSZ), and Complex Partial Seizure (CPSZ)), and the model struggles with seizure types with a low sample count. The main reason behind this behavior would be the memory, which learns long-term relationships related to majority classes, disregarding the minority classes.

B. Proposed Approach

Seizure events are considered non-stationary and dynamic, and result in rhythmic discharges in EEG signals [20]. Most approaches in the literature use manual feature extraction as a first step, even though deep learning has been shown to learn more robust and rich features from raw data. Furthermore, convolutions on stacked EEG channel features limit model interpretability due to spatial and temporal invariance, and disregard the fact that some channels can contain noisy or uninformative features. In this paper, we perform multi-class seizure classification [15] with scalp EEG recordings and attentive multi-channel fusion. Compared to state-of-the-art methods, we explore the viability of using raw EEG signals for seizure classification, avoiding time-consuming pre-processing, and increasing model interpretability. We have evaluated our proposed architecture for multi-class seizure classification and have analyzed the results with a baseline method with a similar experimental protocol and proportional class distribution [15]. The main technical contributions of this paper are:

1) An interpretable deep learning architecture to classify epileptic seizures using raw EEG signals, that can achieve consistent precision, recall, and F1-score over each seizure type, irrespective of the number of training samples. The direct use of EEG signals (without any pre-processing or feature extraction) helps the network to learn rich features through deep learning compared to relying on engineered features.
2) Unlike state-of-the-art methods, separate shallow deep networks are trained for each channel (19 channels from the 10-20 system are commonly used in state-of-the-art methods are used in this study), enabling each network to learn channel-specific features and increasing the interpretability of the network. Therefore, the proposed architecture can be easily extended to process new channels which have different frequency distributions, limiting model re-training to the fusion and classification sub-modules.
3) As all EEG channels may not contain salient information for all examples, an attention-based weighting mechanism operating over temporal and channel-wise data is introduced to mitigate adverse effects and ensure only salient information is passed to the classifier.

II. Methodology

In this paper, we propose a novel deep learning architecture for seizure classification using multi-channel EEG data. Since EEG channels develop a rhythmic activity after the onset of seizures [30], these patterns and their frequency components may contain salient information about the seizure type. As such, we use raw EEG wave-forms rather than engineered features. First, the EEG wave of each selected channel is passed through a convolution block with SincNet [31] filter (parametrized sinc functions) as outlined in Section II-A. Then the resultant features are passed through an attentive fusion block with temporal and channel-wise attention to filter features salient for the objective, as described in Section II-B. Finally, the fused features are passed through a classification network containing a MLP as described in II-C. The overall architecture of the proposed model is illustrated in Figure 2 and 3.

A. Encoder Networks for EEG Channels

In our approach, we have used raw EEG waveforms from 19 channels (see Section III-A) as the input to the deep network. Given that the EEG channels show rhythmic activity following the onset of seizures [30], we have used SincNet based Conv-1D layers to learn channel specific features. The parametrized sinc function based convolutional layer (SincConv layer in Figure 2) provides a compact and customized filter-bank that is tuned for the application domain during model training [31]–[33]. Thus this layer is used as the first layer of our network to learn application specific and meaningful filters. Since the SincConv layer uses band-pass filters (parameterized sinc functions with high and low cutoff frequencies as learnable parameters), it offers flexibility to extract relevant information while providing a clear physical meaning with higher interpretability. Furthermore, SincNet models have shown an ability to achieve better performance compared to standard CNNs on raw waveforms, with faster convergence [31].

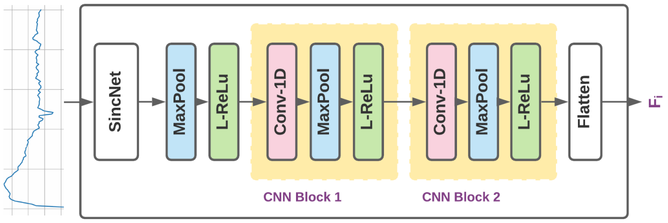


Fig. 2. Encoder for channel i: The encoder network (a customized SincNet model) used to learn features from the raw EEG signal uses a combination of SincConv (parametrized sinc function based convolutions) and 1D-convolutional layers for the i^{th} channel. Identical networks are trained for all channels, and the flattened feature vectors $f_i \in$ (0-18) are passed to the fusion component. All “Max Pooling”, “Leaky-ReLu” and “Conv-1D” layers share the same configuration.

Consider $x_i[n]$ as a discrete EEG segment of length 1 second, where $i \in (0, 1, \ldots, 18)$ is the input to block i which learns features from channel i. First, the above 1s segment is passed through the SincConv layer which performs convolutions using a predefined function g with learnable parameters θ, in contrast to the standard CNN convolutions with Finite Impulse Response filters [34], as in Equation 1,

$$y_i[n] = x_i[n] * g_i[n, \theta_i]. \tag{1}$$

The function g (filter bank with rectangular band-pass filters) is defined as two low-pass filters in the frequency domain as in Equation 2 where $f_{(i,1)}$, $f_{(i,2)}$ and $rect$ are the low cutoff frequency, high cutoff frequency and a rectangular function [31]. The time domain representation of the function g, G, is given by Equation 3 [31]. The cutoff-frequencies are initialized between 0 and $f_s/2$, where f_s is the sampling frequency,

$$G_i[f, f_{(i,1)}, f_{(i,2)}] = rect(\frac{f}{2f_{(i,2)}}) - rect(\frac{f}{2f_{(i,1)}}); \tag{2}$$

$$\begin{aligned} g_i[n, f_{(i,1)}, f_{(i,2)}] &= 2f_{(i,2)}\, sinc(2\pi f_{(i,2)} n) \\ &\quad -2f_{(i,1)}\, sinc(2\pi f_{(i,1)} n) \\ where\; sinc(x_i) &= \frac{sin(x_i)}{x_i}. \end{aligned} \tag{3}$$

The sampling rate, number of filters, and the length of the filters in the SincConv layer are set to 250, 80, and 125 respectively. The output of the SincConv layer is then passed through a max-pooling layer with size 2, followed by a Leaky ReLU activation with alpha of 0.2. The resultant feature is passed through 2 1D-CNN blocks with 40 filters with a filter length of 5, followed by a max pool layer and a Leaky ReLU activation (alpha of 0.2). Finally, the feature from the last layer

