THREE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK BASED DETECTION OF EPILEPTIC SEIZURES FROM VIDEO DATA

Aidan Boyne[1], Hsiang J. Yeh[2], Anthony K. Allam[1], Brandon M. Brown[1], Mohammad Tabaeizadeh[2], John M. Stern[2], R. James Cotton[3,4], Zulfi Haneef[1,5]

1) Department of Neurology, Baylor College of Medicine, Houston, Texas, USA
2) Department of Neurology, University of California at Los Angeles, Los Angeles, California, USA
3) Shirley Ryan Ability Lab, Chicago, Illinois, USA
4) Department of Physical Medicine and Rehabilitation, Northwestern University, Chicago, Illinois, USA
5) Neurology Care Line, Michael E. DeBakey VA Medical Center, Houston, Texas, USA

Corresponding Author

Name: Zulfi Haneef, MD

Address: 7200 Cambridge St #9A, Houston, TX, 77030, USA

E-mail: zulfi.haneef@bcm.edu

Phone: (713) 798-0980

Word count

Manuscript: 3038

Abstract: 300

Character count for the title: 100

Tables: 3

Figures: 3

Ethical publication statement: We confirm that we have read the Journal's position on issues involved in ethical publication and a□rm that this report is consistent with those guidelines.

Data Availability Statement: Code used to generate the results in this study is based on the MMaction2 toolbox and is publicly available at https://github.com/aidanboyne/CVEpilepsy under the Apache 2.0 license. The video data is not available due to patient privacy issues.

Acknowledgements: None.

Funding: ZH receives research support from Neuro Events Lab (NEL) and the Dept. of Defense (DoD), CDMRP Virtual Post-Traumatic Epilepsy Research Center Faculty Award HT9425-24-1-0355. RJC receives funding from the Research Accelerator Program of the Shirley Ryan AbilityLab. The remaining authors have no conflicts of interest.

Disclosures: No commercial conflicts of interest are reported by the authors.

Author Contributions: Aidan Boyne – Writing original draft, methodology, software, investigation, formal analysis, review and editing; Hsiang J. Yeh – software, formal analysis, validation; Anthony K. Allam – conceptualization, data curation, software; Brandon M. Brown – data curation; Mohammed T. Fesharaki – data curation; John Stern – resources, R. James Cotton – methodology, supervision; Zulfi Haneef – project administration, supervision, review and editing.

Patient Consent Statement: Not applicable

Clinical Trial Registration: Not applicable

ABSTRACT

Objective: Seizure detection in epilepsy monitoring units (EMU) is essential for successful seizure characterization and for the clinical assessment of drug-resistant epilepsy. Automated video analysis using machine learning provides a promising aid for seizure detection with resultant reduction in the resources required for diagnostic monitoring. We employ a novel approach, using a 3D convolutional neural network to identify seizures from EMU videos.

Methods: A two-stream inflated 3D-ConvNet architecture classified video clips as a seizure or not a seizure. A pretrained action classification model was fine-tuned on 11 hours of video data containing 49 tonic-clonic seizures from 25 patients monitored at a large academic hospital (site A) using leave-one-patient-out cross-validation. Performance was evaluated by comparing model predictions to ground-truth annotations obtained from video-EEG review by an epileptologist on videos from site A and on a separate dataset from a second large academic hospital (site B).

Results: The model classified previously unseen videos from site A with a mean accuracy of 0.944 ± 0.064 and area under the receiver operating curve (AUC) score of 0.991 ± 0.015. The mean latency between seizure onset and seizure detection was 0.61 ± 2.01 seconds, which surpasses the prior state-of-the-art. A mean accuracy of 0.964 ± 0.034 and AUC score of 0.984 ± 0.049 were achieved by the model trained and tested at site B. Cross-site evaluation (site A model tested on site B data and vice versa) demonstrated generalizability of the model with a mean decrease in AUC score of -0.156 from same-site evaluation.

Significance: Our model demonstrates high performance in detection of epileptic seizures from video data using spatiotemporal 3D-CNN models and outperforms prior similar models identified in the literature. This study provides a foundation for future work in real-time EMU seizure monitoring and possibly for reliable and cost-effective at-home detection of tonic-clonic seizures.

KEYWORDS: Epilepsy, seizure detection, neural network, machine learning, video

KEY POINTS

- We evaluate a video-based 3-D CNN for seizure detection in patients undergoing evaluation in an EMU at 2 large academic hospitals.
- Our video-only model provides highly accurate detection of tonic-clonic seizures with low detection latency.
- The underlying model architecture requires no video-preprocessing and is generalizable across two EMUs.

1. INTRODUCTION

Accurate and timely detection of epileptic seizures is critical for safe and efficient video-EEG evaluations and would improve patient safety at home when alone (1). Evaluation in an epilepsy monitoring unit (EMU) is the standard evaluation for drug-resistant epilepsy and requires continuous monitoring by a care provider for safety. However, EMU access is often insufficient for the care needs of a community, and EMU admission is both expensive and time intensive for hospital staff (2). Alternatives to the EMU, such as wearable multimodal seizure detection devices using accelerometers and surface electromyography, have shown promise but have high false alarm rates and are often impractical for prolonged use (3,4).

Recent advances in computer vision and deep learning offer a potential solution to these limitations by providing non-contact seizure detection via video data analysis. Automated detection of seizures based only on video would allow at-home seizure monitoring and alert systems. Accurate and precise detection software could form the basis of at home and inpatient seizure alarms and play a role in reducing sudden unexpected death in epilepsy (5).

