A Convolutional Long Short-Term Memory Based Neural Network for Epilepsy Detection from EEG

Md. Nurul Ahad Tawhid, Siuly Siuly *Member, IEEE*, Tianning Li

***Abstract*—Epilepsy is a severe neurological disorder characterized by recurrent seizures, which increases the risk of death three times more than normal. Currently Electroencephalography (EEG) has emerged as a highly promising technique for the diagnosis of epilepsy. The majority of current EEG-based epilepsy detection research have employed a variety of deep learning-based models, but most of the approaches suffer from poor generalizability, optimal design and performance rates. To address these issues, this study aims to develop an efficient framework based on deep spatio-temporal neural network called convolutional long short-term memory (ConvLSTM) for epilepsy detection from EEG signals. In the proposed model, firstly standard 19 channel EEG data are selected and resampled at 256Hz, and then those signals are segmented into three-second time frames. Afterward, the segmented data are fed as input to the ConvLSTM model for identifying epileptic patients from normal subjects. To generalize the proposed model, we have tested it on two different datasets with varying population sizes. We have used the five-fold cross validation and leave-one-out cross validation schemes to eliminate the experiment's biases. To further validate the proposed framework, we have carried out various ablation studies. The experimental results demonstrate that the proposed model outperforms the current state-of-the-art results for the studied datasets, making it suitable for use as an automated system for the diagnosis of epilepsy.**

***Index Terms*—Epilepsy, ConvLSTM, Deep Learning, EEG**

I. Introduction

EPILEPSY is a chronic noncommunicable common neurological disease that has affected 50 million people of all ages worldwide [1]. It causes recurrent seizures, lack of consciousness and loss of control of bladder or bowel function which may lead to severe physical injury and death. Although the risk of premature death with epilepsy is three times higher than the normal population, 70% of the people with epilepsy (EP) could live normally if properly diagnosed and treated [1]. To do so, we need to develop an automatic system for early detection and prediction of epilepsy. Monitoring brain activity is one way to detect and predict the abnormality in the human brain [2]–[5]. Electroencephalogram (EEG) is a tool to capture the electrical activity of the brain by placing electrodes on the skull surface and it is one of the most popular tool due to non-invasiveness, fairly economic, and wide availability for clinicians [6]. This EEG signals are visually analyzed by the expert clinicians for patterns related to epilepsy, but this process is time-consuming, repetitive, costly, subjective and error-prone [7]. Additionally, EEG signal pattern for epilepsy is quite diverse, and it can vary significantly between patients and over time for the same patient. Moreover, three quarters of the people with epilepsy from developing countries have very little access to medical treatment due to its high cost and less availability [1], [2].

For those constraints, several researchers have developed automated systems to detect epilepsy [8]. Although those approaches achieved a higher level of accuracy [8], yet they have several drawbacks. First of all, most of these methods require manual feature extraction, which is complex and hard to verify as it requires designing new features [9]–[11]. Secondly, these EEG signals contains artifacts like eye blinks, swallowing and muscle activity which changes the extracted features and also affects the classification performance [12]. Thirdly, most of the automatic epilepsy detection systems are tested on small dataset, which has a negative impact on the performance and robustness of those systems [13]. Fourthly, majority of the researches have verified their model using a single dataset, which lefts the question on the model's generalization. Therefore, in this study we have developed a deep learning (DL) based framework for epilepsy detection using EEG signal and tested on multiple datasets.

We have addressed the mentioned issues by using DL techniques and validated the proposed model on two different epilepsy datasets of variant sizes. DL technique automatically extracts features from the EEG signal and performs classification tasks by learning from those extracted features, which removes the requirements of extracting complex hand-crafted features [9]. DL techniques adopts data and automatically discovers a hierarchy of features [2]. Although DL methods achieved promising outcomes in epilepsy detection in several studies [11], yet several enhancements can be accomplished by using different DL models. Among those DL based studies, many studies have used convolutional neural networks (CNN), like Gomes *et al.* [14] proposed a seizure detection system using CNN on the image representation of EEG signals. They used overlapping window segments to generate image and data

This work was funded by 2021 University of Southern Queensland (USQ) Faculty Collaboration Grant. *Corresponding author: Md. Nurul Ahad Tawhid; Tianning Li*

Md. Nurul Ahad Tawhid and Siuly Siuly are with the Institute for Sustainable Industries & Liveable Cities, Victoria University, Footscray, VIC 3011, Australia (e-mail: md.tawhid1@live.vu.edu.au; siuly.siuly@vu.edu.au).

Tianning Li is with the School of Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia (e-mail: Tianning.Li@usq.edu.au).

augmentation to train their CNN model and tested the model on two epilepsy datasets with promising results. Xiong *et al.* [15] also used the similar approach as [14] by representing the EEG signal into 2D image using fast Fourier transform (FFT) on sliced signal and then EfficientNet neural network was used to perform the classification of those 2D images. These approaches looked for the same pattern across EEG samples from various patients without considering temporal patterns in the EEG signal. Moreover, the designed network contains very complex and deep architecture to capture the necessary information related to epilepsy. This increased network size requires more parameters to train using a small dataset that may increase the probability of over fitting [2]. Additionally, the CNN model has a disadvantage of the inability to extract temporal information from EEG signal. To use the temporal information of the EEG signals, the long short-term memory (LSTM) model was used for epilepsy detection [16], [17]. To use the feature extraction capability of the CNN model in a noisy environment and temporal feature extraction capability of the LSTM model, Hussain *et al.* [2] used hybrid one dimensional CNN-LSTM model for epileptic seizure detection where the first layer of the model is a CNN layer followed by a maxpooling layer and then a LSTM layer. This model requires the input data to be converted into a single dimension and pre-processed using discrete wavelet transform (DWT) and FFT for making the data input ready.

In this study, we have used a two-dimensional (2D) convolutional long short-term memory (ConvLSTM) model to perform the classification of 2D EEG data for epilepsy detection. The hybrid CNN-LSTM model of [2] used CNN and LSTM in two different layers. In contrary, ConvLSTM combines CNN and LSTM in a single layer to extract spatiotemporal information from time-series data, which is why we have used this model for EEG data classification. Previously, Yang *et al.* [18] have used ConvLSTM for seizure recognition, but they have pre-processed the EEG data and converted them into image before feeding those into their model which required extra pre-processing time and manual feature extraction overhead. To reduce this extra pre-processing complexity and manual feature extraction overhead, we have used the raw EEG data as input in our proposed model. We have tested our proposed model on two publicly available datasets to validate its dataset independence, as well as the population size of the dataset. A five-fold cross-validation technique is used to validate the model's performance on the full dataset. Also, we have tested the leave-one-out cross-validation (LOOCV) to check the model's performance for each subject in the dataset without using it in the model training.

Below are the lists of this study's most significant contributions:

1) Design and validate a new efficient and automatic ConvLSTM based framework for epilepsy classification.
2) Explore the performance of the proposed model on two epilepsy datasets of variant population sizes.
3) Improve the classification accuracy compared to existing methods on the same dataset.
4) Explore the performance of the proposed model by using both five-fold cross validation and also the leave-one-out cross-validation scheme.
5) Validate the proposed model using different ablation studies and data augmentation techniques.

The remainder of this article is organized as follows. Section II gives details of the method used in this study. Our experimental results are described in Section III. Finally, Section IV gives the conclusions to this study.

II. Methodology

In this proposed system, we have developed a deep learning-based epilepsy classification system. To do so, we first collected the raw EEG data from publicly available data sources. Then those data are pre-processed to make input ready for the deep learning model. After that those signals are segmented into small time frames and fed into the proposed ConvLSTM based deep learning model. Finally, the classification performance of the system is measured using different performance evaluation matrices. An overview of the proposed system is given in Fig 1. Details of those steps are discussed in the below subsections.