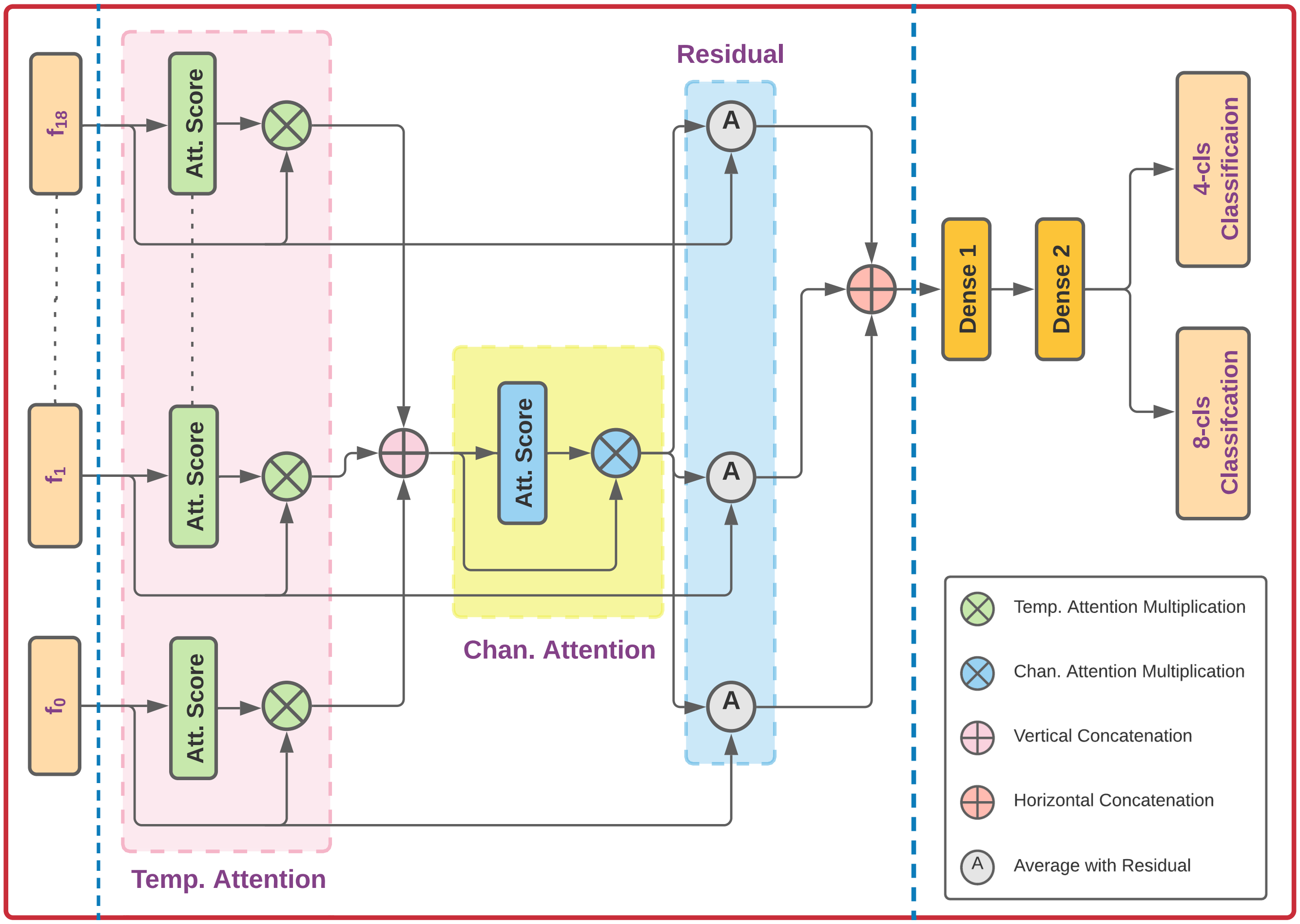


Fig. 3. The attentive feature fusion and classification architecture proposed in this paper. The figure illustrates the fusion of 3 channels, though the proposed approach uses 19 channels in the same manner. First, the output feature from each uni-channel encoder (see Section II-A) is weighted with uni-channel attention scores and stacked (vertical concatenation). Since the attention is applied to weight the importance of the channel towards the objective, the spatial arrangement of channels based on the spatial location of the EEG electrodes is not considered. Then each channel feature is weighted with channel importance calculated through an attention mechanism that is jointly trained with the network (refer to Section II-B). Then, the input feature vector and the corresponding attended feature vector from each model are averaged, followed by a naive concatenation of features. Finally, the fused features are passed through two dense layers with 512 units followed by 4-class (3-majority classes and all others as one class) and 8-class (all seizure types) classification outputs (refer to Section II-C).

of the block is flattened (f_i) to obtain a 1-D feature vector to be passed to the fusion layer.

B. Temporal and Channel-Wise Attentive Fusion

Multi-channel feature fusion is the second component of our proposed method. Considering the channel variations, the information contained at all time-steps across all channels will not be equally important for the final objective due to noise and redundancy, which can adversely affect network performance. Therefore, we propose an attentive fusion model to remove extraneous information.

First, the temporal importance of each feature in the feature vector f_i (flattened feature vector from channel i) is calculated via an attention score (using a $sigmoid$ function) as per Equations 4 and 5, where α_i and b_i define the temporal attention score and the attended feature,

$$\alpha_i = sigmoid(f_i), \tag{4}$$

$$b_i = f_i \cdot \alpha_i. \tag{5}$$

The $sigmoid$ over a vector gives the importance of each element towards the objective, considering elements are not mutually exclusive, while the $softmax$ distributes the probability across all elements. Since there can be multiple informative elements in each vector, we have used a $sigmoid$ activation to consider each feature independently as $softmax$ can struggle to capture multiple relevant elements.

Then, the attended feature vectors from each channel are stacked together and channel-wise attention is applied. We have used $softmax$ attention to weight the channel importance,

$$\beta = softmax(s), \tag{6}$$

$$c = f_{ca}(s, \beta). \tag{7}$$

where s, β, f_{ca} and c refer to the stacked feature vector, the channel-wise attention score, and the function to calculate attention weighted feature vectors. Since this process uses a $softmax$ activation (importance/weights over each channel adds to one), it also provides strong regularization.

Then the feature vector, f_i, and the corresponding attended vector, c_i, from c are averaged. Finally, each feature vector is concatenated (o_f) and passed to the classification layer

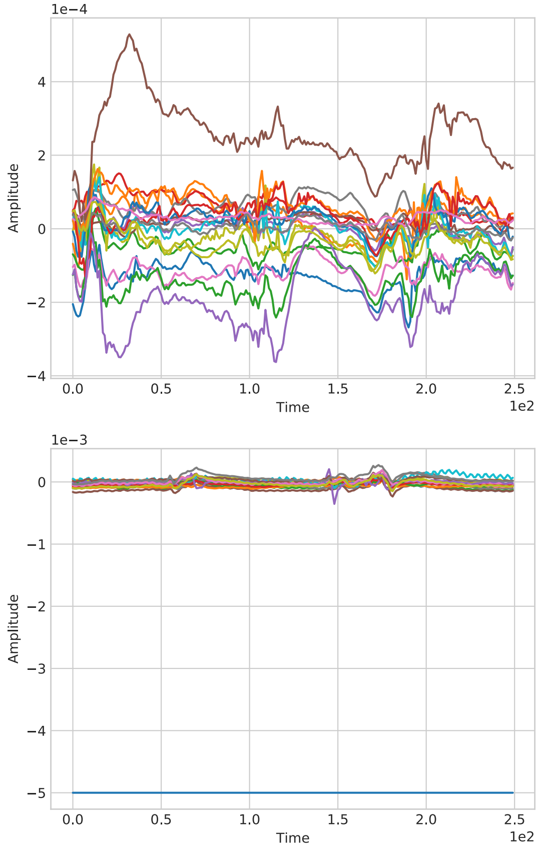


Fig. 4. Two 1 second segments from 19 channels for the ABSZ (top) and TCSZ (bottom) seizure classes, after the re-sampling described in III-A.

(see Equation 8 where ⊕ refers a concatenation operation) as described in Section II-C,

$$o_f = \frac{(f_{i=1} + c_{i=1})}{2} \oplus \frac{(f_{i=2} + c_{i=2})}{2} \oplus \ldots. \quad (8)$$

Figure 4 illustrates the importance of having an attentive mechanism to select salient features from signal segments. Considering the bottom example, it is observed that all channels carry similar information in the EEG, except for one channel which has a much lower amplitude. In this scenario, this information can have an adverse impact on the learned features from the deep network which may impact recognition performance. The proposed channel attention mechanism can decrease the above risk by weighting the channels based on attention scores (a measure of salience) as described in Equation 6. When considering the top example, a rhythmic activity is observed in most EEG channels (a rhythmic activity with multiple frequency components is observed in EEG channels after seizure onset [28], [30]). Even though the proposed encoder networks are capable of handling these varying signal frequencies, all features may not carry significant information for classification. Therefore, to filter out uninformative temporal features from the encoder, temporal attention is applied as in Equation 4.