The numerous advantages of video-based seizure detection have made the subject an area of active research. Detection of the characteristically abnormal movement and/or posturing during seizures underlies the majority of preexisting work. Older methods primarily relied on tracking key points of interest, such as they joints and eyes, or measuring the amplitude and direction of limb and body movement (6–8). Performance of these models was generally low and was likely limited by machine learning techniques such as support vector machines that were able to utilize only a tiny fraction of the data available in the video recording. However, deep learning allows researchers to take advantage of the larger amount of information available in the video data. Many models using convolutional neural networks (CNN) have produced impressive results in seizure detection and classification tasks (9–13) and have largely outperformed other model architectures (14). In this investigation, we present a novel approach that uses a 3D-CNN-based algorithm for the automated detection of seizures from video input.

2. METHODS

2.1 Data Acquisition

All data were acquired from patients undergoing long-term video-EEG monitoring in the EMU at the Baylor St. Luke's Medical Center. The study was approved by the Institutional Review Board at Baylor College of Medicine. Videos were recorded at 30 frames per second at a resolution of 640x480 pixels using a SONY SNC EP520 camera mounted on the ceiling of the EMU patient room. These cameras automatically acquire color (RGB) video during high light conditions and greyscale video in low light conditions. Audio was also recorded but was not used in training or testing of the model. Videos corresponding to electro-clinical seizures underwent manual review by experts for visual evidence of generalized and focal seizures containing visible tonic-clonic and were segmented into seizure and non-seizure sections accordingly. Videos without any visually identifiable tonic-clonic seizures were excluded from the analysis. Two additional videos were excluded from the final analysis based on the patient being out of the video's frame, overly obstructed, or out of focus. EEG readings were captured using conventional scalp EEG

montages on a Nihon Kohden system using the international 10-20 system of electrode placement. EEG data was used only for labeling the videos and was not used in the training or testing of the model.

2.2 Dataset Preparation

A total of 21 hours of video data was collected from 25 patients. Clinical annotations were used to identify EMU videos for potential inclusion and sections of the video containing any visually evident seizure activity were marked. After excluding monitoring sessions without movement, 49 videos from the 25 patients remained, totaling to 11 hours and 4 minutes of usable video data. This data consisted of 55 minutes of seizure activity (8.3%) and 10 hours and 9 minutes (91.7%) of non-seizure activity based on the manual review described in section 2.1. (**Table 1**).

Patient	Video ID	Seizure Length (s)	Seizure %	Non-Seizure Length (s)	Non-Seizure %	Total Length (s)	Color
1	P1-V1	59	8.76%	614	91.24%	673	GS
	P1-V2	62	9.27%	607	90.73%	669	RGB
	P1-V3	58	5.21%	1055	94.79%	1113	RGB
2	P2-V1	78	6.18%	1185	93.82%	1263	RGB
3	P3-V1	94	11.80%	703	88.20%	797	RBG
4	P4-V1	70	7.82%	826	92.18%	896	RBG
	P4-V2	80	5.96%	1263	94.04%	1343	RBG
	P4-V3	89	13.15%	588	86.85%	677	RBG
5	P5-V1	79	11.83%	589	88.17%	668	GS
6	P6-V1	57	3.45%	1595	96.55%	1652	RBG
7	P7-V1	78	6.83%	1064	93.17%	1142	GS
8	P8-V1	50	5.74%	821	94.26%	871	RBG
9	P9-V1	81	21.80%	291	78.20%	372	RBG
10	P10-V1	75	20.53%	290	79.47%	365	RBG
11	P11-V1	34	10.58%	287	89.42%	321	GS
	P11-V2	44	9.97%	397	90.03%	441	RBG
	P11-V3	31	10.20%	273	89.80%	304	GS
	P11-V4	39	11.24%	308	88.76%	347	RBG
	P11-V5	38	8.61%	403	91.39%	441	RBG
	P11-V6	40	13.55%	255	86.45%	295	RBG
12	P12-V1	62	14.34%	370	85.66%	432	RBG
13	P13-V1	65	6.34%	961	93.66%	1026	RBG
14	P14-V1	66	6.42%	962	93.58%	1028	GS
	P14-V2	75	9.79%	691	90.21%	766	GS
	P14-V3	69	16.35%	353	83.65%	422	RBG
	P14-V4	67	5.90%	1068	94.10%	1135	GS
15	P15-V1	75	3.94%	1831	96.06%	1906	RBG
16	P16-V1	65	5.87%	1042	94.13%	1107	RBG
17	P17-V1	80	16.44%	407	83.56%	487	RBG
	P17-V2	68	8.33%	748	91.67%	816	RBG