A. EEG data collection

We have used two publicly available EEG datasets for validating the proposed system. The first dataset we have used is from Universidade Federal do Para, Brazil (UFP) [19]. It contains resting-state EEG data from 14 subjects. Seven of those 14 subjects are diagnosed with epilepsy (4 males, 3 females with an average age of 39.5±6.4 years and 24±7 years, respectively) and the remaining subjects are healthy controls (HC) with similar age and sex groups. EEG data is collected from standard 20 channels settings at a sampling frequency of 256Hz.

The second dataset we have used in this study is from the Temple University Hospital (TUH) EEG Epilepsy Corpus (TUEP) [20]. It is a small part of a large clinical EEG signal collection of 30,000 subjects recorded at TUH, Philadelphia. The TUEP dataset contains around 1500 sessions of different duration from 187 subjects (88 patients and 99 HC subjects, age range from 17 to 88 years). EEG signals are recorded using at least 19 electrodes from standard 10-20 montage [21]. Most of the recordings are sampled at a frequency of 256Hz. Demographic information of those datasets are given in Table I.

TABLE I: Demographic data of the used datasets.

Dataset	Subjects / (EP/HC)	Channel	Sampling Frequency	Segments (EP/HC)
UFP	14(7/7)	20	256	1248/1235
TUEP	187(88/99)	19-21	250/256/400	43978/39770

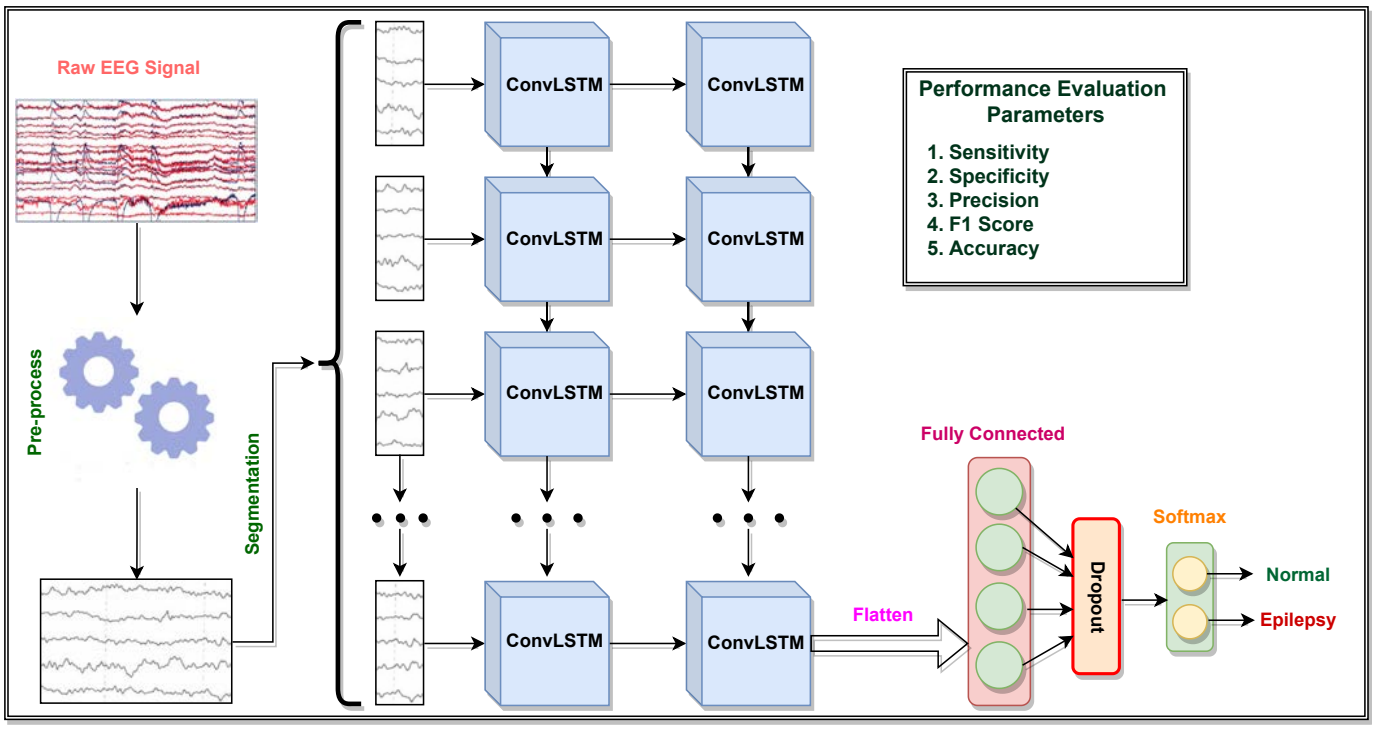


Fig. 1: Schematic illustration of the proposed system and steps involved in the analysis.

B. Pre-processing the data for input ready

In this step, we pre-processed the EEG data to make those signals input ready for the deep learning model. As we see from Table I, both the datasets have a different number of recording channels and sampling frequency. So, to classify them using a single deep learning model, we need to bring those two datasets into a common standard. To do so, at first we kept the data from 19 channels (Fp1, Fp2, F3, F4, F7, F8, C3, C4, T7, T8, P3, P4, P7, P8, O1, O2, FZ, CZ, PZ) of standard 10-20 electrode montage [21]. After that, to mitigate signal sampling frequency differences, we resampled the signals of the second dataset to 256Hz so that both the dataset have the same sampling frequency.

C. EEG signal segmentation

Data shortage is a major issue in the analysis of EEG data using deep learning models. To mitigate this issue, several authors have used the data segmentation technique in this field of study [5], [7], [22]. In this approach, EEG recordings are segmented into small time frames and labelled with the same label as the original. This makes a boost in the sample size of the dataset. In this study, we have split the EEG recording of each subject into three seconds (3s) time frame to capture the representative features from those small segments, as our previous study showed that this segment size is enough for EEG classification [22]. This segmentation process creates EEG signals of size 19x768 (19 channels x 256 samples/second x 3 seconds).

D. Feature extraction and classification using deep learning

To extract features from the EEG data and perform classification, we have used deep learning-based model named Convolutional Long Short-Term Memory (ConvLSTM). It was introduced to deal with precipitation nowcasting [23] by combining convolutional neural network (CNN) and long short-term memory (LSTM). It is capable of extracting spatiotemporal information from time-series data, which makes it suitable for EEG signal classification. It is a recurrent layer like the LSTM, except its internal matrix multiplication operations are replaced by convolution operations. As a consequence, rather than being a 1D vector containing features, the data that flows through the ConvLSTM cells retains the input dimension (3D in our instance) and preserve all the spatial information. The ConvLSTM uses a convolution operator in state-to-state and input-to-state transitions to determine the future state of a specific cell in the grid based on the inputs and past states of its local neighbours [23]. The key equation of the internal operation of ConvLSTM cell in given in 1:

$$\begin{aligned} li_t &= \sigma(W_{xi} * \chi_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\ C_t &= f_t \circ C_{t-1} + i_t \circ tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + Wco \circ C_t + b_o) \\ H_t &= o_t \circ tanh(C_t) \end{aligned} \quad (1)$$

Here, χ_1, χ_2, ..., χ_t are cell inputs, C_1, C_2, ..., C_t are cell outputs, H_1, H_2, ..., H_t are hidden states, i_t, f_t, o_t

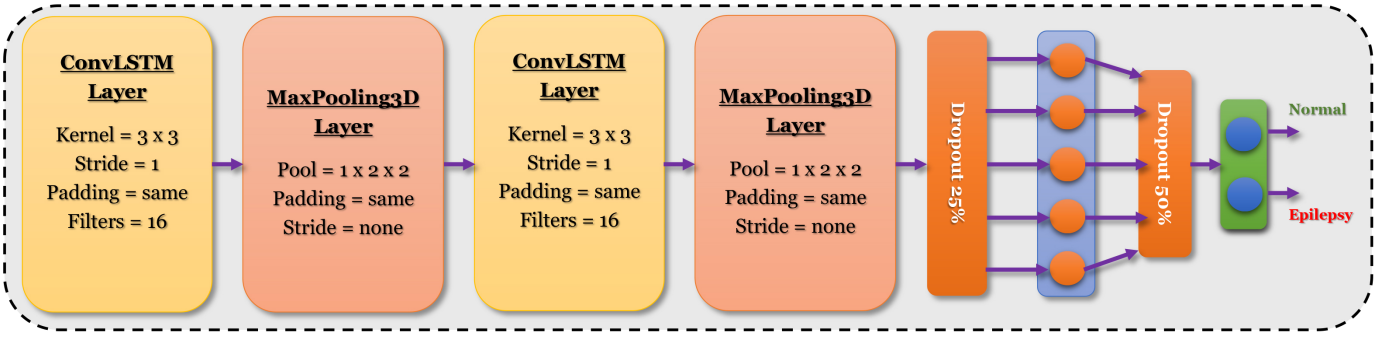


Fig. 2: Schematic diagram of the proposed ConvLSTM model.

are 3D tensor gates, and '*' and ∘ denotes the convolution operation and Hadamard product, respectively.