C. Classification Network for Fused Channel Features

The final component of the model operates over high-level features from the fused data. Since the fused data preserves the temporal and channel-wise relationships in the EEG data, different deep net architectures can be used. However, we have selected a Deep Neural Network (DNN) with multiple fully connected layers, which resulted in the highest accuracy,

$$o_d = ReLU\ (\ ReLU\ (o_f w_{d_1} + b_{d_1})\ w_{d_2} + b_{d_2}). \quad (9)$$

First, the fusion output (o_f) is passed through two successive fully connected layers with 512 neurons in each, with a ReLU activation as per Equation 9, where w_{d_i} and b_{d_i} refer to the weight and the bias of layer i respectively. The output of the second dense layer (o_d) is then passed to the classification layers. In this work, we use two classification layers as illustrated in Figure 3 where the layers classify the input EEG segment to 1). 8 classes (8-cls classification) and 2). 4 classes; FNSZ, GNSZ, CPSZ and all others types (4-cls classification). The second classification layer is trained along with the network to reduce the confusion between minority classes and the majority classes via a related sub-task.

TABLE I
TUH SEIZURE DATASET (V1.4.0) STATISTICS

Seizure	Patients	Events	Duration (s)	Samples
FNSZ	108	992	73,446	145,492
GNSZ	44	415	34,348	68,106
TCSZ	11	50	5,630	11,189
MYSZ	2	3	1,312	2,620
CPSZ	34	342	33,088	65,684
ABSZ	12	99	852	1,562
SPSZ	2	44	1,534	3,002
TNSZ	2	67	1,271	2,445

III. DATASET AND EXPERIMENTAL SETUP

In this section, we briefly discuss the statistics of the TUH Seizure Corpus V1.4.0 (Table I), the EEG segmentation procedure (Section III-A), and the experimental setup (Section III-B).

For the experiments on seizure classification with the proposed deep learning architecture, we use a subset of the Temple University Hospital EEG (TUH EEG) database [29], the TUH EEG Seizure Corpus (TUSZ) v1.4.0 [35] (this is, to the best of our knowledge, the only freely available dataset that contains seizure type annotations suitable for seizure classification). The dataset contains 2, 012 seizure events from different patients. Seizure events are assigned to one of 8 seizure classes: Focal Non-Specific Seizure (FNSZ), Generalized Non-Specific Seizure (GNSZ), Simple Partial Seizure (SPSZ), Complex Partial Seizure (CPSZ), Absence Seizure (ABSZ), Tonic Seizure (TNSZ), Tonic-Clonic Seizure (TCSZ) and Myonic Seizure (MYSZ). Statistics on the above seizure types including the number of seizure events, and total recording time (sum of the duration of each seizure event of a particular class rounded to the nearest minute) are given in Table I.

Since EEG recordings contain both seizure events and background signals, we used the following segmentation and processing techniques to obtain the dataset used in our experiments. Since the sampling rate and number of channels are not consistent among the recordings in TUSZ v1.4.0, the following configurations were considered to ensure uniformity. As the number of channels in the recordings varies from 24 to 36, 19 channels common to all recordings were selected, with the selected channels following the 10-20 layout of scalp EEG electrodes. Furthermore, the TUSZ samples contain EEG recordings at a different sampling rate (512 Hz rather than 250 Hz). As such, we re-sampled all EEG recordings with a seizure event to 250 Hz. Then, seizure events were extracted from the original EEG based on the time interval and the sampling rate (if the seizure event is present from T_1 to T_2, a segment from bit positions $T_1 \times S_{freq}$ to $T_2 \times S_{freq}$ was extracted, considering the bit representation of the EEG signal where S_{freq} refers to the sampling frequency). The re-sampling process is the only pre-processing step used by the proposed approach, and is only required to ensure consistency in the data.

A. Dataset Generation and Statistics

Following the state-of-the-art methods [15], [25], [28], we have segmented the seizure events into $1s$ segments (250 samples), and sampled using a $0.5s$ (125 samples) sliding window (segmentation is achieved with bit representations where $1s$ of EEG signal is represented by 250 samples at a sampling rate of 250 Hz). Therefore the shape of each segment is $(19, 250)$ where 19 and 250 refer to the number of channels and the length of each channel in samples. The number of samples obtained from a seizure event can be calculated using Equation 10,

$$N = \frac{F_{round}(T_n \times S_{freq} - T_s \times S_{freq})}{T_w \times S_{freq}} + 1, \tag{10}$$

where N, F_{round}, T_n, T_s and T_w refer to the number of sub-segments from the event, a function that rounds a given integer to the nearest multiplication of a given integer, the duration of the seizure event, the duration of the segment and the sliding window length respectively. The total number of segments for each seizure type is given in Table I.

B. Experimental Setup

Since the number of seizure events and the patient counts are heavily imbalanced (refer to MYSZ in Table I), a patient independent or seizure type-wise classification is inappropriate when all 8 seizure classes are considered. Therefore, we follow a 5-fold cross-validation method for evaluation. However, if the folds are obtained randomly from the segmented EEG events, each fold may have wildly different sample counts due to the significant differences in the duration of seizure events. Therefore, we segment the seizure events as described in Section III-A and then randomly divide them into 5 folds where the proportional class distribution of the dataset after splitting is approximately equal in each fold. Since all the segments are shuffled and randomly divided among the train and test sets [15], the temporal relationship among consecutive segments is ignored under the assumption that each segment contains sufficient data variation to represent the seizure. Thus, the temporal relationship among successive segments is not modelled in this work.

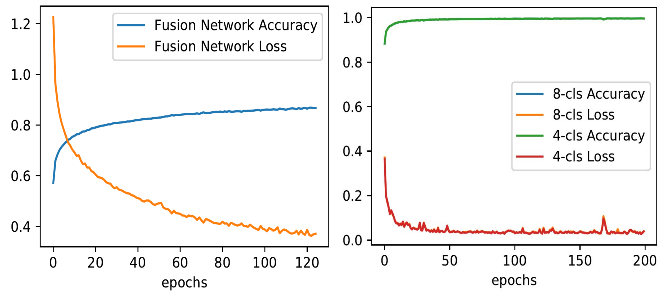


Fig. 5. Variation in training accuracy and loss during the training phase of the fusion network and the classification network. The 8-cls accuracy and the 4-cls accuracy in the right figure corresponds to the two-objective classification in the classification network (refer Section II-C).

First, channel-wise encoder networks are trained together with naive fusion followed by dense classification. With this, each sub-network is capable of learning the best features towards the objective while keeping the interaction with other channels minimal. Then, the naive fusion layer is replaced with the proposed attentive fusion and the network is fine-tuned with the new fusion layer (refer to Section II-B), followed by the dense classification to ensure that the fusion layer is capable of learning salient features for seizure classification from different EEG channels. This two stage training approach is used due to hardware and memory limitations. With more computational resources, the network can be trained end-to-end.