	P17-V3	84	4.45%	1806	95.55%	1890	RBG
	P17-V4	69	30.45%	158	69.55%	227	RBG
	P17-V5	73	19.99%	292	80.01%	365	GS
	P17-V6	48	25.11%	143	74.89%	191	RBG
18	P18-V1	43	4.69%	874	95.31%	917	RBG
19	P19-V1	98	8.13%	1108	91.87%	1206	RBG
	P19-V2	125	9.02%	1261	90.98%	1386	GS
20	P20-V1	66	11.13%	527	88.87%	593	RBG
	P20-V2	44	6.31%	653	93.69%	697	RBG
21	P21-V1	198	12.21%	1424	87.79%	1622	RBG
22	P22-V1	65	6.36%	957	93.64%	1022	RBG
23	P23-V1	70	7.11%	915	92.89%	985	RBG
	P23-V2	61	9.00%	617	91.00%	678	RBG
	P23-V3	48	5.46%	831	94.54%	879	RBG
	P23-V4	62	10.37%	536	89.63%	598	RBG
24	P24-V1	61	5.58%	1032	94.42%	1093	GS
25	P25-V1	41	6.04%	638	93.96%	679	RBG
	P25-V2	57	11.01%	460	88.99%	517	RBG
	P25-V3	52	9.22%	512	90.78%	564	RBG
Total		**3293**	**8.26%**	**36589**	**91.74%**	**39882**	

Table 1: Breakdown of EMU monitoring video footage by patient and event. Each event is captured in a separate .mp4 video file. Fellowship trained epileptologists reviewed each video along with the corresponding EEG trace to classify segments of the video as seizure or non-seizure. Abbreviations: GS, Greyscale; RGB, Red Green Blue.

The data was then divided into 3-second-long clips. Clips were labeled as seizure if the clip contained at least one second of seizure-related movements. During training, an equal number of seizure clips and randomly selected non-seizure clips were selected to ensure an even balance between the two classes, a standard practice when training machine learning models (15). Testing was performed using the original balance of seizure and non-seizure videos present in the EMU data shown in **Table 1** to more closely approximate the predominance of non-seizure video in a real-world setting.

2.3 Model Structure

A two-stream inflated 3D CNN (I3D) was used as the base model for video classification (**Figure 1**). This model architecture had not previously been used for seizure detection but demonstrated excellent results when used to classify human action. The architecture is described in detail in the seminal work by Carreira and Zisserman (16). The structure of the model in the context of seizure detection is briefly described here.

The model consists of an initial 3D convolution layer followed by 2 spatiotemporal max-pooling layers. The 3D convolution layers allow the model to capture both spatial (image edges, textures, and shapes) and temporal (change in frames over time) features from the EMU videos. Output from the 3D convolution is compressed by the pooling layers and fed through four residual

layers composed of multiple Bottleneck3D blocks which progressively reduce spatial dimensions while increasing the depth of the feature maps. Finally, the features are used as input for the classification head which outputs a probability ranging from 0 to 1 where 0 represents no likelihood of seizure activity and 1 represents certain seizure activity.

The two streams of data used for classification were a raw RGB or greyscale video and the same video fed through an optical flow algorithm. The raw video stream provided spatial features such as objects and textures to the model, while the optical flow stream captured motion between consecutive frames of the video. The predictions of each stream were averaged to provide the final predicted probability of seizure activity within the video (**Figure 1**). Each video clip fed to the model was classified as seizure if the final predicted probability was greater than 0.5.

