A DEEP LEARNING ARCHITECTURE FOR EPILEPTIC SEIZURE CLASSIFICATION BASED ON OBJECT AND ACTION RECOGNITION

Tamás Karácsony[1] *Anna Mira Loesch-Biffar*[2] *Christian Vollmar*[2]
Soheyl Noachtar[2] *João Paulo Silva Cunha*[1,3,*]

[1] Center for Biomedical Engineering Research, INESC TEC, Porto, Portugal
[2] Epilepsy Center, Department of Neurology, University of Munich, Munich, Germany
[3] Faculty of Engineering (FEUP), University of Porto, Porto, Portugal

ABSTRACT

Epilepsy affects approximately 1% of the world's population. Semiology of epileptic seizures contain major clinical signs to classify epilepsy syndromes currently evaluated by epileptologists by simple visual inspection of video. There is a necessity to create automatic and semiautomatic methods for seizure detection and classification to better support patient monitoring management and diagnostic decisions. One of the current promising approaches are the marker-less computer-vision techniques. In this paper an end-to-end deep learning approach is proposed for binary classification of Frontal vs. Temporal Lobe Epilepsies based solely on seizure videos. The system utilizes infrared (IR) videos of the seizures as it is used 24/7 in hospitals' epilepsy monitoring units. The architecture employs transfer learning from large object detection "static" and human action recognition "dynamic" datasets such as ImageNet and Kinectics-400, to extract and classify the clinically known spatiotemporal features of seizures. The developed classification architecture achieves a 5-fold cross-validation f1-score of 0.844±0.042. This architecture has the potential to support physicians with diagnostic decisions and might be applied for online applications in epilepsy monitoring units. Furthermore, it may be jointly used in the near future with synchronized scene depth 3D information and EEG from the seizures.

Index Terms— Computer Vision, Deep learning, Action recognition, Epileptic seizure semiology, Diagnostic support

1. INTRODUCTION

Epilepsy is a neurological condition that affects people of all ages and has no geographic, social or racial boundaries. It affects more than 70 million people, nearly 1% of the population worldwide. Globally, an estimated 2.4 million people are diagnosed with epilepsy each year [1]. It is defined as "a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain" [2] causing, among other manifestations, uncoordinated movements. Seizure semiology is well established in the evaluation of epilepsy patients, especially when considering epilepsy surgery [3]. Thus, to classify epilepsy syndromes, motion semiology of epileptic seizures is a major source of clinical signs that are qualitatively evaluated through video-EEG systems in epilepsy monitoring units (EMUs) [4, 5]. For quantitative semiology characterization to support epilepsy diagnosis, automated and semi-automated computer-vision analysis approaches have been a promising methodology [6], but still depend on massive human interaction [7]. In EMUs it is advantageous to use infrared (IR) video streams to ensure 24/7 monitoring, including nights, when patients are asleep and no RGB video can be acquired [8, 9, 10, 11]. In this paper, an end-to-end deep learning approach is proposed for binary classification of Frontal (FLE) vs. Temporal Lobe Epilepsies (TLE) based solely on seizure motion extraction from IR video from our jointly collected NeuroKinect 3.0 database [12, 13]. The architecture employs transfer learning from large object detection, "static", ImageNet dataset [14] and human action recognition, "dynamic", Kinetics-400 dataset [15] to extract spatiotemporal features of seizures and classify them as FLE or TLE. An encouraging 5-fold cross-validation f1-score of 0.844±0.042 was obtained, indicating this approach performs better than previous methods reported in the literature.

Table 1: Comparison to other deep learning based publications in this domain. (ETLE - Extra Temporal Lobe Epilepsy, MTLE - Mesial Temporal Lobe Epilepsy)

Author	Classes	Performance	Notes
Achilles et al. [17, 18]	Seizure No seizure	AUC: 0.78	Single frame approach (posture recognition)
Ahmedt-Aristizabal et al. [19]	MTLE ETLE	Average accuracy: 53.39%-56.31%	Face body and hand inputs, very high std
Maia et al. [21]	TLE ETLE	AUC 0.65	Probably overfits
This work	**TLE FLE**	**f1-score: 0.844±0.042 AUC: 0.90±0.04**	-

2. RELATED WORK

There are a very limited number of studies utilizing deep learning approaches to video based seizure classification, in contrast to applications using EEG signals. Previous deep learning based approaches are summarized in Tab. 1. Our joint Munich-Porto research group [17, 18] has proposed a convolutional neural network (CNN) based epilepsy classification with IR and depth video input, however the method neglects the temporal information of the seizures. A hierarchical approach proposed by Ahmedt-Aristizabal et al. [19, 20] divides the processing into three main parallel threads. Despite attempting to extract the temporal information with LSTM layers it did not achieve outstanding classification due to the lack of data compared to the wide variety in semiological patterns. An other work

*Corresponding author: jcunha@ieee.org; This project was partially financed by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, through national funds, and co-funded by the FEDER, where applicable, and by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia within project POCI-01-0145-FEDER-028618 (PTDC/CCI-COM/28618/2017).