Since we have followed a two-stage training process due to hardware and memory limitations, the variation in accuracy and loss during the learning phase of the fusion network and the classification network are illustrated in Figure 5. Due to very high class imbalance in the dataset, we haven't used a validation split since it can further reduce the available data for training (see Table I). However, to obtain a fair evaluation, we have trained both the fusion network and the classification network to a fixed number of epochs on the training set and then evaluated on the testing set at the end of the training cycle.

All network models were trained with the Adam optimizer with a learning rate of 0.001 and a batch size of 80. The performance of the network is evaluated using class-wise precision, class-wise recall, class-wise F1-score, and weighted F1-score due to the heavily imbalanced dataset. The evaluation metrics are calculated using following equations where TP_i, FP_i, FN_i, TN_i and, N_i refer to true positives, false positives, false negatives true negatives and, number of samples for the seizure type i respectively [36].

$$precision_i = \frac{TP_i}{TP_i + FP_i} \tag{11}$$

$$recall_i = \frac{TP_i}{TP_i + FN_i} \tag{12}$$

$$specificity_i = \frac{TN_i}{TN_i + FP_i} \tag{13}$$

$$F1{-}score_i = \frac{2 \times precision_i \times recall_i}{precision_i + recall_i} \tag{14}$$

$$weighted\ F1{-}score = \frac{\sum_{i=0}^{n} (N_i \times F1{-}score_i)}{\sum_{i=0}^{n} N_i} \tag{15}$$

TABLE II
PERFORMANCE OF STATE OF THE ART METHODS FOR SEIZURE CLASSIFICATION ON TUH-TUSZ V1.4.0 CORPUS

Baseline Method	Network Input (Preprocessing Method)	Data Type	F1-score
SGD [24]	Correlation Coefficient Matrix from FFT	sEEG 20 montages	0.807
XGBoost [24]	Correlation Coefficient Matrix from FFT	sEEG 20 montages	0.851
KNN [24]	Correlation Coefficient Matrix from FFT	sEEG 20 montages	0.901
ResNet50 [24], [15]	FFT	sEEG 20 montages	0.723
AlexNet [37]	Concatenated STFT Spectrograms	sEEG 19 unipolar channels	84.1% (WA)
SeizureNet [14]	Saleincy Encoded Spectrograms	sEEG 20 montages	0.94
stSENet [38]	EEG and Spectrograms	sEEG 20 channels	0.937
Plastic NMN [15]	FFT	sEEG 20 montages	0.945
Ours	Raw EEG Signal (No Preprocessing)	sEEG 19 channels in 10-20 system	0.967

Class weights were used in training the initial encoder networks, where the weights were calculated as per Equation 16,

$$W_i = \begin{cases} \log \dfrac{\mu * \sum_{j=0}^{n} N_j}{N_i} & if \log \dfrac{\mu * \sum_{j=0}^{n} N_j}{N_i} > 1 \\ 1 & if \log \dfrac{\mu * \sum_{j=0}^{n} N_j}{N_i} < 1, \end{cases} \tag{16}$$

where W_i, μ, n, N_i refer to the weight of the class i, smoothing factor (set to 0.15), the total number of classes and number of samples in the training set for class i. This approach is followed to smooth the weights since if the weights are assigned based on the magnitude alone, it can adversely affect the learning of more common classes due to the very high class imbalance. Both training and testing were carried out on a computer with an NVidia M40 GPU, 30 GB of memory, and 6 CPU cores.

IV. Results and Discussion

A detailed analysis of the experimental results and a comparison with the state of the art methods is provided in this section. We have selected the most significant baseline approaches based on their performance (primarily F1-score), dataset used, and the proportional distribution of selected seizure events and the task (seizure classification only) to enable a fair comparison. Weighted F1-score is considered as the evaluation metric in the majority of the state-of-the-art methods due to the high class imbalance in the TUSZ corpus [15], [24], [25]. Furthermore, we have evaluated the seizure type-wise performance of the proposed model in terms of Precision and Recall (equivalent results are not available for other baseline systems in the literature, and hence are not reported here). Table II summarizes our experimental results along with selected state-of-the-art techniques.

It is observed that different approaches for seizure classification have used a different number of seizure classes and varied dataset configurations, with different sample distributions. This makes a direct comparison challenging. Furthermore, differences in the duration of seizure events has been observed in the literature, which results in different sample distributions in the dataset after any pre-processing and splitting into folds. Therefore, we primarily compare our work with [15], since the proportional distribution of samples among classes is approximately equal, even though [15] uses a 7 class configuration (making our task harder to solve).

We have achieved a 0.9664 ± 0.003 weighted F1-score with a weighted precision, weighted recall, weighted accuracy and weighted specificity of $96.65\% \pm 0.3$, $96.65\% \pm 0.3$, $96.65\% \pm 0.3$ and $98.08\% \pm 0.3$ respectively. Table III summarizes the seizure type-wise performance of our proposed method in terms of precision, recall, F1-score, accuracy and specificity. The lowest precision, recall and F1-scores are achieved for TNSZ while the highest is achieved for MYSZ. TNSZ being a specific subclass of GNSZ may be the reason behind the comparatively low performance, due to the confusion with the majority class. The reason for MYSZ achieving the highest recognition with an even lower sample count may be due to the richness of information and unique patterns in the EEG signals, which helps the network to learn a better decision boundary from other seizure types. It is observed that the proposed method is capable of obtaining consistent results over each seizure type, highlighting the reliability and robustness of the proposed approach. Due to this, we are able to achieve higher unweighted accuracy in our approach compared to state-of-the-art methods, most of which fail to obtain consistent results for all seizure types, and report low performance for at least for one seizure type. Furthermore, it is critical to identify minority seizure types (with fewer samples), since generally deep networks tend to focus on majority classes. Therefore, it is important for models to have a high recall for the minor classes, enabling the network to capture these rare events. Our method surpasses most state-of-the-art methods in terms of recall of the minority classes, achieving high recognition.

Figure 6 illustrates a confusion matrix for a selected test fold (all folds achieve similar results). Considering the minority classes in the dataset, ABSZ, TCSZ, and TNSZ have shown a slight confusion among themselves and GNSZ, while CPSZ and SPSZ show the highest confusion with FNSZ. The reason for this behavior would be that all the seizure types are not clinically disjoint as described in [15]. CPSZ and SPSZ are considered as specific subclasses of FNSZ; and ABSZ, TCSZ, and TNSZ are considered to be specific sub-classes of GNSZ; explaining the results we have obtained.