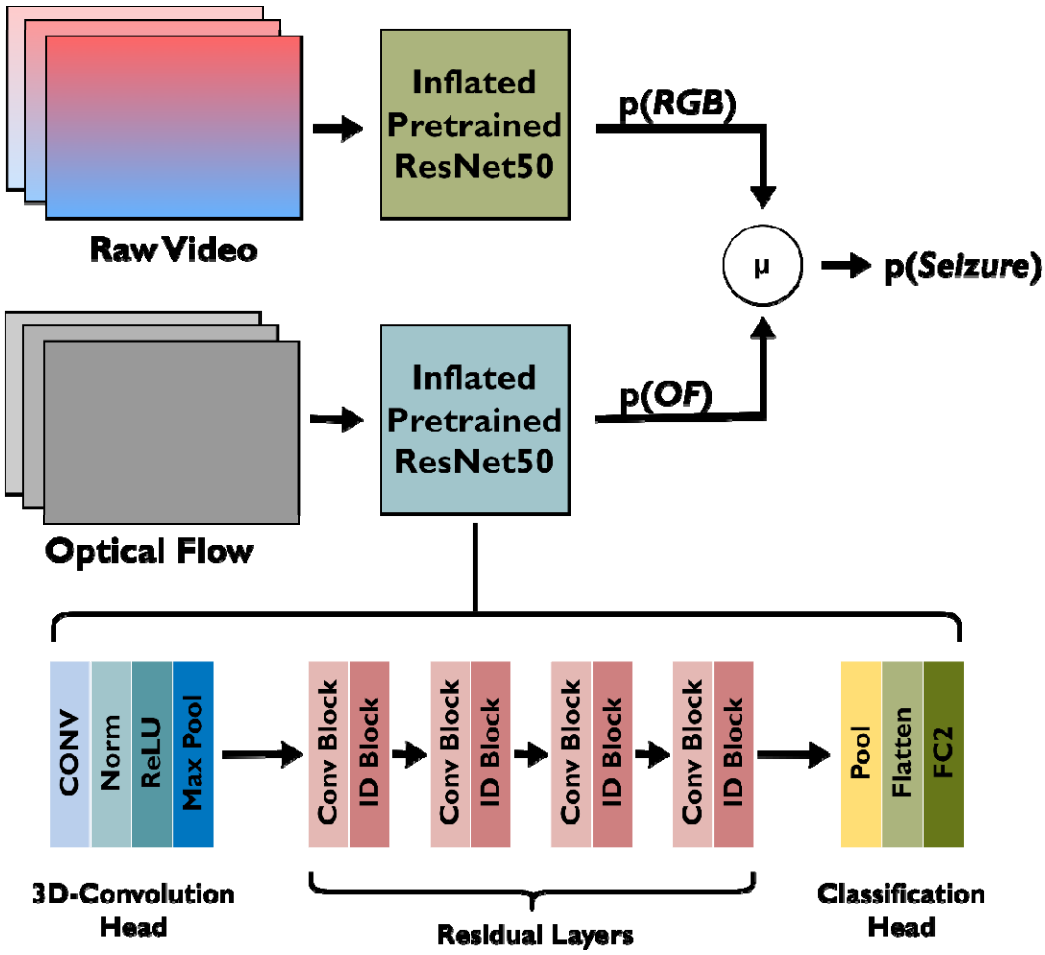

Figure 1: The I3D model consists of two inflated ResNet50 sub-models. These sub-models were fine-tuned on Image Net Data and the entire I3D model was then fine-tuned on the Kinetics Human Action Video 400 dataset *(17,18)*. The 3D-convolutional head extracts spatiotemporal features, stabilizes the input distribution, and introduces non-linearity to the model before reducing the size of input and passing it to the residual layers. These layers are defined by residual connections that allow the model to skip certain layers, leading to a more stable and convergent end result. Finally, the output of the residual layer is passed to the classification head which outputs a probability from 0 (not seizure) to 1 (seizure). 3-second video clips taken from the EMU were then fed to the model training pipeline which

automatically calculates OF for each clip before sending the OF and RGB videos to the appropriate submodel.

(Abbreviations: RGB, red-green-blue; p(RGB), Raw video stream predicted seizure probability; p(OF), optical flow stream predicted seizure probability; μ, arithmetic mean; p(Seizure); overall model predicted seizure probability; CONV, convolutional layer; Norm, normalization layer; ReLU, rectified linear unit activation function; ID Block, identity block; FC2, fully connected 2 class layer.)

2.4 Model Training and Evaluation

Two identical 2D-ResNet50 models pretrained on ImageNet (18), a gold standard image dataset for object recognition tasks, were used as the backbone of our seizure detection model. Weights from the models were inflated to allow the models to process spatiotemporal (3D) video data. The inflation process copies the initial model's 2D kernels multiple times along a third axis to make 3D kernels, analogous to stacking multiple 2D squares on top of one another to form a 3D cube. The temporally inflated ResNet50 models were then fine-tuned using data from the Kinetics Human Action Video 400 dataset, a dataset of various labeled videos often used to train models to recognize human activities, resulting in a general-purpose action recognition I3D model (17). This model is publicly available and has demonstrated high performance on action recognition tasks (16,19). Lastly, the output layer of the individual ResNet50 sub-models was shrunk from 400 to 2 outputs to allow for binary (seizure or non-seizure) classification of videos.

We used the MMaction2 training with a cross-entropy loss function to train the model on our seizure dataset (20). The framework is optimized for training and testing human action recognition models and was well suited for fine-tuning the I3D model for seizure detection. All training and validation were performed using a GeFORCE RTX 4090 graphics card [NVIDIA Corporation, Santa Clara, CA].

Cross-validation was used to mitigate any overfitting and provide a robust estimation of model performance for unseen subjects while still training the model on as much data as possible. We performed leave-one-patient-out (LOPO) cross-validation. Each fold consisted of 15 epochs of training, validation, and evaluation of the test dataset. The models with the highest accuracy and AUC score on the validation set were then used to calculate the performance of the model on unseen patients from the test set (**Figure 2**).

2.5 Cross-Institutional Training and Evaluation

After training and testing at the initial site (site A), the code for model training and testing was sent to the University of California, Los Angeles (site B). Site B then trained and cross-validated another series of seizure detection models on a separate dataset collected at their institution. This segment of the study was independently approved by the Institutional Review Board at the University of California, Los Angles. All site B videos were recorded using a SONY EP580 camera at an original resolution of 1080x1920 pixels and comprised of 3.13 total hours of EMU video data with 22.7**%** of the video data exhibiting seizure activity as identified by an epileptologist (**Supplementary Table 1**). All videos were then cropped to 4:3 aspect ratio and downscaled and rotated to match the 640x480 pixel resolution at site A. The model structure,

initial pre-trained weights, and cross-validation process were identical between the two sites. These results were used to test the generalizability of the underlying model architecture and