The proposed model consists of two ConvLSTM blocks each followed by a 3D max-poling layer. Each ConvLSTM layer has 16 filters with a kernel size of 3x3 and a stride of 1x1. Max-poling layers have a pool size of 1x2x2 with the same padding. The last ConvLSTM block is followed by a dropout layer of 25%, which is followed by a fully connected layer and a dropout layer of 50%. For ConvLSTM layers, we have used hyperbolic tangent function (tanh) as activation function and rectifier or ReLU as recurrent activation function. The final activation layer uses a softmax activation function to activate one of two outputs i.e normal or epilepsy. To compile the model, categorical cross entropy loss function is used with adam optimizer. A detailed structure of the proposed model is given in Fig 2.

E. Performance evaluation criteria

To evaluate the performance of the proposed framework on two EEG datasets, we have used five parameters that are used in this field of study namely, sensitivity (Sen), specificity (Spec), precision (Prec), F1 score (F1), and accuracy (Acc). These criteria allow to predict the behavior of the classifiers on the tested data [24]–[28].

Moreover, to mitigate the biases of the prediction result and reduce the over-fitting problem, we have used five-fold cross-validation technique. In this process, the dataset is divided into five equal or nearly equal parts and four of them are used to train the model and the rest of the fold is used for testing. This process is repeated five times so that each of the EEG segments is tested once [5].

To further evaluate the model's performance, we have also tested the model using leave-one-out cross-validation (LOOCV) technique. In this process, all the segments from a subject are left out from training phase and the trained model is used to perform the prediction on the left-out segments [29]. This process is repeated on all the subjects in the dataset. Although it is a computationally intensive method, it yields an accurate and unbiased measure of model performance [29].

III. Results and discussion

This section starts with a detail experimental setup for the proposed system and then the results of the experiments are discussed in the later subsection.

A. Experimental setup

As stated in the methodology section, after preprocessing, the signal data are segmented into 3s segments. This process produces a total of 1248 and 43978 segments for patients from dataset UFP and TUEP, respectively. While for healthy subjects, those numbers are 1235 and 39770 for UFP and TUEP, respectively. After segmentation, these segments are arbitrarily divided into five equal or nearly equal sub parts to perform the 5-fold cross-validation for the proposed DL model. The experimental model is trained with four subparts and the rest is used to validate the model. This process is repeated for five times so that each sub part is used for validating the model exactly once. Results of this 5-fold cross-validation process show the overall performance of the model on the full dataset as well as reduce the overfitting and biasing result problems. The model is trained with 50 epochs for UFP dataset as they get overfitting after that, while for TUEP dataset, we have used 500 epochs to train it as it is a large dataset. We have used mini-batch mode for batch size selection, which is popular for faster learning [5]. We have tested four different batch sizes (32, 64, 128 and 256) during the training process of the model.

B. Results

In this study, we have evaluated the proposed model using a 5-fold cross-validation technique. Experiments are carried out with different batch sizes to check the effect of batch size on the proposed model. In this study, four batch sizes are evaluated: 32, 64, 128 and 256, and their results are compared. For evaluation of the proposed model, five evaluation parameters (Sen, Spec, Prec, F1 and Acc) are calculated and reported in Table II for the tested two datasets. Fold wise average values are reported in bold face for different batch sizes.

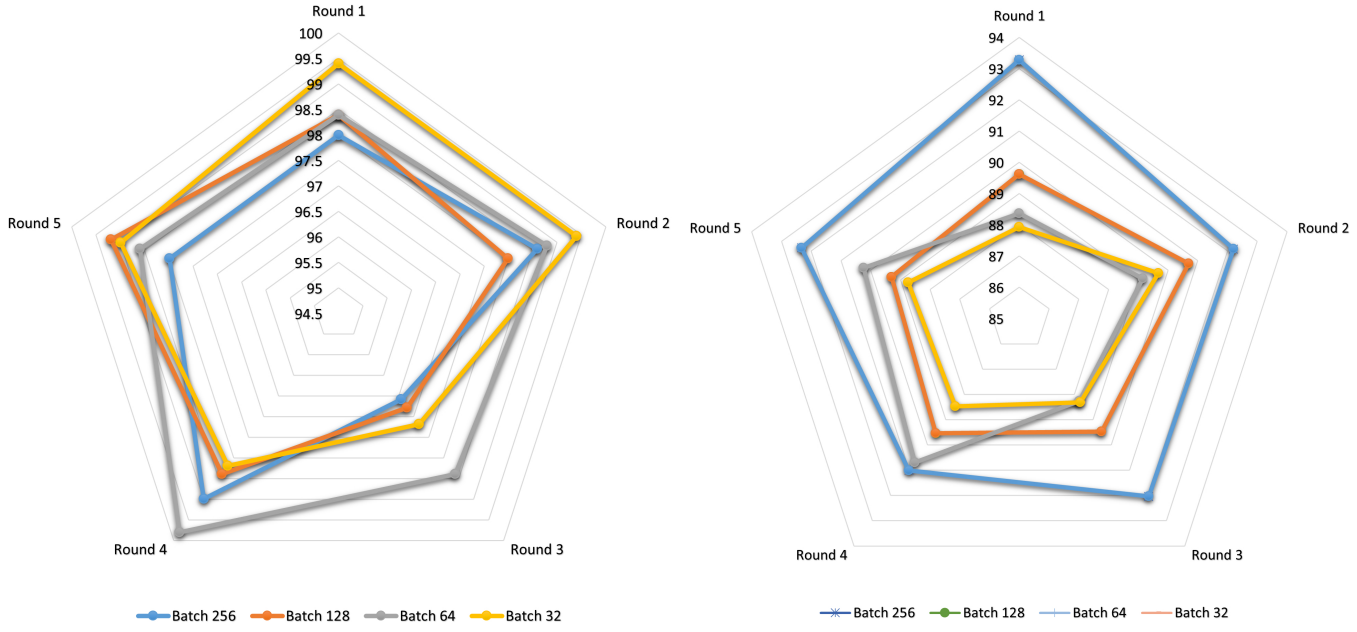
From Table II, we can see that for UFP dataset, average accuracy decreases with the increase of the batch size, while for TUEP dataset, accuracy increases with the

TABLE II: Detailed results of the proposed model on two datasets for different batch sizes.

