Epileptic Seizure Classification with Patient-level and Video-level Contrastive Pretraining

Chin-Jou Li[1], Chien-Chen Chou[2], Yen-Cheng Shih[2], Li-Chuan Kuo[1], Yu-Te Wang[3*], Aileen McGonigal[4*], Hsiang-Yu Yu[2*], Jen-Cheng Hou[3*], Yu Tsao[3*†]
[1]National Taiwan University, Taiwan
[2]Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[4]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia
[*]Equal advising [†]Corresponding author: yu.tsao@citi.sinica.edu.tw

***Abstract*—Accurate classification of epileptic seizure types through seizure semiology analysis demands significant clinical expertise. While previous studies have employed various action recognition modules, the scarcity of labeled clinical videos has hindered the deployment of larger models. In this study, we explore unlabeled data to pretrain a transformer-based model with contrastive loss, taking advantage of the information that circumvents the need for additional annotation from medical professionals. We maximize the similarity between embeddings from the same patient and video while minimizing those from different patients and videos. Subsequently, a classification head was finetuned to distinguishing temporal lobe epilepsy (TLE) and extratemporal lobe epilepsy (exTLE), achieving a 5-fold accuracy of 0.93 and an F1 score of 0.88 on the video level (N = 57). Our results outperformed other state-of-the-art seizure classification models, demonstrating the efficacy of our approach. This suggests potential applications in clinical practice, where unlabeled data could serve as a valuable aid in improving seizure classification accuracy and patient care.**

***Index Terms*—epilepsy, seizure classification, contrastive learning, action recognition**

I. Introduction

Epilepsy is a prevalent chronic neurological condition marked by abnormal brain activity and recurrent seizures, affecting approximately 50 million individuals. A main clinical goal is to correctly localize the likely region of cerebral organization, which is particularly important for individuals with focal drug-resistant epilepsy, who may benefit from epilepsy surgery with a view to controlling seizures [1]. Clinicians wish to distinguish between temporal lobe epilepsy (TLE) and extra-temporal epilepsy (exTLE), since these involve different patterns of seizure organization that necessitate different presurgical evaluation strategies for correct brain localization [2]. In addition, TLE and exTLE may have differences in surgical prognosis, especially if brain imaging does not show an underlying lesion [3], [4]. Thus, advances in deep learning-based seizure video analyses that can accurately discriminate between TLE and exTLE need to be further developed, amongst key challenges for the neuro-engineering scientific community [5].

However, labeling medical datasets poses a challenge in adopting newer computer vision techniques. In contrast to general datasets, which can be annotated through crowd-sourcing without domain knowledge, medical data lacks this feasibility. Despite the increasing accessibility of cameras for clinical recordings, the high cost of building large datasets hinders applying state-of-the-art computer vision models, which often require substantial data for training. Hence, leveraging "free" labels, such as the timestamp of a medical image, relevant body parts, or patient information, becomes crucial for enhancing performance. These naturally provided labels, incurring no additional cost, serve as valuable resources for model training, particularly through contrastive learning.

This study aims to enhance the epileptic seizure video classification of TLE and exTLE. By utilizing unlabeled videos directly from the epilepsy monitoring unit, we explore the applicability of more advanced model architectures through contrastive pretraining. We argue that utilizing less expensive unlabeled data enhances seizure video classification, as it provides more information about implicit semiology through capturing additional seizure episodes in similar environments.

Firstly, inspired by Singh *et al.* [9], which implemented contrastive learning for action recognition tasks on two different levels, we developed a contrastive pretraining strategy where the loss function operates at both patient and video levels. The goal is to emphasize the consistency of human behavior at video and patient levels.

Secondly, we pretrained a transformer-based model to extract embeddings from video clips. Compared to previous models based on convolutional neural networks (CNNs), transformers encode longer clips, extending to 10 seconds in our case. Our research revealed that incorporating more temporal information into each segment surpasses the efficacy of previous methods, which relied on recurrent neural networks (RNNs) to aggregate temporal information from short clips.

Our results shed light on the potential of contrastive pretraining using unlabeled data. Video-level contrastive loss enhances the continuity of embedding from the same video, while patient-level contrastive loss contributes to identifying consistent semiology from the same patient. With the proposed pretraining objective applied to a transformer-based model, competitive results for seizure classification can be achieved by adding a finetuned classification head.