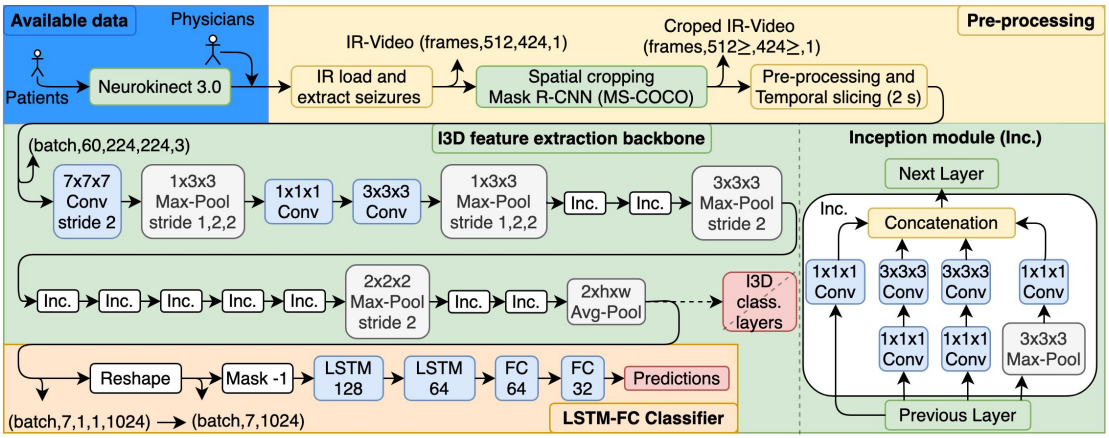


Fig. 1: Architecture pipeline, including the available data from Neurokinect 3.0 [13], the pre-processing algorithm (See section 3.3), the I3D feature extraction backbone (See section 3.4, [16]) and the LSTM-FC classifier (See section 3.5).

proposed from our group [21] utilizes the Neurokinect 3.0 IR video dataset as well. This approach is a combination of Inception-V3 feature extraction and fully connected classifier, which is an architecture designed for object recognition, therefore it also excludes temporal features of the seizure. Moreover, it probably fails to deal with the class imbalance and overfits to one class. Thus, following the DL current literature, previous related papers considered mostly spatial information with relatively modest performance levels. Based on the clinical knowledge that seizure type semiology is highly dependent on motion dynamics, we aim in the present paper to take advantage of spatiotemporal features in addition the habitual spatial (static) information DL approaches use. In a certain way, we performed a "clinical knowledge" transfer to the proposed architecture so that our DL pipeline better mimics the qualitative clinical knowledge accumulated by epileptologists that use motion dynamics to classify seizure types.

3. METHODS

3.1. Data acquisition

The dataset was acquired with the NeuroKinect 3.0 system implemented at the EMU of the University of Munich [12, 13]. This system is a three-bed Kinect v2 3Dvideo-EEG system developed for epileptic seizure monitoring. Kinect v2 acquires multiple streams of data namely, 1920x1080 HD-RGB, 512x424 infrared (IR) and depth videos, and 3D body joint information, with a sampling rate of 30 fps. During the night, no HD-RGB is available. When a seizure occurs the data is exported with a timestamp for posterior analysis and labeling by clinical professionals.

3.2. Extracted dataset

From the above described Neurokinect 3.0 database IR videos were extracted for this study, as the proposed pipeline is intended to be used for 24/7 (day and night) near-online monitoring. Two classes were defined, the first consists of Frontal Lobe Epilepsies (FLE), including left FLE and right FLE, and the other class is Temporal Lobe Epilepsies (TLE), including left TLE and right TLE. The main metrics of the extracted dataset are described in Tab. 2. It should be noted that the number of seizures is very imbalanced (85 FLE, 41 TLE) and needs to be addressed (see methods).