Roy *et al.* evaluated the applicability of different machine learning algorithms along with a single deep model (ResNet50) [23], [24] for seizure classification. Their results indicated that k-Nearest Neighbours (k-NN) had surpassed all the other methods including SGD, XGBoost, and ResNet50, where both SGD and ResNet50 have achieved significantly

TABLE III
AVERAGE SEIZURE-WISE PERFORMANCE OF THE PROPOSED METHOD

Seizure	Precision	Recall	F1-score	Accuracy	Specificity
FNSZ	96.8% ± 0.6	97.4% ± 0.3	0.971 ± 0.003	97.4% ± 0.3	97.00% ± 0.55
GNSZ	96.2% ± 0.4	95.1% ± 0.9	0.956 ± 0.004	95.1% ± 0.9	98.88% ± 0.13
TCSZ	96.8% ± 0.5	97.2% ± 0.5	0.970 ± 0.002	97.2% ± 0.5	99.88% ± 0.02
MYSZ	99.3% ± 0.6	98.6% ± 0.8	0.989 ± 0.007	98.6% ± 0.8	99.99% ± 0.01
CPSZ	96.6% ± 0.7	96.5% ± 0.6	0.965 ± 0.003	96.5% ± 0.6	99.04% ± 0.22
ABSZ	98.9% ± 0.8	94.6% ± 1.7	0.967 ± 0.007	94.6% ± 1.7	99.99% ± 0.0
SPSZ	97.5% ± 1.0	96.8% ± 1.1	0.971 ± 0.007	96.8% ± 1.1	99.97% ± 0.01
TNSZ	95.3% ± 1.6	95.1% ± 1.7	0.952 ± 0.012	95.1% ± 1.7	99.96% ± 0.01

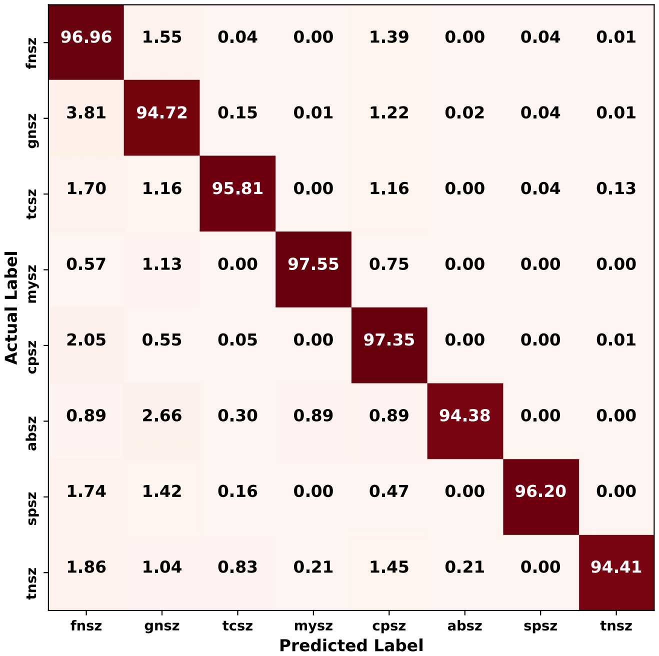


Fig. 6. Confusion matrix of the proposed approach for one fold (all folds achieve similar results). It is observed that all seizure classes have achieved comparatively similar recognition accuracies, even with the highly imbalanced dataset (refer Section III).

lower F1-scores. The main reason behind this behavior is the domain mismatch when using transfer learning with ResNet50. Since they re-train only the last layer of the original architecture (which was pre-trained on image-net), it may not be able to learn a meaningful representation from the FFT based features. However, we train our CNN networks from scratch (shallow networks make the training fast and the high sample count helps the model to generalize) for each channel, enabling the network to learn more meaningful feature representations.

Sriraam *et al.* proposed an AlexNet based approach for seizure classification with concatenated spectrograms as the input [37]. The results indicate that basic CNN architectures are able to obtain comparable (with AlexNet [21]) or higher (with VGG16 & VGG19 [22]) results compared to deep-nets using transfer learning, demonstrating the value of carefully designed CNN architectures over transfer learning when the domain mismatch is large. In contrast to both the above methods, we use raw EEG waves instead of engineered features (such as Discrete Fourier Transform (DFT) and STFT), and we have used carefully designed, interpretable deep networks to learn more robust features from the signals directly. Furthermore, the trade off between the frequency resolution and the time resolution of STFT can result in feature loss when EEG channels show different behaviors.

SeizureNet [14] utilizes a multi-spectral feature sampling-based ensemble CNN architecture to improve convergence and reduce over-fitting of deep networks. This has resulted in higher performance compared to previous CNN based methods. However, the complex architecture has resulted in a high parameter count and inference time compared to our proposed method. Furthermore, the pre-processing of the EEG samples (salience encoded spectrograms) takes considerable time, while our method achieves higher performance in terms of F1-score without any pre-processing. Furthermore, noisy signals may adversely affect the generated inputs while the attention used in our method is aimed at reducing this effect.

stSENet [28] utilized a Squeeze and Excitation network [39] based deep model for seizure type classification by using multi-level spectral and multi-scale temporal analysis. This method has been able to surpass most conventional machine learning based methods in-terms of recognition, however [15], [25] have achieved better performance through simpler networks. Furthermore, the method has achieved comparatively low performance in-terms of class-wise recognition accuracy for majority classes compared to our proposed method. The NMN approach of [15] surpassed all the above in performance. Both FNSZ and GNSZ achieved 100% recall, but the precision of the two classes remains low resulting in a reduced F1-score. When the minority classes are considered (such as ABSZ, TNSZ, and SPSZ), it is observed that the majority of them obtained a lower recall, even though they have obtained very high precision. Therefore, it is observed that the above method is not capable of achieving consistent results over every class in terms of precision, recall and F1-score. Furthermore, the overall unweighted accuracy of the system remains low compared to our model due to the very low classification accuracies for minority classes. The main reason behind this is the use of memory, where the memory tends to learn features for majority classes which helps to improve the overall performance (the contribution of minority classes towards weighted accuracy is low compared to the major classes). Although the hybrid models proposed in [28] have been able to surpass other baseline methods, they have not achieved consistent performance over every seizure class (e.g. ABSZ) due to the class imbalance in the dataset. Furthermore, the utilization of ConvLSTM2D and Conv3D layers operating over spectrograms limits the interpretability of the deep network, while increasing the model complexity. However, as illustrated in Figure 6, our model is capable of achieving comparable precision and recall over all classes, even with the high-class imbalance, resulting in a higher accuracy and F1-score compared to [15], [28]. Therefore, it is evident that our model is able to learn more informative and rich features.

An additional advantage of the proposed method is the interpretability of the model when compared to state-of-the-art methods. Since the input signal is $1D$, we need to apply a feature transformation to get a spectrogram or similar to feed the data to a $2D$ or $3D$ convolutional network, resulting in a changed data representation and changes in what is modelled by the network. Furthermore, applying CNNs ($2D$ convolutions) to spectrograms is fundamentally different to applying CNNs to images which they are originally designed for. $2D$ filters share weights across the dimensions of the image based on the assumption that an image carries the same meaning, regardless of the location of the patch being analysed at any given point. However, spectrograms have fundamentally different axes (time and frequency) where the horizontal and vertical location of any feature is vital in determining the meaning. Therefore, the probable spatial variance of spectrograms can reduce the interpretability of the deep network when using CNNs on spectrograms [26].

A $2D$ convolution operation over a spectrogram will respond to the same shape at any frequency while a filters in SincConv layer can select specific frequencies. To get similar behaviour of selecting just a small frequency band with either $2D$ or $3D$ kernels, we need to have a very high number of layers to obtain a sufficiently large receptive field, which can become harder to interpret. When considering Convolution+LSTM based approaches, similar issues in interpreting to the $2D$ convolutional layer arise; and to provide an input to the LSTM, the data (which is not truly a sequence following the convolution layer operations) must be mapped to a sequential form. As LSTMs themselves are hard to interpret, and the input is not necessarily a sequence, model interpretation is challenging compared to $1D$ convolution based approaches operating over the raw signal.

Due to the use of separate encoder networks for each channel, the approach can be easily extended when more channels need to be added to the network, each of which may have different characteristics. The parameter count and the inference time have a significant impact on the applicability of the proposed approach in real-world applications, especially in low resource environments. Since we are using separate shallow encoder networks ($\sim 29,000$ parameters per encoder) to learn features from an EEG channel, our proposed architecture contains approximately 10 million parameters and the inference time is approximately 15 ms. Furthermore, all the above discussed approaches require a pre-processing step which is not generally GPU accelerated, adding additional time component to the inference.