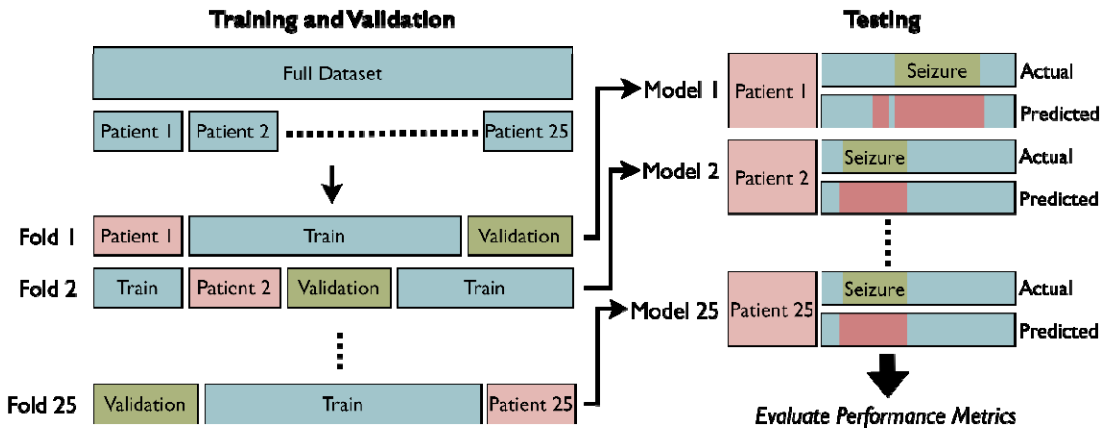

Figure 2: Model training and testing strategy. **a)** 25-fold, leave-one-patient-out cross-validation. In each fold, all videos from a single patient were set aside as the test set. All remaining videos were split into training and validation sets at a ratio of 80:20. Testing, training, and validation sets were then balanced as discussed in section 2.2 before training the model. **b)** Full video testing. The best performing models as determined by validation-set performance during cross-validation were used to classify all clips in each video sequentially to mimic use in a clinical scenario. Models were selected to classify videos that had never been seen in the training process to ensure valid performance metrics.

training process.

After evaluating the site B models on the site B dataset, the model with the best overall performance from site A was evaluated on the dataset from site B. Likewise, the best performing model from site B was evaluated on the dataset from site A. This process was designed to test the generalizability of a fully trained model to a dataset with different environmental conditions such as camera type, camera location, and bed setup.

2.6 Measures

Performance of the model was evaluated using accuracy, area under the receiver operator curve (AUC) score, sensitivity, specificity, and detection latency (14). Sensitivity is defined as the number of clips correctly identified as seizure divided by the total number of seizure clips. Specificity is defined as the number of clips correctly identified as non-seizure divided by the total number of non-seizure clips. Equations for performance measures are given in **Supplementary Figure 1**.

Detection latency is defined as the time in seconds from the clip marking the actual start of the seizure to the first clip classified as seizure by the model. In cases where the model predicts seizure prior to the actual start of seizure activity, detection latency is defined as zero seconds. Due to the randomized order of video clips during the cross-validation process, detection latency is only reported for the full-video tests. It is also important to note that the dataset used during

cross-validation was balanced to include an even number of seizure and non-seizure video clips in accordance with standard practice for model training. Performance data for the cross-validation process is reported separately from model testing on full seizure videos which include a much lower proportion of seizure video clips to more closely reflect the real-world EMU setting.

3. RESULTS

3.1 Model Performance during Cross-Validation

Metrics from the top performing epoch from each of the 25 folds are available in **Supplementary File 1**. Across the 25 top performing epochs, average accuracy was 95.94% ± 0.75% and average AUC was 0.988 ± 0.004. These results represent classification performance of the model on isolated 3-second video clips, with an even balance of seizure and non-seizure clips. The full results of the cross-validation process are presented in comparison to the prior best-in-class model in **Table 2**. None of the clips used in cross-validation testing were used during training of the corresponding model.

	Metric	Mean ± Standard Deviation	Minimum - Maximum
Our Model	Accuracy	95.94% ± 0.75%	94.23% – 97.58%
	AUC	0.988 ± 0.004	0.976 - 0.997
	Sensitivity	0.963 ± 0.012	0.938 - 0.990
	Specificity	0.956 ± 0.015	0.919 - 0.981
Yang et al.	Sensitivity	0.88 ± 0.05	-
	Specificity	0.92 ± 0.02	-

Table 2: Summary of test evaluation metrics achieved by models at site A during the cross-validation process and comparison to prior best in class 2D-CNN Long Short Term Memory (LSTM) model by Yang et al. (10).

3.2 Full-Video Classification Performance

The best performing models from the LOPO cross-validation process were used to sequentially classify all video clips from each of the 49 EMU recordings from site A without balancing the data to better emulate clinical use. Summary statistics for model performance across the recordings are shown **Table 3**. Across the videos, the average accuracy was **93.82% ± 6.91%** and the average AUC was **0.990 ± 0.017**. Metrics for individual videos are shown in **Figure 4**. Visualizations of the model predictions throughout representative recordings are shown in **Figure 5**. Classification accuracy did not vary significantly with the relative duration of seizure

(R^2 = -0.043, p-value = 0.770), total video length (R^2 = 0.076, p-value = 0.606) as measured by the Pearson correlation coefficient. Classification accuracy was also not significantly different between RGB vs Greyscale video (p-value = 0.285) as measured by the Mann-Whitney U test. Average detection latency for the model was 0.61s ± 2.01 with a median detection latency of 0.00s.