Batch size		UFP Dataset					TUEP Dataset				
	Rounds	Sensitivity	Specificity	Precision	F1	Accuracy	Sensitivity	Specificity	Precision	F1	Accuracy
256	R1	97.61	98.37	98.39	0.98	97.99	93.98	92.50	93.32	0.94	93.28
	R2	98.41	98.78	98.80	0.99	98.59	93.63	90.63	91.65	0.93	92.20
	R3	95.36	97.69	97.41	0.96	96.58	92.88	91.11	92.06	0.92	92.04
	R4	99.19	98.80	98.79	0.99	98.99	92.22	89.71	90.66	0.91	91.01
	R5	97.98	97.98	97.98	0.98	97.98	93.64	90.84	91.98	0.93	92.32
	Average	**97.71**	**98.32**	**98.27**	**0.98**	**98.03**	**93.27**	**90.96**	**91.93**	**0.93**	**92.17**
128	R1	98.41	98.37	98.41	0.98	98.39	90.70	88.41	89.72	0.90	89.62
	R2	98.41	97.55	97.64	0.98	97.99	91.80	89.46	90.54	0.91	90.69
	R3	97.47	96.15	95.85	0.97	96.78	89.11	89.87	90.71	0.90	89.47
	R4	96.76	100.00	100.00	0.98	98.39	90.49	88.51	89.51	0.90	89.54
	R5	99.19	99.19	99.19	0.99	99.19	91.08	87.30	88.95	0.90	89.30
	Average	**98.05**	**98.25**	**98.22**	**0.98**	**98.15**	**90.64**	**88.71**	**89.89**	**0.90**	**89.72**
64	R1	99.20	97.56	97.65	0.98	98.39	89.48	87.11	88.56	0.89	88.36
	R2	98.41	99.18	99.20	0.99	98.79	89.62	88.62	89.65	0.90	89.15
	R3	98.73	98.08	97.91	0.98	98.39	89.42	86.96	88.39	0.89	88.26
	R4	99.60	100.00	100.00	1.00	99.80	91.56	89.74	90.63	0.91	90.69
	R5	98.79	98.39	98.39	0.99	98.59	91.05	89.34	90.56	0.91	90.24
	Average	**98.95**	**98.64**	**98.63**	**0.99**	**98.79**	**90.23**	**88.35**	**89.56**	**0.90**	**89.34**
32	R1	99.60	99.19	99.21	0.99	99.40	88.96	86.78	88.24	0.89	87.93
	R2	99.21	99.59	99.60	0.99	99.40	91.13	88.08	89.37	0.90	89.68
	R3	97.89	96.54	96.27	0.97	97.18	88.13	88.53	89.50	0.89	88.32
	R4	97.98	98.39	98.37	0.98	98.19	88.99	87.92	88.87	0.89	88.48
	R5	100.00	97.98	98.02	0.99	98.99	89.86	87.48	88.96	0.89	88.74
	Average	**98.94**	**98.34**	**98.29**	**0.98**	**98.63**	**89.41**	**87.76**	**88.99**	**0.89**	**88.63**

increase of the batch size. This is due to the size of the subjects in the datasets. Since UFP dataset has a small number of subjects in it, that's why increasing the batch size causes over fitting and reduces the accuracy. On the other hand, TUEP dataset is a large dataset that has an impact on the training batch size.

To compare the impact of the batch size on the accuracy of different testing rounds, we have plotted the round wise accuracy values for different batch sizes on the tested datasets in Fig. 3. Here, Fig. 3a and Fig. 3b shows the accuracy comparison for dataset UFP and TUEP, respectively. From Fig. 3a and Table II, we can see that, for UFP, a single round highest accuracy of 99.80% is achieved for round 4 with a batch size of 64 and the lowest accuracy of 96.58% is received in round 3 for batch size 256. For TUEP, those values are 93.28% and 88.26% for round 1 with batch



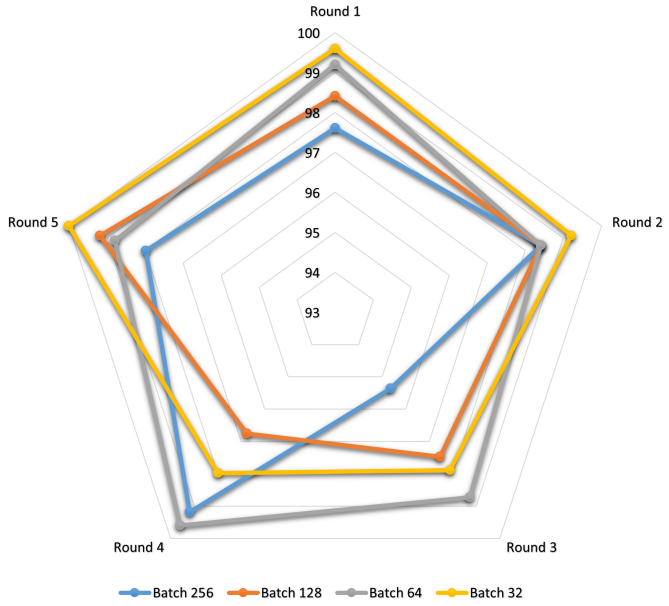
(a) UFP Dataset

(b) TUEP Dataset

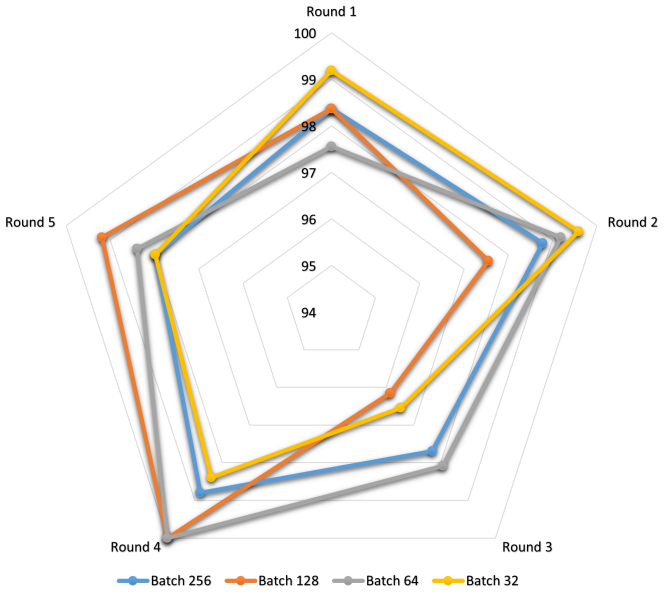
Fig. 3: Radar chart visualization of accuracy comparison for five folds with four training batch sizes for the tested datasets. Data has five folds (Round 1, 2, .., 5) and each polygon is a multivariate data point for a training batch size.

size 256 and round 3 with batch size 64, respectively. Five round average highest accuracy for UFP and TUEP are 98.79% and 92.17% for batch sizes 64 and 256, respectively. While for five round average lowest accuracy for UFP and TUEP are 98.03% and 88.63% for batch sizes 256 and 32, respectively.

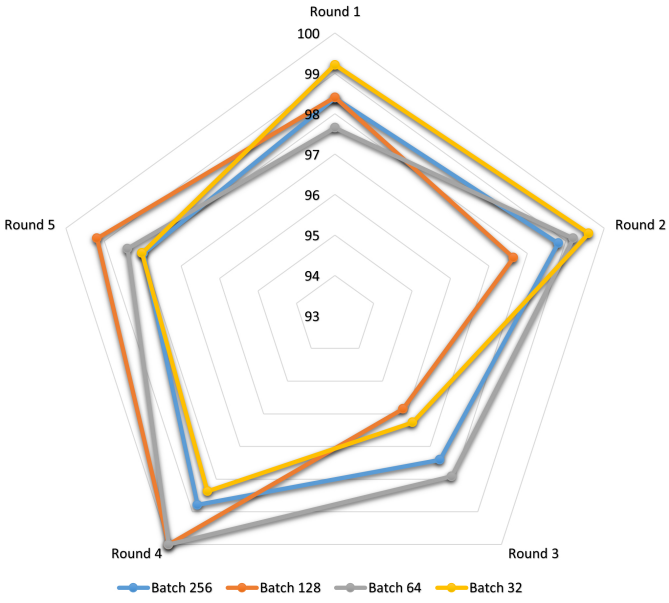
Also, we have calculated and compared four performance parameters (sensitivity, specificity, precision and F1 score) for both the datasets and plotted them in Fig. 4 and Fig. 5. For UFP dataset, we have single round highest sensitivity of 100% in round 5 with batch size 32 and a lowest sensitivity of 95.36% for round 3 with batch size 256. In case of five round average, 98.95% and 97.71% are the highest and lowest sensitivity value for UFP with batch size 64 and 256, respectively. On the other hand, for TUEP dataset, the highest single round sensitivity value is 93.98% for round 1 with batch sizes 256 and the lowest value is 88.13% for round 3 with batch size 32. Five round average highest and lowest sensitivity values for TUEP are 93.27% and 89.41% with batch size 256 and 32, respectively. Since sensitivity (also known as true positive rate) refers to the number of truly identified patients from HC subjects and the proposed model has high sensitivity values for both the datasets which indicates that the