V. Ablation Studies

In this section, we present several ablation studies that have been carried out to identify the:

1) the filters learnt through SincNet;
2) representation of feature learnt by the deep network;
3) importance of using attention to learn salient features;
4) impact of dimension reduction in fusion and the value of using two objectives in classification; and
5) selection of filter length and size for encoder modules.

A. Filters Learnt Through SincConv Layer

Figure 7 illustrates the STFT of selected filters from the first layer (SincConv layer) of different sub-networks used for different channels. We have used 'librosa' [40] to generate spectrograms. A significant difference in the filter length has been observed across different filters, though most of have learned one prominent band pass filter. The filter banks from all the channels contain similar filter shapes to those shown in Figure 7, however the proportions, magnitudes, and number of prominent bands differ from channel to channel highlighting the importance of having different networks for each channel. Since our feature extraction networks for each channel are shallow, the trade-off between runtime and the importance of the subtle differences of each channel are minimized.

B. Representation of Deep-Net Features

Figure 8 presents a non-linear and non-deterministic dimensionality reduction visualization of high dimensional features from the last dense layer before the classification in phase 2 (refer to Section III-B) using t-distributed Stochastic Neighbor Embedding (t-SNE) with a perplexity of 40. The proposed approach for seizure type classification has been able to identify a clear decision boundary, even with the high dimensional data, for all the majority classes. However, several outlier points are still visible, especially near the decision boundaries of FNSZ, GNSZ, and CPSZ. It is observed that the proposed method is capable of learning a clear decision boundary for most minority classes in the corpus except for MYSZ. Furthermore, the dense clusters suggest that the proposed method is capable of learning a consistent feature representation from the input EEG channels. However, we do not observe a perfect decision boundary due to the challenging nature of the problem.

C. Importance of Using Attention to Learn Salient Features

In this study, we evaluate the impact of the attention module in the proposed fusion architectures on the classification performance. Since the initial training is carried out in two phases (training encoder networks with proposed fusion; followed by training the classification network with extracted features) due to hardware limitations, we have compared network performance at the end of both training phases to best illustrate the role of attention. Table IV presents performance after phase 1 of training with and without attentive fusion.

It is observed that the attentive fusion has achieved slightly higher precision, recall and F1-score compared to the naive fusion, indicating the importance of having attention to weight the channels and features, which helps the network to learn a better decision boundary. Furthermore, it is observed that applying attention (both $sigmoid$ and $softmax$ based) to the dense layers of the classification network results in a slight

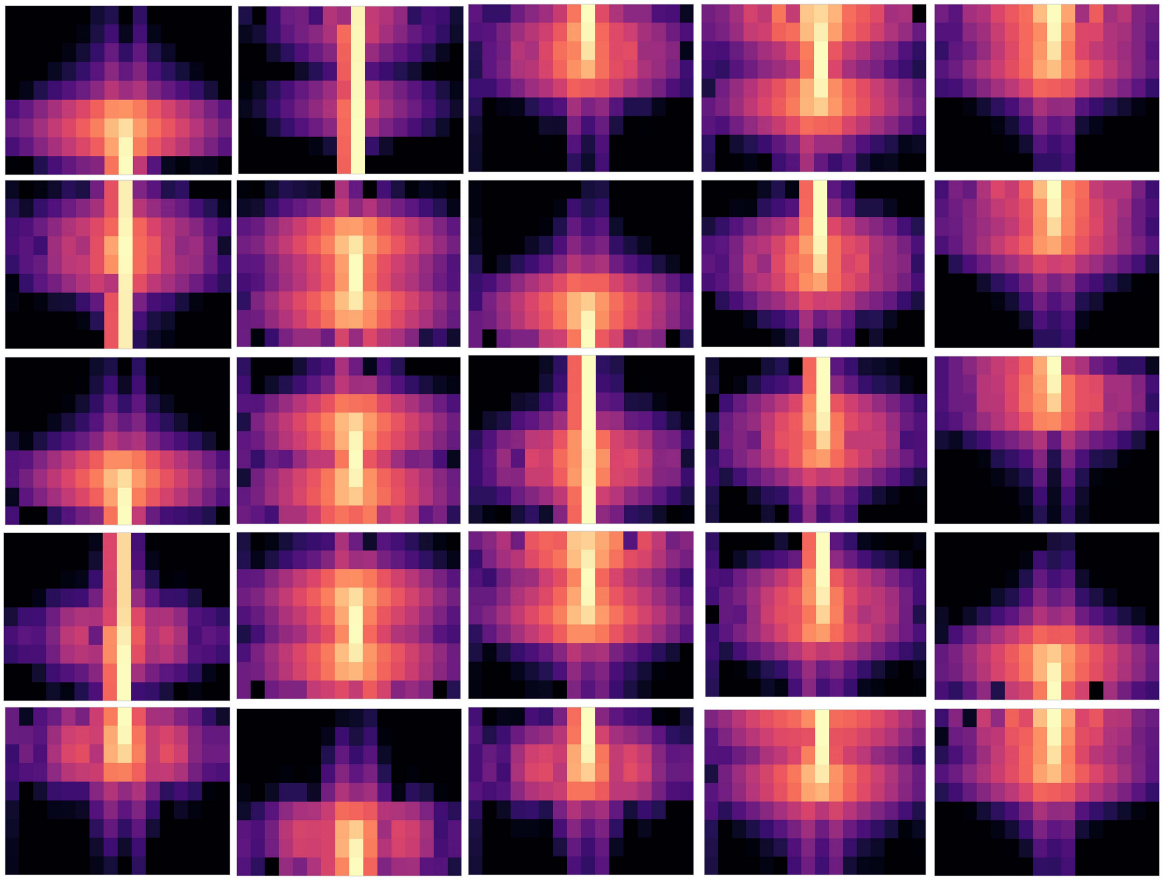


Fig. 7. Visualization of STFT of the first 5 filters (rows) from the first convolution layer (Sinc function based convolution) from the encoder networks corresponding to; FP1, O1, F8, P4 and CZ channels of 10-20 system (columns).

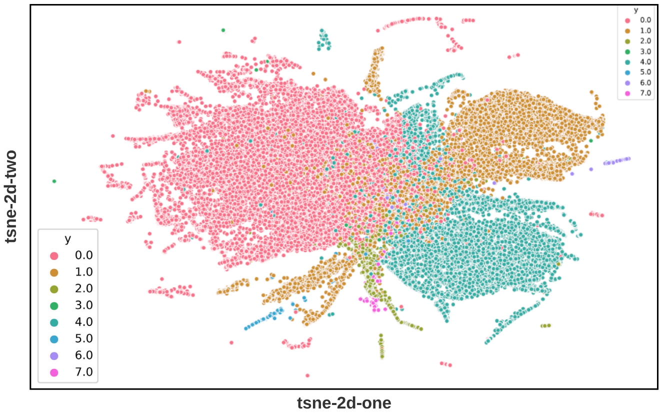


Fig. 8. t-stochastic Neighbour Embedding (t-SNE) plot obtained using the dense feature vector from the fully connected layer before the classification layer in the proposed network. The numbers in the legend correspond the seizure types presented in the Table I in the same order.