II. Related Work

A. Epilepetic Seizure Video Classification

Researchers have employed computer vision modules for seizure detection, type prediction, and cerebral localization [5]. Moro *et al.* [6] utilized 3D-CNN to differentiate between non-hyperkinetic seizures and sleep-related paroxysmal events. Hou *et al.* [8] proposed a multi-stream framework using keypoints and appearance from body and face to classify epileptic seizures and psychogenic non-epileptic seizures. Aristizabal *et al.* [7] proposed a system capable of computing motion signatures including body, face, and hand semiology. Karácsony *et al.* [11] utilized CNN-based models to extract video features followed by an RNN to differentiate epileptic seizures in frontal lobe epilepsy, temporal lobe epilepsy, and non-epileptic events. Similarly, Pérez-García *et al.* [12] used a pretrained spatiotemporal CNN and RNN to classify seizures into focal onset seizures and focal to bilateral tonic-clonic seizures.

B. Contrastive Learning

Contrastive learning has been shown to effectively provide supplementary information for guiding classification tasks [13]. This approach involves pulling positive pairs closer and negative pairs farther apart within the same batch. Singh *et al.* [9] extended this concept to semi-supervised training on both instance and group levels, enhancing the performance of a general-case action recognition on a ResNet-18 backbone. In medical imaging, including X-ray and Magnetic Resonance Imaging (MRI) datasets [14], [15], [16], contrastive loss has demonstrated success. Notably, in medical video tasks, Kumar *et al.* [17] implemented frame-level self-supervised learning, highlighting the applicability and effectiveness of contrastive loss in medical video analysis.

III. Methodology

A. Problem Setup

Our task involves a portion of labeled (D_l) dataset and the majority unlabeled (D_u) in terms of epilepsy type. Videos were segmented into non-overlapping clips, each lasting 10 seconds. Clips in D_u, denoted as c_u, come with the information for the respective patient (p) and video (v). For the labeled set D_l, classification labels (y) are provided for each clip. Therefore, $D_u = \{c_{u_i}, p_i, v_i\}_{i=1}^{M}$, and $D_l = \{c_{l_i}, p_i, v_i, y_i\}_{i=1}^{N}$ where M, N represent the total number of clips in each set.

Our model, depicted in Fig. 1, comprises a transformer-based pretrained model, denoted as Θ, and a subsequent classification head, denoted as H. D_u is used for pretraining and D_l for training.

B. Contrastive Pretraining

Contrastive loss involves data defined as different types, forming positive pairs (c_+) from the same type and negative pairs (c_k) from different types. Each positive pair is selected once to calculate the contrastive loss. The distance between two clips' embeddings is computed using cosine similarity, regulated by a temperature hyperparameter τ. Fig. 2 illustrates the two losses in our loss function.

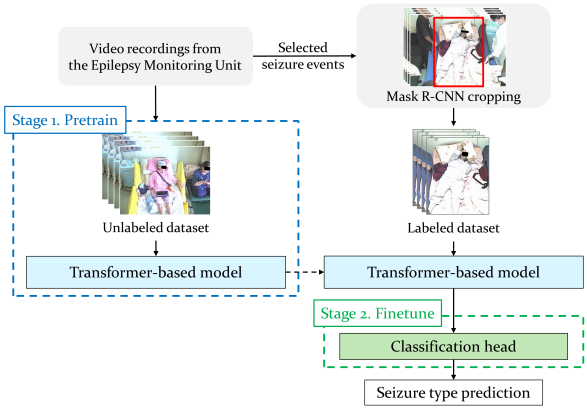

Fig. 1. Overview of our model structure. A transformer-based model first undergoes contrastive pretraining on the unlabeled dataset, and then a classification head is added on top for classifying the labeled dataset.

1) Video-level Contrastive Loss: Assuming that clips from the same seizure recording are similar, a video-level contrastive loss is applied. In a batch of clips from D_u, the video-level contrastive loss ($\mathcal{L}_{vc}$) aims to encourage the similarity between embeddings from the same video's clips and separate embeddings from different videos. A pair of clips (c_i, c_j) forms a positive pair if they are from the same video ($v_i = v_j$), otherwise it forms a negative pair. Specifically, for a positive pair selected among $V_+ = \{(c_i, c_+)\}$ where $v_+ = v_i$, its corresponding negative pairs are $V_k = \{(c_i, c_k)\}$ where $v_k \neq v_i$. The loss is expressed as given in (1):

$$\mathcal{L}_{vc}(c_i, c_+) = -\log \frac{\exp(\Theta(c_i) \cdot \Theta(c_+)/\tau)}{\sum_{k=0}^{|V_k|} \exp(\Theta(c_i) \cdot \Theta(c_k)/\tau)} \quad (1)$$