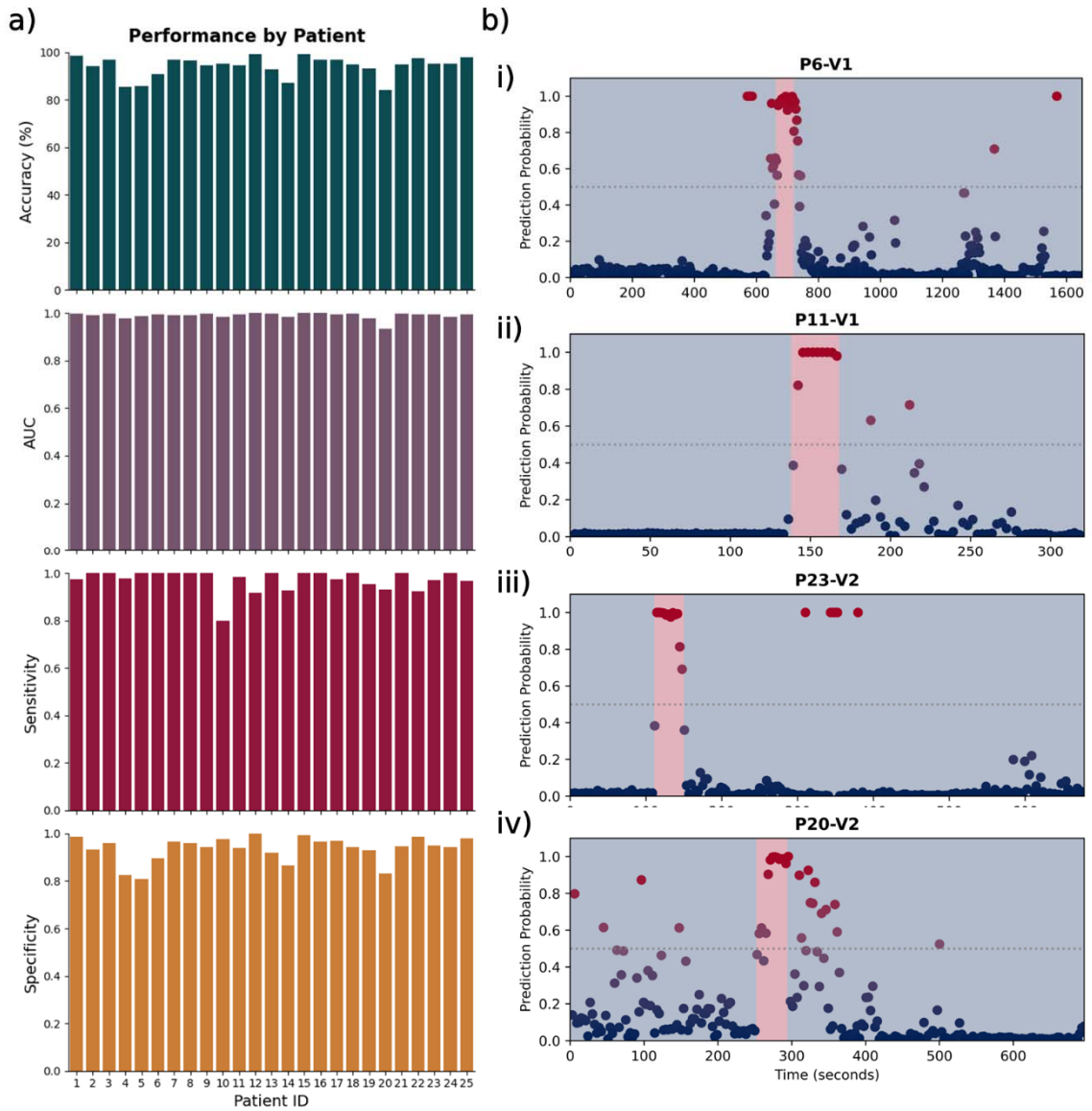

Figure 3: **a)** Performance of models from site A for each patient. The model with the highest AUC score from each cross-validation fold was used to classify the full EMU videos from the fold's corresponding patient. For example, the first model trained on the data from patients 2-25, then classified all video clips comprising the three EMU videos from patient 1 (P1-V1, P1-V2, and P1-V3). **b)** Visualization of model performance across representative EMU videos at site A. Blue background color represents true negative (non-seizure) while red background color represents true positive (seizure). The position and color of the points overlayed indicate model prediction at the corresponding timepoint in the video as a probability between 0 (non-seizure, blue) and 1 (seizure, red). (i)-(iii) demonstrate the robust sensitive and specific seizure detection capabilities of the model across videos

of varying lengths. Performance does not vary significantly across nocturnal greyscale video in (i) vs daytime color video in (ii, iii). The model obtained the worst AUC score on video (iv).

3.3 Generalizability of Architecture and Training Scheme

To validate the model architecture and training scheme on an independent data set, the untrained I3D model and code for training and LOPO cross-validation were sent to site B. The results from the cross-validation process are available in **Supplementary File 1**. This site B model was then used to sequentially classify all 19 EMU videos from site B with a high accuracy (**96.44% ± 3.42%**) and AUC (**0.984 ± 0.049)**. Summary statistics for the site B model are listed in **Table 3**. The site B model did not significantly deviate from the accuracy (p-value = 0.449) or AUC score (p-value = 0.401) of the model trained and tested at site A as determined by the Kolmogorov-Smirnov test.

3.4 Cross Institutional Testing:

To investigate the generalizability of already trained models on novel datasets with different environmental conditions, the model with the best performance during cross-validation at site A was sent to site B and used to classify the 19 EMU videos from site B. Likewise, the best performing model from site B was sent to classify the 49 EMU videos from site A. The mean decrease in AUC score from same-site to cross-site testing was -0.156. Average metrics from both trials are listed in **Table 3**.

4. DISCUSSION

Our results demonstrate that video-only seizure detection using a spatiotemporal 3-D CNN model is feasible and highly accurate. The I3D architecture, initially developed for human action recognition, is highly effective in detecting tonic-clonic seizures using only video data. Furthermore, our results show the efficacy of using transfer learning between different clinical sites to fine-tune a large, pretrained activity classification model for clinical tasks with only a modest sized dataset. This finding builds upon previous work showing that video-based methods can robustly capture and analyze epileptic seizures (14).