TABLE IV
AVERAGE PERFORMANCE OF THE NETWORK (PHASE 1) WHEN NAIVE FUSION AND THE PROPOSED ATTENTIVE FUSION IS USED

Fusion	Precision	Recall	F1-score
Attentive	89.2% ± 0.3	89.1% ± 0.3	0.891 ± 0.003
Naive	87.1% ± 0.4	86.9% ± 0.5	0.870 ± 0.004

TABLE V
PERFORMANCE VARIATION IN THE PURPOSED SYSTEM WITH AND WITHOUT ATTENTION IN THE CLASSIFIER NETWORK

Classifier	Precision	Recall	F1-score
Without Attn.	96.7% ± 0.3	96.7% ± 0.3	0.966 ± 0.003
With Attn.	96.0% ± 0.4	96.0% ± 0.5	0.960 ± 0.005

reduction in the overall performance of the network (refer to Table V). The main reason for this is the richness of the fused features which can be directly mapped to the objective, which is adversely affected by the additional constraints introduced applied through attention.

D. Impact of Two Objectives on Classification

In this section, we evaluate the importance of the proposed multi-task classification approach, as explained in Section II-C, by comparing the proposed approach to an equivalent model using a single 8-class classification objective. Due to the high number of samples for the majority classes, it is highly likely to have confusion between minority classes and majority classes, reducing the precision of the minority classes. Therefore in the proposed approach, we have considered all the minority classes as one seizure class and the majority classes as different classes. Due to this approach, class imbalance for the 4-class objective is reduced, resulting in adjustments in the weights of the last dense layer (Dense 2 in Figure 3) to learn a better decision boundary among the majority and minority classes. The performance of the above approaches (with and without the

TABLE VI
PERFORMANCE OF THE PURPOSED SYSTEM WITH/WITHOUT TWO OBJECTIVE CLASSIFICATION

Classification	Precision	Recall	F1-score
Two Objectives	96.7% ± 0.3	96.7% ± 0.3	0.966 ± 0.003
Single Objective	96.2% ± 0.2	96.2% ± 0.3	0.962 ± 0.003

TABLE VII
DIFFERENT PARAMETER CONFIGURATIONS FOR THE SINCNET MODEL AND THE CORRESPONDING RESULTS IN LEARNING STAGE 1 WITH NAIVE FUSION. n_{f1}, l_{f1}, n_{f2}, l_{f2}, n_{f3} AND, l_{f3} REFER TO THE NUMBER OF FILTERS (n) AND FILTER LENGTH (l) IN *SincConv* LAYER ($f1$), FIRST $Conv-1D$ LAYER ($f2$) AND, SECOND $Conv-1D$ LAYER ($f3$) RESPECTIVELY

Config.	n_{f1}	n_{f2}	n_{f3}	l_{f1}	l_{f2}	l_{f3}	Acc.
1	80	40	40	127	5	5	87.17%
2	20	20	20	129	32	8	86.98%
3	40	40	40	64	16	8	85.28%
4	10	10	10	127	16	8	82.90%
5	20	20	20	127	5	5	82.12%
6	5	5	5	127	5	5	74.69%
7	40	40	40	127	5	5	83.88%

two objective classification) is given in Table VI. The results indicate that comparatively higher results can be obtained by utilizing two classification objectives which helps to learn a better decision boundary between majority and minority classes.

E. Parameter Selection for Encoder Networks (SincNet Modules)

In this section, we discuss the basis for the selection of the main parameters of the SincNet module which we have used as the encoder for each channel. We consider the following SincNet module parameters: n_{f1}, the number of filters in SincConv layer, l_{f1}, the filter length of the SincConv layer, n_{f2}, the number of filters in the first Conv-1D layer, l_{f2}, the filter length of the first Conv-1D layer, n_{f3}, the number of filters in the second Conv-1D layer and, l_{f3}, the filter length of the second Conv-1D layer (refer to Figure 2 for the diagram of the encoder) and the results are illustrated in Table VII. We have followed the original SincNet paper [31] when obtaining the possible parameters and the parameters were set to $n_{f1} = 80$, $n_{f2} = 60$, $n_{f3} = 60$, $l_{f1} = 251$, $l_{f2} = 5$ and, $l_{f3} = 5$ in the original implementation [31]. However, since our sampling rate is $250Hz$ and the segment length is $1s$, we have decided to reduce l_{f1}. Since it is nearly impossible to find the globally optimal parameters for a deep network due to the higher number of parameters and long training time, we have followed an experimental trial and error method. We have evaluated our encoder network with a selection of what we consider to be sensible configurations based on our problem's characteristics, and have tested these on one fold and selected the best parameter configuration (80, 40, 40, 127, 5 and 5) for our full model. We have kept all other parameters of the SincNet model similar to the original paper and all the channel encoders share the same parameter configurations.

VI. CONCLUSION

In this paper, we propose a novel deep learning architecture for epileptic seizure classification from raw EEG data. Compared to most state-of-the-art techniques that rely on hand-crafted inputs such as spectrograms and STFT, we use the raw EEG signals and deep learning to identify and extract relevant information which also aids interpretability. Since each EEG channel can contain information relevant to the final objective, we use all channels in the 10-20 system as input, where independent networks of identical architecture are learnt for each channel. Since the network can grow significantly with the number of channels used, we make each network compact such that the method is applicable in low resource environments. Furthermore, as the richness of the features extracted by the uni-channel encoders varies depending on the quality of the signal, we use an attentive fusion method to extract salient information for classification.

Experiments on the largest publicly available seizure dataset, TUH-TUSZ v1.4.0, demonstrates that the proposed method is capable of achieving significant performance improvements over methods with hand-crafted inputs. The SincNet based channel specific encoders help the network to learn domain and channel specific features, while the attentive fusion helps to learn salient channel-wise and temporal features. Furthermore, 1D convolutions and parametrized sinc function-based convolutions improves the overall model interpretability. Due to the low number of parameters and fast inference time, the proposed approach can be used in real-world applications and in low-resource environments. While noisy signals from EEG channels can adversely affect the recognition task increasing false positives, the carefully designed attention over temporal features and channels helps to compensate. Furthermore, this approach can be easily extended to incorporate more EEG channels with different sampling frequencies, obtained through different acquisition hardware.

REFERENCES

[1] W. Penfield and T. C. Erickson. (1941). *Epilepsy and Cerebral Localization: A Study of the Mechanism, Treatment and Prevention of Epileptic Seizures*. [Online]. Available: https://books.google.com.au/books?id=aqVrAAAAMAAJ

[2] J. S. Duncan, J. W. Sander, S. M. Sisodiya, and M. C. Walker, "Adult epilepsy," *Lancet*, vol. 367, no. 9516, pp. 1087–1100, Apr. 2006.

[3] T. Siddharth, P. Gajbhiye, R. K. Tripathy, and R. B. Pachori, "EEG-based detection of focal seizure area using FBSE-EWT rhythm and SAE-SVM network," *IEEE Sensors J.*, vol. 20, no. 19, pp. 11421–11428, Oct. 2020.

[4] R. S. Fisher *et al.*, "Operational classification of seizure types by the international league against epilepsy: Position paper of the ILAE commission for classification and terminology," *Epilepsia*, vol. 58, no. 4, pp. 522–530, Apr. 2017.

[5] M. M. Goldenberg, "Overview of drugs used for epilepsy and seizures: Etiology, diagnosis, and treatment," *Pharmacy Therapeutics*, vol. 35, no. 7, p. 392, 2010.

[6] T. U. Syed *et al.*, "Can semiology predict psychogenic nonepileptic seizures? A prospective study," *Ann. Neurol.*, vol. 69, no. 6, pp. 997–1004, Jun. 2011.