2) Patient-level Contrastive Loss: As videos from the same patient belong to the same epilepsy type, patient-level contrastive loss is employed to guide pretraining from a higher level. Similar to the video-level loss, embeddings of clips from the same patient are pulled closer, while others are pushed farther away. In a batch of clips from D_u, the patient-level contrastive loss ($\mathcal{L}_{pc}$) uses positive pairs (c_i, c_j) where the two clips are from the same patient (i.e., $p_i = p_j$); negative pairs are formed otherwise. For a positive pair selected among $P_+ = \{(c_i, c_+)\}$ where $p_+ = p_i$, its corresponding negative pairs are $P_k = \{(c_i, c_k)\}$ where $p_k \neq p_i$. The loss is expressed as given in (2):

$$\mathcal{L}_{pc}(c_i, c_+) = -\log \frac{\exp(\Theta(c_i) \cdot \Theta(c_+)/\tau)}{\sum_{k=0}^{|P_k|} \exp(\Theta(c_i) \cdot \Theta(c_k)/\tau)} \quad (2)$$

The overall pretraining loss function is then defined as (3):

$$\mathcal{L}_{prtn} = \mathcal{L}_{vc} + \mathcal{L}_{pc} \quad (3)$$

C. Finetuning for Epilepsy Type Classification

After pretraining, the weight of the transformer-based model is frozen for the finetuning step. A classification head (H) is added on top. It is optimized with standard cross entropy loss

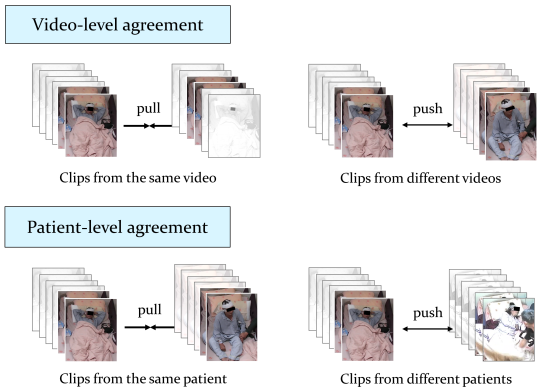

Fig. 2. Demonstration of two contrastive losses. Video-level agreement is implemented through video-level contrastive loss, where clips from the same video form positive pairs and those from different videos form negative pairs. Similarly, patient-level contrastive loss aims for patient-level agreement, pulling the embeddings of clips from the same patient closer while pushing others away.

($\mathcal{L}_{CE}$) for our two-class classification. The loss is formulated as given in (4):

$$\mathcal{L}_{CE} = -\sum_{i=0}^{K} [y_i \log p_{i_0} + (1 - y_i) \log p_{i_1}] \tag{4}$$

Here, K stands for the batch size, while p_{i_j} represents the predicted probability for clip c_i belonging to class j, which came from from $H(\Theta(c_{l_i}))$.

IV. Experiment

A. Data Collection

Clinical seizure videos were gathered from the Epilepsy Monitoring Unit at the Department of Neurology, Taipei Veterans General Hospital, Neurological Institute, Taiwan. Participants were selected from those who underwent elective video-electroencephalography (VEEG), adhering to standard clinical protocols for seizure assessment. Ethics approval was obtained from the institutional review board (IRB number: 2022-07041BC), and participants provided informed consent.

For the unlabeled dataset, we gathered 731 videos capturing both daily activities and seizure onsets from 40 patients, amounting to approximately 22 hours of recordings. Subsequently, a labeled dataset was created as a subset of the unlabeled dataset, encompassing 57 videos from 15 patients. This labeled dataset included two seizure types: temporal lobe epilepsy (TLE) and extratemporal lobe epilepsy (exTLE); 42 videos belong to TLE and 15 belong to exTLE. To align with the pretrain model's input specifications, we segmented each video into multiple 10-second clips. The labeled dataset underwent additional Mask R-CNN [18] cropping to remove surroundings by using the hospital bed's location in the first frame. The resulting unlabeled and labeled datasets comprised 8985 clips and 241 clips respectively. A visual representation of our data collection method is shown in Fig. 1.

B. Implementation Details

1) Pretraining: We chose TimeSformer [10] as our transformer-based action recognition model. It replaced the convolution operator with self-attention, learning spatiotemporal features directly from a sequence of frame-level patches, therefore faster to train and can be applied to longer video clips. Specifically, we employed TimeSformer-L featuring a 96-frame input and a spatial crop of 224×224 pixels. We excluded the final linear prediction layer, using 768-dimensional vectors from the transformer as the clip embeddings. The model weight was initialized with a given checkpoint pretrained on Kinetics400 [21]. The entire unlabeled dataset D_u was used, requiring 4 GPUs to achieve a batch size of 8. Each training epoch took approximately 1.5 to 2 hours. The pretraining process was conducted for 5 epochs.