To the best of our knowledge, our model's performance for detecting tonic-clonic seizures exceeds all reported video-only methods, including the prior best in class 2D-CNN LSTM model by Yang et al. (10). Like Yang et al., we employed LOPO cross-validation, a method shown to be a more reliable estimator of model performance compared to k-fold cross-validation that does not respect patient boundaries (21). This approach ensures that the model's performance metrics are not inflated by data leakage and provides a more realistic assessment of the model's generalizability to new patients.

One of the important contributions of this work is the demonstration of the model's generalizability across different clinical environments. The model architecture and training pipeline developed at Site A yielded similarly high performance when used to train and evaluate models at site B. This finding suggests that our pipeline can be effectively deployed across different EMUs with relatively little site-specific training data. In the case of site B, less than 4

hours of training data were required to reach comparable performance to the original site, which used approximately 11 hours of data for training. This aspect of our study is crucial for practical implementation, as it implies that the model can be scaled and adapted to new clinical settings with minimal effort, a key consideration for widespread adoption (22).

		Site A Model – Site A Data		Site B Model – Site B Data	
		Mean ± Standard Deviation	*Minimum - Maximum*	*Mean ± Standard Deviation*	*Minimum - Maximum*
Same Site Evaluation	**Accuracy**	93.82% ± 6.91%	69.42% - 100%	96.44% ± 3.42%	85.79% – 99.48%
	AUC	0.990 ± 0.016	0.901 - 1.000	0.984 ± 0.049	0.782 – 1.000
	Sensitivity	0.969 ± 0.056	0.750 - 1.000	0.847 ± 0.207	0.177 – 1.000
	Specificity	0.933 ± 0.080	0.622 - 1.000	0.984 ± 0.029	0.909 – 1.000
		Site A Model – Site B Data		**Site B Model – Site A Data**	
		Mean ± Standard Deviation	*Minimum - Maximum*	*Mean ± Standard Deviation*	*Minimum - Maximum*
Cross Site Evaluation	**Accuracy**	84.02% ± 10.12%	56.21% – 95.69%	91.76% ± 5.57%	74.60% – 98.67%
	AUC	0.844 ± 0.130	0.581 – 0.994	0.818 ± 0.143	0.270 – 0.994
	Sensitivity	0.624 ± 0.251	0.156 – 1.000	0.323 ± 0.238	0.000 – 0.931
	Specificity	0.908 ± 0.095	0.725 – 1.000	0.986 ± 0.029	0.847 – 1.000

Table 3: Summary of test evaluation metrics achieved by the models at both sites when classifying full videos. Same site evaluation metrics represent performance of models from site A classifying site A video and models from site B classifying site B video. The models at each site were evaluated only on videos not seen in the training process. The models were then transferred and used to evaluate EMU videos without any further training at the other institution. This performance is detailed in the Cross site evaluation metrics.

Though the underlying architecture transferred well between institutions, the model trained at site A failed to fully generalize to site B data and vice versa. In particular, the drop in AUC of about 0.15 and 0.17 suggest that models were not able to fully generalize to new features in the EMU environment not seen at the other institution. Nonetheless, the models performed significantly better than chance, and it remains unclear if more training data or multi-institutional datasets could close the gap between same-institution and cross-institution performance. Federated learning, a technique designed to allow for model training on multi-institutional data without the need to centralize videos in one location, is a promising solution to address legal and privacy concerns associated with sharing protected health information and facilitate multi-institutional collaboration (23).

The high video processing speed exceeding 1,800 frames per second and consistently low detection latency mark a notable advancement for video-based seizure detection using neural-network based modes. Though the current model pipeline does not include features for real-time detection from streamed video data, these processing speeds suggest that real-time detection from one or more video feeds could be possible. Moreover, the lack of pre-processing requirements, such as cropping or other modifications to the video data, minimizes variations between institutions and reduces processing time, further enhancing the model's potential applicability in real-time scenarios such as the EMU or home seizure monitoring.

This work focused only on tonic-clonic seizure detection, and it is likely that the model would not perform as well for seizures with limited movement such as purely tonic, absence, or focal without progression to tonic-clonic. Future work including a broader array of seizure types during training could potentially address this limitation in our model. Training and validation on home seizure detection setups with variable layout and quality is also an essential next step for a truly general-purpose model. Other areas for future work include clinical trials for the implementation and evaluation for real-time, I3D model based seizure detection in the EMU. Integration of audio data into model designs, as done by several existing investigational and commercial seizure detection algorithms, may also increase overall model performance and generalizability (24).

5. CONCLUSION

Our study underscores the feasibility of using a spatiotemporal 3-D CNN model for video-only detection of tonic-clonic seizures with high accuracy and speed. The model's high processing speed, minimal pre-processing requirements, and robust performance across different clinical environments make it a potentially valuable tool for real-time seizure monitoring. Future enhancements, including larger and more varied datasets, the integration of audio data and the use of federated learning, can further augment the model's capabilities and extend its applicability in diverse clinical settings. By leveraging advanced computer vision techniques, we can significantly improve the accuracy and efficiency of seizure detection, ultimately enhancing patient care and clinical outcomes in epilepsy management.