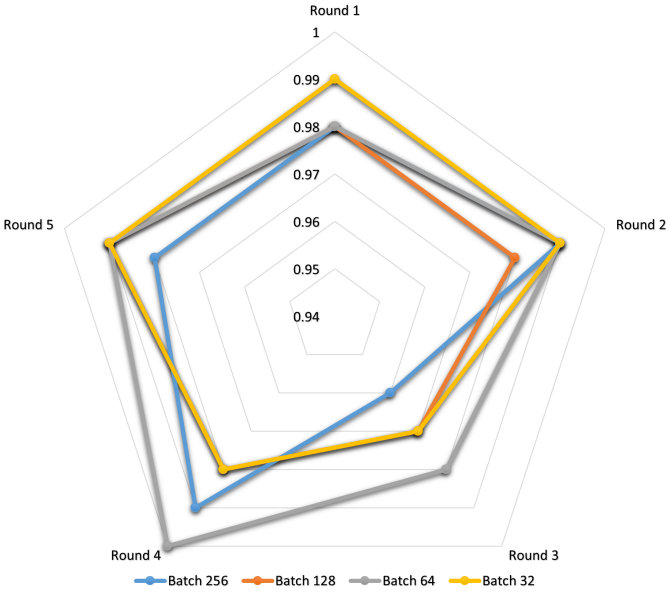
(a) Sensitivity



(b) Specificity



(c) Precision



(d) F1 score

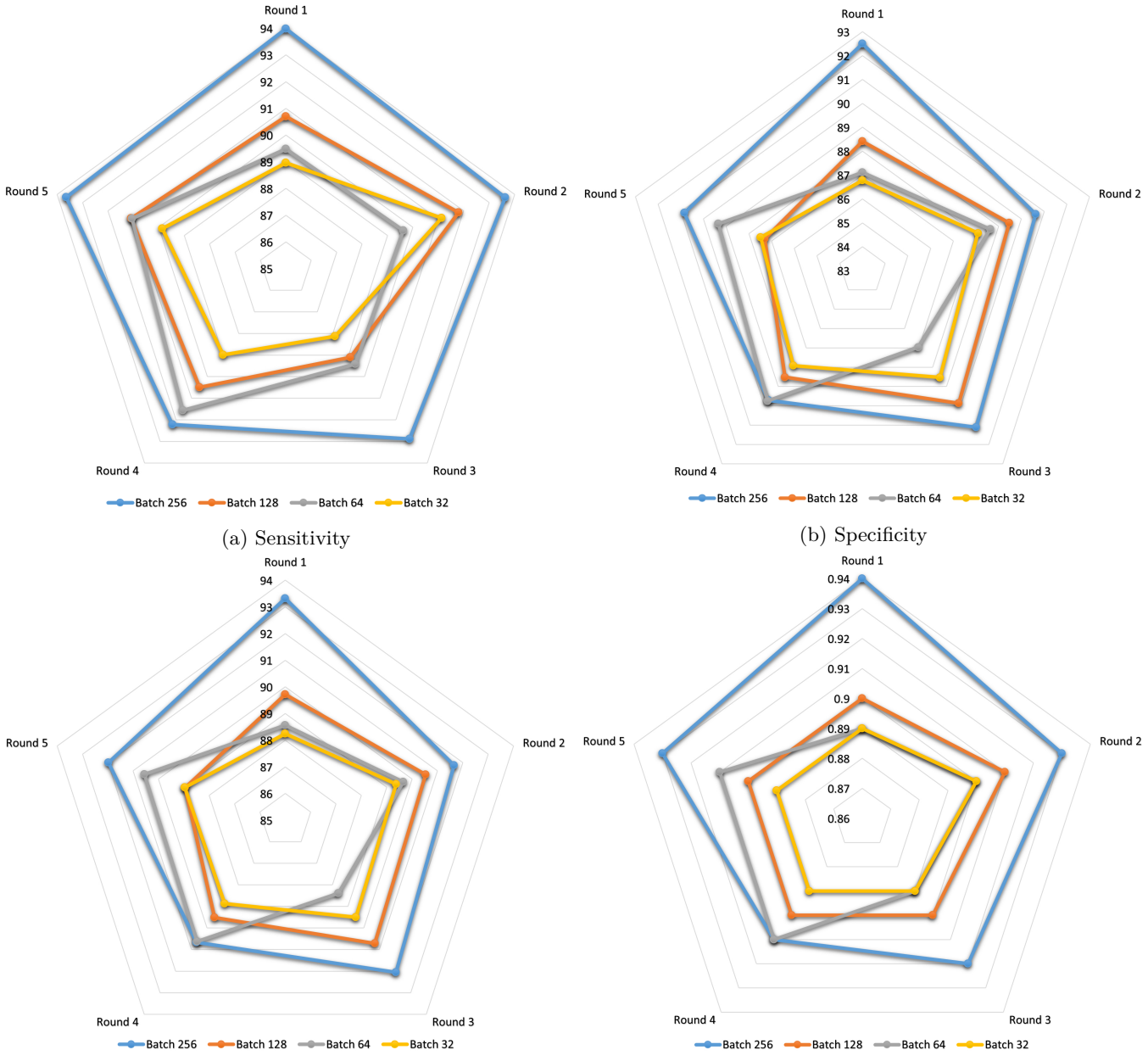
Fig. 4: Radar chart visualization of four evaluation parameters for five folds with four training batch sizes of UFP dataset. Data has five folds (Round 1, 2, .., 5) and each polygon is a multivariate data point for a training batch size.

proposed model is suitable for EEG signal analysis to detect epilepsy.

The second evaluation parameter we have considered is specificity (also known as true negative rate) which indicates the ability to classify the HC subjects from patient subjects. For UFP dataset, our proposed model has a single round highest specificity value of 100% for several testing rounds while for TUEP dataset, it is 92.50% for round 1 with batch size 256. The lowest specificity values for UFP is 96.15% for round 3 with batch size 128 and for TUEP, it is 86.96% for round 3 with batch size 64. Five round average highest and lowest specificity for UFP are 98.64% and 98.25%, and for TUEP are 90.96% and 87.76%, respectively.

The next parameter we have used to measure the classification performance of the proposed model is precision (also termed as positive predictive value), which is the percentage of original patients among the identified patients. From Table II, Fig. 4c and Fig. 5c we can see that, for UFP dataset, average highest and lowest precision values are 98.63% and 98.22% for batch size 64 and 128, respectively. For TUEP dataset, those values are 91.93% and 88.99% for batch sizes 256 and 32, respectively. The high value of the precision indicates that the proposed model is capable of identifying genuine patients from the test set.

We have calculated the last evaluation parameter named



(a) Sensitivity

(b) Specificity

(c) Precision

(d) F1 score

Fig. 5: Radar chart visualization of four evaluation parameters for five folds with four training batch sizes of TUEP dataset. Data has five folds (Round 1, 2, .., 5) and each polygon is a multivariate data point for a training batch size.

F1 score, which is the harmonic mean of precision and recall and an important performance measure for the classification task. For UFP, round wise F1 score varies from 0.98 to 1.00 with an average highest of 0.99 and lowest of 0.98. For TUEP, round wise F1 score varies from 0.89 to 0.94 with the average highest and lowest values are 0.93 and 0.89, respectively. Round wise comparison of those values are given in Fig. 4d and Fig. 5d

Finally, we have plotted the sensitivity in the y-axis and (1-specificity) in the x-axis of a line chart to create the receiver operating characteristic (ROC) curve for both the datasets as shown in Fig. 6. Here, we have only plotted the ROC curve for the best performing batch sizes which are batch size 64 for UFP dataset and batch size 256 for TUEP dataset. From the figure we can see that for UFP dataset, the curve is close to the point (0,1) and for TUEP dataset it is much closer to the (0,1) point. This indicates that the proposed model is a good classifier for both datasets.