2) Finetuning: We implemented a classification head with multiple linear layers to progressively reduce dimensions from 768 to 2. The softmax operation is conducted before calculating the cross entropy loss for epilepsy classification. The classification was performed with a 5-fold cross-validation by patient, each fold trained for 10 epochs. Video-level results were computed by voting the clip-level prediction.

3) Baselines: Three alternative models, as mentioned in related work, were reproduced to compare with our approach. They all share a common structure, featuring a CNN-based feature extractor coupled with an RNN. The first baseline is from Karácsony *et al.* [11]. The second baseline substitutes the first baseline's I3D model with an R(2+1)D module. The third baseline is developed by Pérez-García *et al.* [12],

C. Performance Comparison

Table I presents the performance scores of the evaluated models. Our best model outperformed others with an increase of at least 13% in accuracy and 17% in F1 score. Achieving an accuracy of 0.93 and an F1 score of 0.88, our approach effectively captures essential features for epilepsy classification. Unlike previous models, our approach undergoes pretraining with unlabeled data, which benefits the downstream task. These results show that pretraining not only enhances overall effectiveness but also improves the efficiency of finetuning on the labeled dataset.

Additionally, our results stand out without utilizing RNN or other modules to aggregate clip embeddings. This shows the capability of transformer-based models to capture spatial and temporal information. While previous studies relied on CNNs for feature extraction, their limitations arise when dealing with long clips, and the temporal relationship is confined to adjacent frames. Even when combined with an RNN, there's a risk of information loss if not adequately extracted by the CNN. In contrast, the transformer structure facilitates the utilization of longer clips and offers greater flexibility in exploring spatial-temporal information.

D. Ablation Study

To assess the utility of contrastive learning, we conducted an ablation study, the results of which are summarized in Table I.

TABLE I
PERFORMANCE OF VIDEO-LEVEL CLASSIFICATION

Model	Acc.	F1
Karácsony *et al.* [11] (I3D)	0.796	0.708
Karácsony *et al.* [11] (R(2+1)D)	0.664	0.587
Pérez-García *et al.* [12]	0.771	0.667
Ours		
- w/o pretraining	0.825	0.750
- Video	0.860	0.790
- Patient	0.912	0.857
- **Video+patient**	**0.930**	**0.882**

Video-level classification scores of baseline models and our model. The best scores are indicated in bold for each metric. The first three rows represent previous methods constructed and trained on our labeled dataset. The last row shows the result of our default approach using contrastive pretraining on both video and patient levels, while the rest present different configurations for the ablation study.

Four configurations were adopted: no pretraining, video-level contrastive loss only ($\mathcal{L}vc$), patient-level contrastive loss only ($\mathcal{L}pc$), and the full pretraining objective ($\mathcal{L}_{prtn}$).

As shown in Table I, it illustrates a discernible trend of increasing scores across different scenarios. The model without any pretraining objective has a decent performance outperforming the baseline models. If the video/patient-level contrastive loss was included in pretraining, a performance boost could be obtained in accuracy (+3%, +8%) and F1 score (+4%, +10%). This demonstrates how encouraging the consistency of embeddings from the same video and semiology from the same patient in pretraining could be helpful for the downstream task. Adopting both losses gives a further improvement in accuracy (+10%) and F1 score (+13%).

V. CONCLUSION

This study presents a novel approach for classifying epileptic seizures by pretraining a transformer-based model with unlabeled clinical videos. The pretraining is conducted based on contrastive learning by emphasizing the consistency of human behavior at video and patient levels. The ablation study underscores the effectiveness of each contrastive loss at both levels. Our model demonstrates competitive accuracy and F1 score when compared to other state-of-the-art models.

VI. ACKNOWLEDGEMENTS

The authors would like to acknowledge the funding from Taipei Veterans General Hospital (V112C-143).

REFERENCES

[1] Thijs, R. D., Surges, R., O'Brien, T. J., and Sander, J. W., "Epilepsy in adults," *The Lancet*, vol. 393, no. 10172, pp. 689-701, 2019.

[2] Jehi, L., Friedman, D., Carlson, C., Cascino, G., Dewar, S., Elger, C., Engel, J., Jr, Knowlton, R., Kuzniecky, R., McIntosh, A., O'Brien, T. J., Spencer, D., Sperling, M. R., Worrell, G., Bingaman, B., Gonzalez-Martinez, J., Doyle, W., and French, J., "The evolution of epilepsy surgery between 1991 and 2011 in nine major epilepsy centers across the United States, Germany, and Australia," *Epilepsia*, vol. 56, no. 10, pp. 1526-1533, 2015.