6. REFERENCES

1. Baumgartner C, Koren JP, Rothmayer M. Automatic Computer-Based Detection of Epileptic Seizures. Front Neurol. 2018;9:639.

2. Albarrak A. Challenges and Prospects in Epilepsy Monitoring Units: A Comprehensive Review of Logistic Barriers. Cureus. 16(5):e59559.

3. Bhagubai M, Vandecasteele K, Swinnen L, Macea J, Chatzichristos C, De Vos M, et al. The Power of ECG in Semi-Automated Seizure Detection in Addition to Two-Channel behind-the-Ear EEG. Bioeng Basel Switz. 2023 Apr 20;10(4):491.

4. Milosevic M, Van de Vel A, Bonroy B, Ceulemans B, Lagae L, Vanrumste B, et al. Automated Detection of Tonic-Clonic Seizures Using 3-D Accelerometry and Surface Electromyography in Pediatric Patients. IEEE J Biomed Health Inform. 2016 Sep;20(5):1333–41.

5. van Andel J, Thijs RD, de Weerd A, Arends J, Leijten F. Non-EEG based ambulatory seizure detection designed for home use: What is available and how will it influence epilepsy care? Epilepsy Behav. 2016 Apr 1;57:82–9.

6. Karayiannis NB, Tao G, Xiong Y, Sami A, Varughese B, Frost Jr. JD, et al. Computerized Motion Analysis of Videotaped Neonatal Seizures of Epileptic Origin. Epilepsia. 2005;46(6):901–17.

7. Cuppens K, Chih-Wei Chen, Wong KBY, Van De Vel A, Lagae L, Ceulemans B, et al. Using spatio-temporal interest points (STIP) for myoclonic jerk detection in nocturnal video. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. San Diego, CA: IEEE; 2012 [cited 2024 Jan 7]. p. 4454–7. Available from: http://ieeexplore.ieee.org/document/6346955/

8. Li J, Zhen X, Liu X, Ouyang G. Classifying Normal and Abnormal Status Based on Video Recordings of Epileptic Patients. Sci World J. 2014 Apr 8;2014:e459636.

9. Karácsony T, Loesch-Biffar AM, Vollmar C, Rémi J, Noachtar S, Cunha JPS. Novel 3D video action recognition deep learning approach for near real time epileptic seizure classification. Sci Rep. 2022 Nov 15;12(1):19571.

10. Yang Y, Sarkis RA, Atrache RE, Loddenkemper T, Meisel C. Video-Based Detection of Generalized Tonic-Clonic Seizures Using Deep Learning. IEEE J Biomed Health Inform. 2021 Aug;25(8):2997–3008.

11. Moro M, Pastore VP, Marchesi G, Proserpio P, Tassi L, Castelnovo A, et al. Automatic video analysis and classification of sleep-related hypermotor seizures and disorders of arousal. Epilepsia. 2023;64(6):1653–62.

12. Achilles F, Tombari F, Belagiannis V, Loesch AM, Noachtar S, Navab N. Convolutional neural networks for real-time epileptic seizure detection. Comput Methods Biomech Biomed Eng Imaging Vis. 2018 May 4;6(3):264–9.

13. Pérez-García F, Scott C, Sparks R, Diehl B, Ourselin S. Transfer Learning of Deep Spatiotemporal Networks to Model Arbitrarily Long Videos of Seizures. In 2021 [cited 2024 Jan 9]. p. 334–44. Available from: http://arxiv.org/abs/2106.12014

14. Brown BM, Boyne AMH, Hassan AM, Allam AK, Cotton RJ, Haneef Z. Computer vision for automated seizure detection and classification: A systematic review. Epilepsia. 2024;65(5):1176–202.

15. Mooijman P, Catal C, Tekinerdogan B, Lommen A, Blokland M. The effects of data balancing approaches: A case study. Appl Soft Comput. 2023 Jan 1;132:109853.

16. Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In 2017 [cited 2024 Jul 10]. p. 6299–308. Available from: https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html

17. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The Kinetics Human Action Video Dataset [Internet]. arXiv; 2017 [cited 2024 Jul 23]. Available from: http://arxiv.org/abs/1705.06950

18. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition [Internet]. 2009 [cited 2024 Jul 23]. p. 248–55. Available from: https://ieeexplore.ieee.org/document/5206848

19. MMAciton2 Contributors. Action Recognition Models [Internet]. 2020. Available from: https://mmaction2.readthedocs.io/en/latest/model_zoo/recognition.html#I3D

20. MMAction2 Contributors. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark [Internet]. 2020 [cited 2024 Jul 10]. Available from: https://github.com/open-mmlab/mmaction2

21. Ahmedt-Aristizabal D, Fookes C, Nguyen K, Denman S, Sridharan S, Dionisio S. Deep facial analysis: A new phase I epilepsy evaluation using computer vision. Epilepsy Behav. 2018 May;82:17–24.

22. Pediaditis M, Tsiknakis M, Leitgeb N. Vision-based motion detection, analysis and recognition of epileptic seizures—A systematic review. Comput Methods Programs Biomed. 2012 Dec 1;108(3):1133–48.

23. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. arXiv.org. 2016 [cited 2024 Oct 5]. Communication-Efficient Learning of Deep Networks from Decentralized Data. Available from: https://arxiv.org/abs/1602.05629v4

24. Automated detection of nocturnal motor seizures using an audio-video system. Brain Behav. 2022;12(9):e2737.