To further assess the performance of the proposed model, we have also tested it using LOOCV technique where all segments of a subject is left out in the training process and those left out segments are used for testing the trained model. This process is repeated for each subject in the dataset. For UFP dataset, the highest average accuracy of 83.06% is achieved using batch size 32, while for TUEP dataset, it is 86.29% for batch size 128.

Fig. 7 and Fig. 8 shows the detailed subject wise accuracy comparison for different batch sizes for dataset UFP and TUEP, respectively. From the Fig. 7 we can see that 8 out of 14 subject achieved an accuracy greater than 99% and for 4 subject it is grater than 82% and less than 99%.

For TUEP dataset, we achieved an accuracy greater or equal to 99% for 87 subjects and for 21 subjects, it is between 95% and 99%. Among the 187 subjects 140 subjects have an average accuracy greater than 80%.

C. Discussion

A ConvLSTM based epilepsy classification framework using EEG data signal is developed in this research work. In the below subsections, few analysis of the proposed model in different aspects are given.

1) Ablation study

To check the optimality of the proposed model, we have carried out an ablation study on the model using both datasets. In this approach, we have used the proposed model's result for batch size 64 and 256 as baseline accuracy for UFP and TUEP dataset, respectively and have conducted five different ablation studies using same batch sizes and the obtained results are reported in Table III.

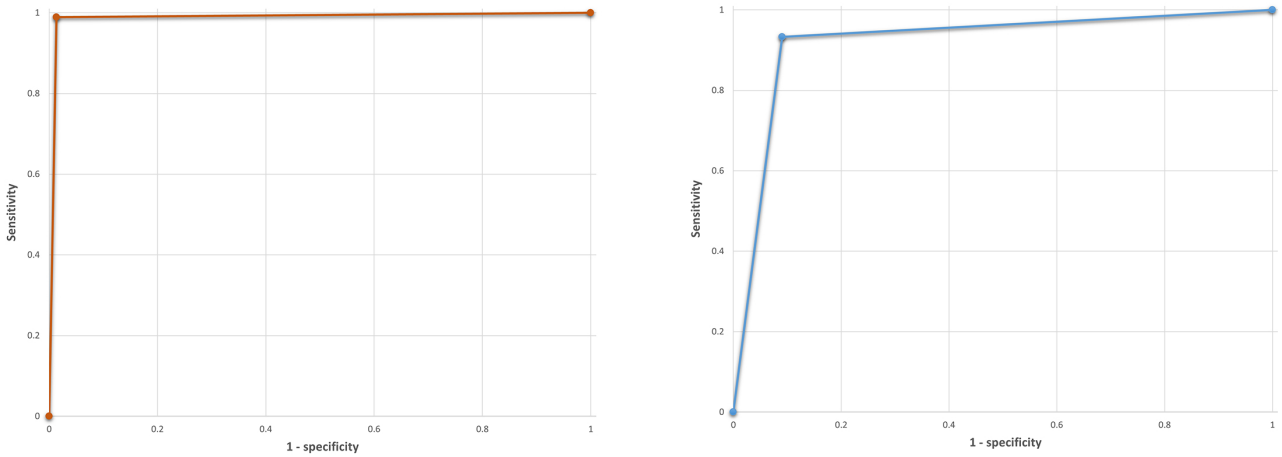
TABLE III: Accuracy comparison for different ablation studies on the tested datasets.

Ablation techniques	**Accuracy (%)**	
	UFP	**TUEP**
Baseline (no ablation)	98.63%	92.17%
Removed the hidden ConvLSTM and pooling layer	96.83%	80.32%
Added new hidden ConvLSTM and pooling layer	89.59%	91.43%
Doubled the neurons in both ConvLSTM layers	98.31%	89.50%
Halved the neurons in both ConvLSTM layers	98.07%	86.68%
Halved the neurons in hidden ConvLSTM layer	98.43%	87.61%
Doubled the neurons in hidden ConvLSTM layer	98.35%	87.81%

From Table III, we can see that increasing or decreasing hidden layer drops the accuracy from the baseline accuracy which indicates that a single hidden layer is enough for separating the two classes. The other ablation method we have tested is changing the number of neurons of the ConvLSTM layers. In baseline model we have used 16 neurons in both the input and the hidden layer. We have achieved an accuracy of 98.31% and 89.50% by doubling the neurons for both layers and an accuracy of 98.07% and 86.68% obtained by halving the neurons for both layers for UFP and TUEP dataset, respectively. We have also tested the ablation model by halving and doubling the neurons in the hidden layer and obtained an accuracy of 98.43% and 98.35%, respectively for UFP dataset and 87.61% and 87.81%, respectively for TUEP dataset.

2) Layer wise t-SNE based feature visualization

As we all know, deep learning models operate like a black box, making it difficult to fully understand their



(a) UFP Dataset (training batch size 64) (b) TUEP Dataset (training batch size 256)

Fig. 6: ROC curve for the tested datasets with best performance configuration.

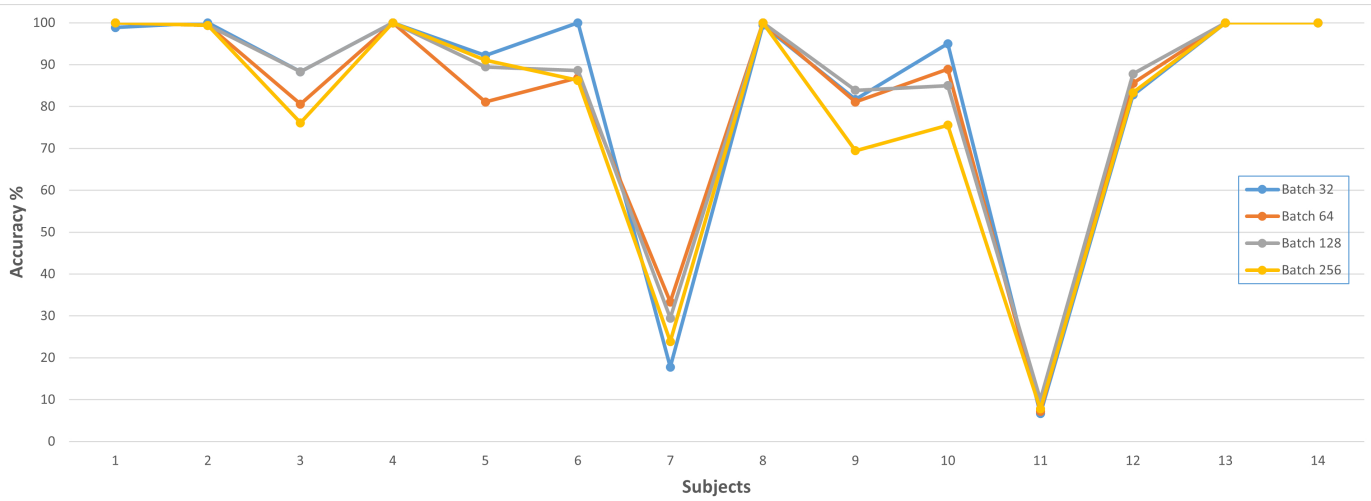


Fig. 7: Subject wise accuracy comparison for different batch sizes in LOOCV technique for UFP dataset.