Table 2: Main metrics of the extracted IR videos dataset (FLE - Frontal Lobe Epilepsy, TLE - Temporal Lobe Epilepsy)

Class name	FLE	TLE	Total
Included seizures	FLE, right FLE, left FLE	TLE, right TLE, left TLE	FLE, TLE
Number of patients	20	15	35
Number of seizures	85	41	126
Tot clinical length [s]	2587	3116	5703
Average clinical length [s]	30.4	76.0	45.3
Minimal clinical length [s]	6.0	6.3	6.0
Maximal clinical length [s]	187.9	225.9	225.9
2 [s] samples	1282	1530	2812
Resolution	512x424 16bit		
Sampling frequency	30 fps		

3.3. Data pre-processing

3.3.1. Video spatial cropping

In order to minimize the unrelated information in the data so classification performance can be maximized, the seizure videos were automatically cropped and later reviewed for quality check, as follows. The first frame of each IR seizure video was segmented with Mask R-CNN [22] with a Keras [23] implementation [24]. The weights for the architecture were pre-trained on MS-COCO dataset, which includes bed and person classes [25]. This automated segmentation provides the bounding boxes of all persons and beds detected on the first frame. Then the bed and person bounding box with the highest confidence is selected and a merged bounding box is created. This resulting bounding box was expanded with +20% in all direction, which was then used to crop automatically the whole video sequence. A visual quality check of the first seconds of each 126 videos was performed (see results).

3.3.2. Pre-processing for feature extraction

To comply with the I3D input requirements [16] the cropped videos are converted from one channel gray (uint8) to RGB (uint8) representation. These are resized with preserving the aspect ratio to have the largest dimension of the frames as 224 pixels with bilinear interpolation and then padded to 224x224 pixels. The pixel values are rescaled between -1.0 and 1.0 (float32) [16].

3.3.3. Creating samples

Due to the class imbalance and the limitation of the available data the seizure videos were temporally sliced to 2[s] (60 frames) samples. The leftover sample, which was less than 60 frames, but more than 9 frames were still included. This resulted in 1282 and 1530 samples for FLE and TLE, respectively (Tab. 2).

3.4. Feature extraction

Features were extracted with a Keras implementation [26] of Inflated 3D Convnet (I3D) [16]. This I3D is a special network designed for human action recognition, especially suitable for spatio-temporal feature extraction of human movements. The architecture is based on Inception-V1 [27], with pre-training on ImageNet [14]. Then the I3D architecture was trained on the Kinetics-400 dataset [15], which consists of 400 human action classes and over 400 clips per class. For feature extraction the last classification layers were removed (Fig. 1). The 9-60 frame length samples were evaluated with the network and a 7 timestep feature vector with 1024 features per timestep was created. Samples with less than 60 frames, however yield less than 7 timesteps, thus these were pre-padded with dummy feature vectors to match the dimensions (7x1024), with a values of -1 for future masking.

3.5. Implemented classifier architecture

Classification was carried out with a long short-term memory (LSTM) classifier, to further exploit temporal features (Fig. 1). The first layer is a masking layer, which masks down the inputs with the dummy -1 value, thus these are not contributing to the optimization process. Two LSTM layers were utilized with 128 and 64 units. The output of the last LSTM layer was classified with two fully connected (FC) layers, with 64 and 32 units. As regularization layers batch normalization (BN) and dropout (DO) was used, with a dropout rate of 0.5. Dropout was applied after batch normalization to prevent variance shift when transferring the model from training to test state [28]. Recurrent dropout was applied with a dropout rate of 0.3. L2 regularization was applied on the kernels of the dense layers and LSTM layers. Furthermore, the LSTM layers recurrent regularization was also L2. Moreover, activity regularizer were applied for the LSTM layers and bias regularization for the FC layers. ReLU activation was used on the dense layers, except for the last binary classification layer with one unit, which used sigmoid activation. The kernels of the dense layers were initialized with He uniform initializer [29]. The full proposed system is illustrated on Fig. 1.

3.5.1. Training methods

Random hyperparameter optimization was used to determine the number of units in the LSTM [32,256] and FC [16,128] layers, the loss function [mean squared error, binary cross-entropy], the DO rate [0,0.5], the recurrent DO rate [0,0.5], the number of epochs [50,2000], and the batch size [50,2000]. The extracted dataset is imbalanced between the two classes, thus weighted mean squared error was used as loss function. It was optimized with Adam optimizer, using hyperparameters suggested by the original paper, such as learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ [30]. The architecture was trained for 2000 epochs with a batch size of 500 samples. To ensure generalization, the architecture with the highest f1 validation score was used, which can be interpreted as early stopping.