[7] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinf.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[8] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, 2019.

[9] Y. Zhang, J. Wu, Y. Liu, Y. Chen, E. X. Wu, and X. Tang, "MI-UNet: Multi-inputs UNet incorporating brain parcellation for stroke lesion segmentation from T1-weighted magnetic resonance images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 526–535, Feb. 2021.

[10] R. Zhang, X. Xiao, Z. Liu, Y. Li, and S. Li, "MRLN: Multi-task relational learning network for MRI vertebral localization, identification, and segmentation," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2902–2911, Oct. 2020.

[11] C.-N. Jiao, Y.-L. Gao, N. Yu, J.-X. Liu, and L.-Y. Qi, "Hyper-graph regularized constrained NMF for selecting differentially expressed genes and tumor classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 3002–3011, Oct. 2020.

[12] Z. Jiang and W. Zhao, "Optimal selection of customized features for implementing seizure detection in wearable electroencephalography sensor," *IEEE Sensors J.*, vol. 20, no. 21, pp. 12941–12949, Nov. 2020.

[13] M. Radman, M. Moradi, A. Chaibakhsh, M. Kordestani, and M. Saif, "Multi-feature fusion approach for epileptic seizure detection from EEG signals," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3533–3543, Feb. 2021.

[14] U. Asif, S. Roy, J. Tang, and S. Harrer, "SeizureNet: Multi-spectral deep feature learning for seizure type classification," 2019, *arXiv:1903.03232*. [Online]. Available: http://arxiv.org/abs/1903.03232

[15] D. Ahmedt-Aristizabal, T. Fernando, S. Denman, L. Petersson, M. J. Aburn, and C. Fookes, "Neural memory networks for seizure type classification," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 569–575.

[16] S. P. Shashikumar, A. J. Shah, Q. Li, G. D. Clifford, and S. Nemati, "A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, 2017, pp. 141–144.

[17] D. Biswas *et al.*, "CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 2, pp. 282–291, Apr. 2019.

[18] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for patient-independent epileptic seizure prediction using scalp EEG signals," *IEEE Sensors J.*, vol. 21, no. 7, pp. 9377–9388, Apr. 2021.

[19] H. Khan, L. Marcuse, M. Fields, K. Swann, and B. Yener, "Focal onset seizure prediction using convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 2109–2118, Sep. 2018.

[20] Y. Li, X.-D. Wang, M.-L. Luo, K. Li, X.-F. Yang, and Q. Guo, "Epileptic seizure classification of EEGs using time–frequency analysis based multiscale radial basis functions," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 2, pp. 386–397, Mar. 2018.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] S. Roy, U. Asif, J. Tang, and S. Harrer, "Seizure type classification using EEG signals and machine learning: Setting a benchmark," 2019, *arXiv:1902.01012*. [Online]. Available: http://arxiv.org/abs/1902.01012

[25] U. Asif, S. Roy, J. Tang, and S. Harrer, "SeizureNet: Multi-spectral deep feature learning for seizure type classification," in *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology*. Cham, Switzerland: Springer, 2020, pp. 77–87.

[26] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," in *Proc. 1st Int. Conf. Deep Learn. Music*, 2017, pp. 37–41.

[27] D. Zeng, K. Huang, C. Xu, H. Shen, and Z. Chen, "Hierarchy graph convolution network and tree classification for epileptic detection on electroencephalography signals," *IEEE Trans. Cognit. Develop. Syst.*, early access, Jul. 27, 2020, doi:10.1109/TCDS.2020.3012278.

[28] T. Liu, N. D. Truong, A. Nikpour, L. Zhou, and O. Kavehei, "Epileptic seizure classification with symmetric and hybrid bilinear models," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2844–2851, Oct. 2020.

[29] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers Neurosci.*, vol. 10, p. 196, May 2016.

[30] A. H. Shoeb and J. V. Guttag, "Application of machine learning to epileptic seizure detection," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 975–982.

[31] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 1021–1028.

[32] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Memory based fusion for multi-modal deep learning," *Inf. Fusion*, vol. 67, pp. 136–146, Mar. 2021.

[33] D. Priyasad, T. Fernando, S. Denman, C. Fookes, and S. Sridharan, "Attention driven fusion for multi-modal emotion recognition," 2020, *arXiv:2009.10991*. [Online]. Available: http://arxiv.org/abs/2009.10991

[34] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.

[35] V. Shah *et al.*, "The temple university hospital seizure detection corpus," *Frontiers Neuroinform.*, vol. 12, p. 83, Nov. 2018.

[36] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[37] Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "A convolutional neural network based framework for classification of seizure types," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2547–2550.

[38] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, "Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 782–794, Apr. 2020.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[40] B. McFee *et al.*, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, vol. 8, 2015, pp. 18–25.

Darshana Priyasad received the B.Sc. (Hons.) degree in engineering, specialized in integrated computer engineering from the University of Moratuwa, Moratuwa, Sri Lanka. He is currently pursuing the Ph.D. degree with the Queensland University of Technology, Brisbane, QLD, Australia. His research interests include deep learning, and computer and machine vision.

Tharindu Fernando (Member, IEEE) received the B.Sc. (special degree in computer science) degree from the University of Peradeniya, Peradeniya, Sri Lanka, and the Ph.D. degree from the Queensland University of Technology (QUT), Brisbane, QLD, Australia. He is currently a Postdoctoral Research Fellow with the SAIVT Research Program, School Electrical Engineering and Computer Science, QUT. His research interests focus mainly on human behavior analysis and prediction.

Simon Denman (Member, IEEE) received the B.Eng. degree in electrical engineering from BIT and the Ph.D. degree in the area of object tracking from the Queensland University of Technology (QUT), Brisbane, QLD, Australia. He is currently a Senior Research Fellow with the Signal Processing, Artificial Intelligence and Video Technologies, QUT. His research interests include intelligent surveillance, video analytics, and video-based recognition.

Sridha Sridharan (Life Senior Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in communication engineering from The University of Manchester, Manchester, U.K., and the Ph.D. degree from the University of New South Wales, Sydney, NSW, Australia. He is currently with the School Electrical Engineering and Robotics, Queensland University of Technology (QUT), Brisbane, QLD, Australia, as a Professor. He is also the Leader of the Research Program in Signal Processing, Artificial Intelligence and Video Technologies, QUT, with a strong focus in the areas of computer vision, pattern recognition, and machine learning. He has authored or coauthored more than 600 articles consisting of publications in journals and more than 600 papers consisting of publications in refereed international conferences in the areas of image and speech technologies from 1990 to 2021 and during this period, he has graduated 75 Ph.D. students in the areas of image and speech technologies. Several of his research outcomes have been commercialized. He has received a number of research grants from various funding bodies, including commonwealth competitive funding schemes, such as the Australian Research Council and the National Security Science and Technology Unit.

Clinton Fookes (Senior Member, IEEE) received the B.Eng. degree in aerospace/avionics and the M.B.A. and Ph.D. degrees from the Queensland University of Technology (QUT), Brisbane, QLD, Australia. He is currently a Professor and the Head of the Discipline for Vision and Signal Processing, Faculty of Engineering, QUT. His research interests include computer vision, machine learning, and pattern recognition areas. He is also a Senior Member of the Australian Institute of Policy and Science Young Tall Poppy, a Senior Fulbright Scholar, and the Winner of the Australian Museum Eureka Prize.