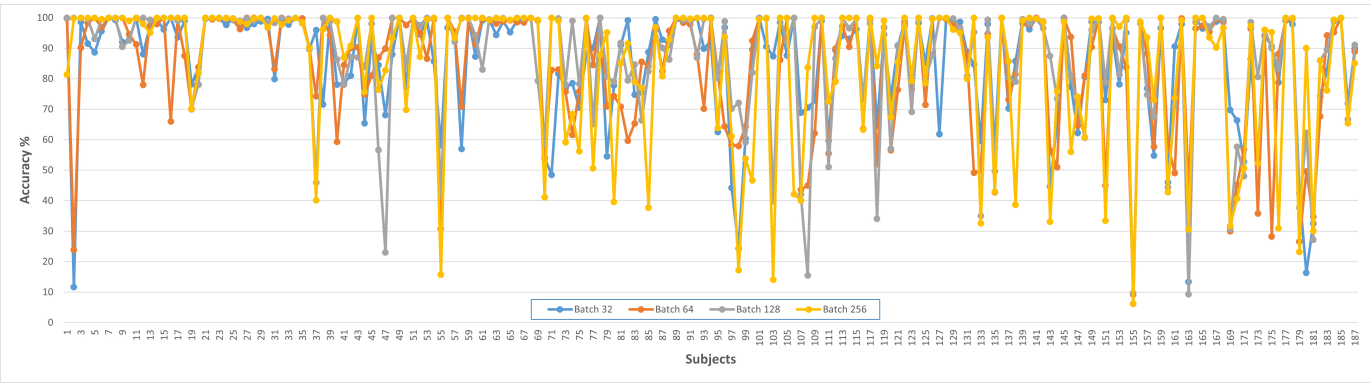


Fig. 8: Subject wise accuracy comparison for different batch sizes in LOOCV technique for TUEP dataset.

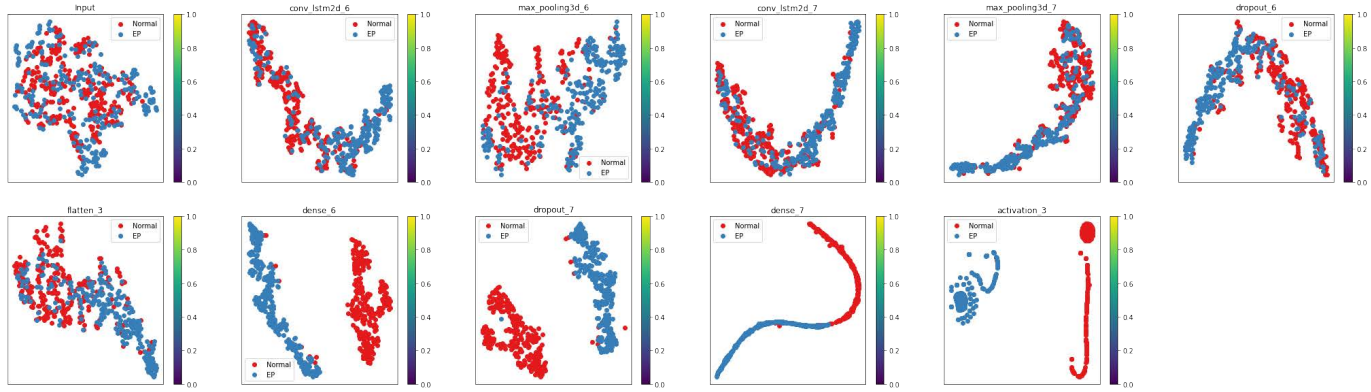


Fig. 9: Layer wise classification process visualization of the proposed model using t-SNE technique.

feature extraction process. T-SNE [30] is an efficient dimension reduction method that allows us to visualize high-dimensional data by mapping it to a two-dimensional space, which is often used to visualize the features extracted by the deep learning models. Fig 9 shows the t-SNE visualization of the classification process of test cases for each layer of the proposed model on UFP dataset. From the figure, we can see that in the input layer, all the subjects' data are mixed up and as the data passes through each layer of the model, they have started forming two cluster of two separate groups and after the final classification layer, they have formed two clearly separable clusters. This indicates that the proposed model is able to separate the EP and normal groups and clusters them into two separable groups.

3) Time complexity analysis of the proposed model

We have carried out our experiments in a PC with AMD Threadripper Pro processor with 256GB RAM and 48GB

graphics. The time-complexity analysis of the proposed model with different setups are given in Table IV.

TABLE IV: Time-complexity analysis of the proposed model for tested datasets with different setups.

Dataset validation	Time / epoch	Training		Validation		Batch size
		Acc.%	Loss	Acc.%	Loss	
UFP 5-fold	3s	99.92	0.002	98.63	0.12	32
	2s	99.81	0.009	98.79	0.20	64
	2s	99.97	0.010	98.15	0.25	128
	1s	99.97	0.009	98.03	0.29	256
TUEP 5-fold	103s	96.06	0.11	88.63	0.39	32
	73s	96.59	0.09	89.34	0.43	64
	53s	97.01	0.08	89.72	0.49	128
	35s	96.93	0.08	92.17	0.26	256

From the table, we can see that with the increase of training batch size, both the training and validation accuracy increased for TUEP dataset and the training time decreased. Similarly, for UFP dataset, training time decreased, and training accuracy increased but there was no steady pattern in the validation accuracy. Considering the above time-complexity analysis, using 256 batch size will gain good accuracy with small training time.

4) Data Augmentation

Data augmentation offers a quick approach to provide extra labelled data to train a network, and has been utilized particularly in the context of deep learning. Rotation, flipping, colour shift, and other similar two-dimensional changes that maintain the integrity of the image and label are frequent examples of data augmentation in the context of images [31]. But due to the nature of the EEG signal, very few data augmentation methods can be used with still keeping similar frequency, spatial, and power components. Here, we have used seven data augmentation methods as used by the authors in [31], which are multiplication, frequency shift, adding noise, flipping data and combination of those four alteration methods. We have tested those augmentation methods on UFP dataset and the obtained results on those augmented data are given in Table V.

TABLE V: Accuracy comparison for different data augmentation techniques on UFP dataset.

Augmentation techniques	**Accuracy (%)**
No augmentation	98.63%
Multiplied signal (Multi)	99.74%
Adding noise (Noise)	99.41%
Flipping the data (Flip)	99.09%
Frequency shifting (Freq)	99.44%
Noise + Flip	99.58%
Noise + Multi	99.79%
Flip + Freq	99.55%

From the Table V, we can see that data augmentation increases the performance of the proposed model and proves its stability to the perturbations.

5) Performance Comparison

Finally, to compare the performance of the proposed model with existing studies that have used the same datasets, we searched for existing studies. For TUEP dataset, authors from the study [32] used this dataset with Tiny-Visual Geometry Group (t-VGG) and its t-VGG Global Average Pooling (t-VGG GAP) variant CNN models and reported an accuracy of 81.42% using t-VGG GAP, while in this study we have achieved an accuracy of 92.17% that surpasses the results of the previous studies. For UFP dataset, as far as we have searched, this study is the first study that have used for binary classification. Details of the comparison are given in Table VI.

IV. Conclusions

In this study, we have proposed deep learning-based framework for classifying epilepsy using EEG data. We have used the ConvLSTM model for extracting features and classifying them into one of two classes (EP vs healthy). ConvLSTM is a combination of CNN and LSTM models which is capable of extracting spatiotemporal information from time-series data. We have tested our model on two different datasets from two publicly available sources and performed both five-fold cross-validation and leave-one-out cross-validation to validate the performance. The experimental results indicate that the proposed ConvLSTM model offers higher performance for both the datasets and outperforms the state-of-the-art existing results for the datasets. The proposed framework has produced an overall correct classification rate of 98.79% and 92.17% for UFP and TUEP dataset, respectively in five-fold cross-validation process, while for LOOCV, those values are 83.06% and 86.29%, respectively.

In ending, the obtained results reveal that the proposed model is robust and extensible and can be used in studies involving EEG data and signal processing techniques. Notwithstanding, the framework's high classification accuracy indicates that EEG data segment as short as 3s is enough for identifying epilepsy disease. In future, this study can be used to develop a real-time application for assisting specialists in automatically and efficiently identifying epilepsy from EEG signal data.