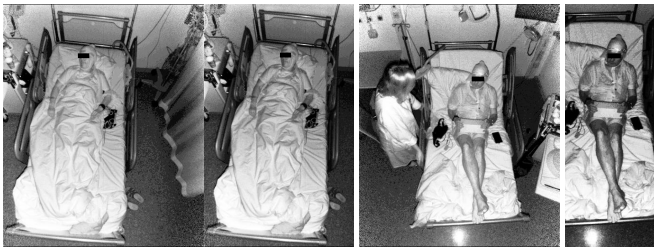
3.6. Cross-validation

In order to prevent data-leakage, such as non-seizure related subject specific features in training and validation set at the same time the seizures were grouped by subjects. Furthermore, it is beneficial to prevent the classifier to overfit on subject specific epilepsy related features also, thus increasing generalization of the classifier. A 5-fold cross-validation was performed. In each fold 16 FLE and 12 TLE subjects were in the training set and 4 FLE and 3 TLE subjects in the validation set. This sorting of the seizures however caused class imbalance, as the available number of seizures per patient varies a lot in the dataset. Thus, temporal slicing of seizures (sec. 3.3.3) and weighted class handling both in training (sec 3.5.1) and evaluation phase (sec. 4.2) were used to address this issue.

4. RESULTS

4.1. Mask R-CNN video cropping

The bounding box detection with Mask R-CNN was visually confirmed on the first seconds of the seizure videos, as there were no available predefined bounding boxes. The designed algorithm successfully detected and cropped the area of interest in 122/126, 96.83% of the cases (Fig. 2). It properly removed the background and unnecessary surrounding scenery, moreover the method performed well in spite of the complex scenery with occlusions, such as blanket or surrounding medical staff. Although the +20% expansion of the cropping box includes some of the surroundings, it is necessary to capture the full scale of seizure actions, even when violent movements happen during the seizure. Detection was not satisfying only in four out of the 126 cases (3.17 %), due to under- or misdetection of the region of interest, when the bed was not detected (2 cases), patient was already in sitting position (1 case) or only a physician was detected (1 case). In this cases the algorithm could not detect properly, due to heavy occlusions both on the patient and the patients' bed. These 4 seizure videos were replaced with raw, uncropped videos for the feature extraction phase. On the other hand, in several other instances the algorithm handled similar scenarios well, such as physicians on the frame, sitting position of the patient or partial occlusions on the bed. The developed cropping method is flexible and can be applied for different EMU setups, regardless of angle of view, orientation and distance to the camera.



(a) Before crop 1 (b) After crop 1 (c) Before crop 2 (d) After crop 2

Fig. 2: Examples of detection and crop, when the bed and the patient was properly detected, surrounding scenery removed, with enough space left to capture the full scale of seizures

4.2. Classification of I3D features

The 5-fold cross-validation metrics was evaluated with macro averages. This average treats both classes the same way, disregarding the class imbalance, thus maximizing the f1-score, precision, recall and other metrics for both classes at the same time. This prevents overfitting for the class with more samples, which may have false suggestions of higher global performance, by disregarding the minority class [31]. The average±standard deviation of the 5-fold cross validation macro average metrics f1-score is 0.844±0.042, precision is 0.857±0.042 and recall is 0.838±0.041. Overall, the metrics suggest that the classifier learned the classes satisfyingly, without highly overfitting any of the classes, in spite of the class imbalance both in training and evaluation sets. However, generally through the 5-fold cross-validation the class with more samples in the fold had better performance. This means that there is a slight bias in favor of the class with more samples in the validation set still exists. Thus, there is some confusion from the minority class to the majority class (Fig. 3).

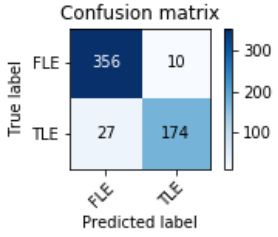


Fig. 3: Confusion matrix of the best fold in the 5-fold cross validation

4.3. Comparison to the state of the art

The achieved results are presented in comparison with the other similar literature contributions in Tab. 1. To the best of our knowledge, the current work outperforms all previous deep learning based approaches to automated epileptic seizure type differentiation. Compared to the work by Achilles et. al. [17, 18], a more complex classification problem was used (FLE vs TLE), instead of just the detection of the presence or absence of seizures. Even with these more specific classes the developed architecture improved the average AUC with 0.12 to 0.90±0.04. The other approaches by Ahmedt-Aristizabalet al. [19] and Maia et. al. [21] use also binary classification of Mesial Temporal Lobe Epilepsy (MTLE), Extra Temporal Lobe Epilepsy (ETLE) and TLE vs. ETLE classes respectively. These classes are still broader, than the ones utilized in this work, moreover they achieved very limited performances compared to the architecture presented in this paper.