[3] Yu, H. Y., Lin, C. F., Chou, C. C., Lu, Y. J., Hsu, S. P., Lee, C. C., and Chen, C., "Outcomes of hippocampus-sparing lesionectomy for temporal lobe epilepsy and the significance of intraoperative hippocampography," *Clinical Neurophysiology*, vol. 132, no. 3, pp. 746-755, 2021.

[4] Noe, K., Sulc, V., Wong-Kisiel, L., Wirrell, E., Van Gompel, J. J., Wetjen, N., Britton, J., So, E., Cascino, G. D., Marsh, W. R., Meyer, F., Horinek, D., Giannini, C., Watson, R., Brinkmann, B. H., Stead, M., and Worrell, G. A., "Long-term outcomes after nonlesional extratemporal lobe epilepsy surgery," *JAMA neurology*, vol. 70, no. 8, pp. 1003-1008, 2013.

[5] Ahmedt-Aristizabal, D., Armin, M. A., Hayder, Z., Garcia-Cairasco, N., Petersson, L., Fookes, C., Denman, S., and McGonigal, A., "Deep learning approaches for seizure video analysis: a review," *arXiv preprint*, arXiv:2312.10930, 2023.

[6] Moro, M., Pastore, V.P., Marchesi, G., Proserpio, P., Tassi, L., Castelnovo, A., Manconi, M., Nobile, G., Cordani, R., Gibbs, S.A. and Odone, F., "Automatic Video Analysis and Classification of Sleep-related Hypermotor seizures and Disorders of Arousal," *Epilepsia*, 2023.

[7] Ahmedt-Aristizabal, D., Sarfraz, M.S., Denman, S., Nguyen, K., Fookes, C., Dionisio, S., and Stiefelhagen, R., "Motion signatures for the analysis of seizure evolution in epilepsy," in *Proc. EMBC*, pp. 2099–2105, 2019.

[8] Hou, J.C., McGonigal, A., Bartolomei,, F. and Thonnat, M., "A multi-stream approach for seizure classification with knowledge distillation," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1-8, 2021.

[9] Singh, A., Chakraborty, O., Varshney, A., Panda, R., Feris, R., Saenko, K. and Das, A., "Semi-supervised action recognition with temporal contrastive learning," in *Proc. CVPR*, pp. 10389-10399, 2021.

[10] Bertasius, G., Wang, H. and Torresani, L., "Is space-time attention all you need for video understanding?" in *Proc. ICML*, vol. 2, no. 3, p. 4, 2021.

[11] Karácsony, T., Loesch-Biffar, A.M., Vollmar, C., Rémi, J., Noachtar, S. and Cunha, J.P.S., "Novel 3D video action recognition deep learning approach for near real time epileptic seizure classification," *Scientific Reports*, vol. 12, no. 1, p.19571, 2022.

[12] Pérez-García, F., Scott, C., Sparks, R., Diehl, B. and Ourselin, S., "Transfer learning of deep spatiotemporal networks to model arbitrarily long videos of seizures," in *Proc. MICCAI*, pp. 334-344, 2021

[13] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, pp. 1597-1607, 2020.

[14] Zhang, Y., Jiang, H., Miura, Y., Manning, C,.D. and Langlotz, C.P., "Contrastive learning of medical visual representations from paired images and text," *Machine Learning for Healthcare*, pp. 2-25, 2022.

[15] Chaitanya, K., Erdil, E., Karani, N. and Konukoglu, E., "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. NeurIPS*, vol. 33, pp.12546-12558, 2020.

[16] Peng, J., Wang, P., Desrosiers, C. and Pedersoli, M., "Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels," in *Proc. NeurIPS*, vol. 34, pp.16686-16699, 2021.

[17] Kumar, V., Tripathi, V., Pant, B., Alshamrani, S.S., Dumka, A., Gehlot, A., Singh, R., Rashid, M., Alshehri, A. and AlGhamdi, A.S., "Hybrid spatiotemporal contrastive representation learning for content-based surgical video retrieval," *Electronics*, vol. 11, no. 9, p.1353, 2022.

[18] He, K., Gkioxari, G., Dollár, P. and Girshick, R., "Mask r-cnn," in *Proc. ICCV*, pp. 2961-2969, 2017.

[19] Carreira, J. and Zisserman, A., "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. CVPR*, pp. 6299-6308, 2017.

[20] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M., "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, pp. 6450-6459, 2018.

[21] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P. and Suleyman, M., "The kinetics human action video dataset," *arXiv preprint*, arXiv:1705.06950, 2017.