References

[1] WHO, "Epilepsy," (Date last accessed February 2021). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy

[2] W. Hussain, M. T. Sadiq, S. Siuly, and A. U. Rehman, "Epileptic seizure detection using 1 d-convolutional long short-term memory neural networks," *Applied Acoustics*, vol. 177, p. 107941, 2021.

[3] S. Siuly, O. F. Alcin, V. Bajaj, A. Sengur, and Y. Zhang, "Exploring hermite transformation in brain signal analysis for the detection of epileptic seizure," *IET Science, Measurement & Technology*, vol. 13, no. 1, pp. 35–41, 2018.

[4] S. Siuly, Y. Li, and Y. Zhang, "Eeg signal analysis and classification," *IEEE Trans Neural Syst Rehabilit Eng*, vol. 11, pp. 141–4, 2016.

[5] M. N. A. Tawhid, S. Siuly, H. Wang, F. Whittaker, K. Wang, and Y. Zhang, "A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from eeg," *Plos one*, vol. 16, no. 6, p. e0253094, 2021.

[6] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust eeg single-trial analysis," *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2007.

TABLE VI: Comparison of the proposed model with the existing best models for epilepsy disease detection.

Study	Database	Subjects	Method	Classifier	Class	Performance
Hussain *et al.* [2]	Freiburg	21	Time-Frequency Domain	CNN+LSTM	Ictal-Preictal-Interictal	Acc: 94.04%
Zhou *et al.* [33]	Freiburg	21	Frequency Domain	CNN	Ictal-Preictal-Interictal	Acc: 92.3%
Zhao *et al.* [34]	CHB-MIT	19	Raw Data	AddNet-SCL	seizure-non seizure	Sen: 94.9
Shen *et al.* [35]	CHB-MIT	23	DWT + time domain	RUSBoosted tree	seizure-non seizure	Acc: 96.38% Sen: 96.15%
Shen *et al.* [35]	Bonn University	10	DWT + time domain	SVM	healthy- seizure free-seizure active	Acc: 97% Sen: 81.42%
Uyttenhove *et al.* [32]	TUEP	200	Raw Data	t-VGG GAP	healthy-epilepsy	Spec: 80.95% Acc: 81.42%
This study	TUEP	187	Raw Data	ConvLSTM	healthy-epilepsy	**Sen: 93.27%** **Spec: 90.96%** **Acc: 92.17%**
This study	UFP	14	Raw Data	ConvLSTM	healthy-epilepsy	**Sen: 98.95%** **Spec: 98.64%** **Acc: 98.79%**

[7] M. N. A. Tawhid, S. Siuly, and H. Wang, "Diagnosis of autism spectrum disorder from eeg using a time–frequency spectrogram image-based approach," *Electronics Letters*, 2020.

[8] S. Supriya, S. Siuly, H. Wang, and Y. Zhang, "Epilepsy detection from eeg using complex network techniques: A review," *IEEE Reviews in Biomedical Engineering*, 2021.

[9] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of eeg signals," *IEEE Access*, vol. 6, pp. 25 399–25 410, 2018.

[10] K. Rasheed, A. Qayyum, J. Qadir, S. Sivathamboo, P. Kwan, L. Kuhlmann, T. O'Brien, and A. Razi, "Machine learning for predicting epileptic seizures using eeg signals: A review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 139–155, 2020.

[11] P. Boonyakitanont, A. Lek-Uthai, K. Chomtho, and J. Songsiri, "A review of feature extraction and performance evaluation in epileptic seizure detection using eeg," *Biomedical Signal Processing and Control*, vol. 57, p. 101702, 2020.

[12] K. Abualsaud, M. Mahmuddin, M. Saleh, and A. Mohamed, "Ensemble classifier for epileptic seizure detection for imperfect eeg data," *The Scientific World Journal*, vol. 2015, 2015.

[13] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–17, 2019.

[14] C. Gómez, P. Arbeláez, M. Navarrete, C. Alvarado-Rojas, M. Le Van Quyen, and M. Valderrama, "Automatic seizure detection based on imaged-eeg signals through fully convolutional networks," *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.

[15] Z. Xiong, H. Wang, L. Zhang, T. Fan, J. Shen, Y. Zhao, Y. Liu, and Q. Wu, "A study on seizure detection of eeg signals represented in 2d," *Sensors*, vol. 21, no. 15, p. 5145, 2021.

[16] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in biology and medicine*, vol. 99, pp. 24–37, 2018.

[17] R. Hussein, H. Palangi, R. K. Ward, and Z. J. Wang, "Optimized deep neural network architecture for robust detection of epileptic seizures using eeg signals," *Clinical Neurophysiology*, vol. 130, no. 1, pp. 25–37, 2019.

[18] Y. Yang, N. D. Truong, C. Maher, A. Nikpour, and O. Kavehei, "Continental generalization of an ai system for clinical seizure recognition," *arXiv preprint arXiv:2103.10900*, 2021.

[19] A. Pereira and J. Fiel, "Resting-state interictal eeg recordings of refractory epilepsy patients," 2019. [Online]. Available: https://data.mendeley.com/datasets/6hx2smc7nw/1

[20] L. Veloso, J. McHugh, E. von Weltin, S. Lopez, I. Obeid, and J. Picone, "Big data resources for eegs: enabling deep learning research," in *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 2017, pp. 1–3.

[21] A. Morley, L. Hill, and A. Kaditis, "10-20 system eeg placement," *European Respiratory Society, European Respiratory Society*, 2016.

[22] M. Tawhid, N. Ahad, S. Siuly, K. Wang, and H. Wang, "Data mining based artificial intelligent technique for identifying abnormalities from brain signal data," in *International Conference on Web Information Systems Engineering*. Springer, 2021, pp. 198–206.

[23] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[24] S. Siuly, Ö. F. Alçin, E. Kabir, A. Şengür, H. Wang, Y. Zhang, and F. Whittaker, "A new framework for automatic detection of patients with mild cognitive impairment using resting-state eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 9, pp. 1966–1976, 2020.

[25] S. Siuly, S. K. Khare, V. Bajaj, H. Wang, and Y. Zhang, "A computerized method for automatic detection of schizophrenia using eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 11, pp. 2390–2400, 2020.

[26] S. Siuly, X. Yin, S. Hadjiloucas, and Y. Zhang, "Classification of thz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 64–82, 2016.

[27] M. N. A. Tawhid and E. K. Dey, "A gender recognition system from facial image," *International Journal of Computer Applications*, vol. 180, no. 23, pp. 5–14, 2018.

[28] E. K. Dey, M. Tawhid, N. Ahad, and M. Shoyaib, "An automated system for garment texture design class identification," *Computers*, vol. 4, no. 3, pp. 265–282, 2015.

[29] C. Sammut and G. I. Webb, Eds., *Leave-One-Out Cross-Validation*. Boston, MA: Springer US, 2010, pp. 600–601.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[31] D. Freer and G.-Z. Yang, "Data augmentation for self-paced motor imagery classification with c-lstm," *Journal of neural engineering*, vol. 17, no. 1, p. 016041, 2020.

[32] T. Uyttenhove, A. Maes, T. Van Steenkiste, D. Deschrijver, and T. Dhaene, "Interpretable epilepsy detection in routine, interictal eeg data using deep learning," in *Machine Learning for Health*. PMLR, 2020, pp. 355–366.

[33] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, and J. Xiang, "Epileptic seizure detection based on eeg signals and cnn," *Frontiers in neuroinformatics*, vol. 12, p. 95, 2018.

[34] Y. Zhao, C. Li, X. Liu, R. Qian, R. Song, and X. Chen, "Patient-specific seizure prediction via adder network and supervised contrastive learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1536–1547, 2022.

[35] M. Shen, P. Wen, B. Song, and Y. Li, "An eeg based real-time epilepsy seizure detection approach using discrete wavelet transform and machine learning methods," *Biomedical Signal Processing and Control*, vol. 77, p. 103820, 2022.