5. DISCUSSION

In EMUs, physicians depend on subjective evaluation of video-EEG data. In order to support diagnosis of epilepsy and reduce the burden of data evaluation it is advantageous to utilize machine learning techniques for seizure detection and classification. The dataset collected at the University of Munich with the NeuroKinect 3.0 system is an excellent initiative to support supervised machine learning with labeled clinical data in this domain. Although the current dataset is still limited and imbalanced, it provides a good basis for initial research and developing proof-of-concept systems. The limitations of the dataset can be mitigated by taking advantage of transfer learning from other large datasets. To extract relevant information from the available videos and minimize unrelated variations, our developed automated cropping algorithm based on the Mask R-CNN achieved encouraging results with successful cropping rate of 96.83%. The classification of the extracted I3D features with the developed LSTM-FC architecture performed above the literature. This classifier managed to handle the imbalanced dataset, with application of regularization techniques and splitting the data into 2 second samples. The results suggest that the developed classifier did not overfit the data, however there is a slight bias in each fold to the class with the majority of the samples, due to the class imbalance in both training and validation sets. Furthermore, the performance was good even with the features extracted from just 2 second snippets from the IR seizure videos. In fact this splitting strategy improved the handling of class imbalance. The seizures for each validation fold originate only from 4 FLE and 3 TLE patients, with approximately 17 and 8 seizures per class respectively, however the number of seizures available varies slightly by patients. With the splitting strategy this resulted in roughly 600 samples in total, which was more balanced due to the different length of the seizures in each class. Even though the I3D network was originally trained on RGB videos (not IR) with different movement classes, the transfer learning approach still performed well. The involvement of information from other significantly larger datasets, by transferring pre-trained weights from similar (static and dynamic) domains proved very useful. Especially the application of I3D network, which is developed for human action recognition found to be smoothly fitting in the proposed architecture. By comparison, previous publications either did not use feature extraction with pre-training or used pre-trained ones for static image recognition [21, 19], which neglect temporal features. In order to improve the feature extraction, the weights of the I3D network could be fine tuned, with the current seizure IR dataset. Moreover, fine tuning the weights would also improve the distinction of seizure specific movements. Classification results could be further improved by acquiring more data, with better balance, including synchronous depth 3D information and EEG.

6. CONCLUSION

An end-to-end deep learning approach was proposed for motion based binary classification of Frontal and Temporal Lobe Epileptic seizures. The system uses IR videos of the seizures that are 24/7 acquired at the EMUs. The developed architecture is based on a combination of Mask R-CNN pre-processing, I3D feature extraction and LSTM-FC classification, heavily utilizing transfer learning from static object detection and dyanamic human action recognition datasets with network architectures available in the literature. It achieved a 5-fold cross-validation f1-score of 0.844±0.042. To the best of our knowledge, the current work outperforms all previous deep learning based approaches to automated epileptic seizure type differentiation indicating a high potential to support physicians with diagnostic decisions. It also shows future potential for online monitoring applications in epilepsy monitoring units.

7. REFERENCES

[1] A. Singh and S. Trevick, "The Epidemiology of Global Epilepsy," *Neurologic Clinics*, vol. 34, no. 4, pp. 837–847, nov 2016.

[2] R. S. Fisher, W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel, "Epileptic Seizures and Epilepsy: Definitions Proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, apr 2005.

[3] S. Noachtar and I. Borggraefe, "Epilepsy surgery: A critical review," *Epilepsy & Behavior*, vol. 15, no. 1, pp. 66–72, may 2009.

[4] S. Noachtar and A. S. Peters, "Semiology of epileptic seizures: A critical review," *Epilepsy & Behavior*, vol. 15, no. 1, pp. 2–9, may 2009.

[5] F. Rosenow, "Presurgical Evaluation of Epilepsy," *Brain*, vol. 124, no. 9, pp. 1683–1700, sep 2001.

[6] F. Fürbass, P. Ossenblok, M. Hartmann, H. Perko, A. Skupch, G. Lindinger, L. Elezi, E. Pataraia, A. Colon, C. Baumgartner, and T. Kluge, "Prospective multi-center study of an automatic online seizure detection system for epilepsy monitoring units," *Clinical Neurophysiology*, vol. 126, no. 6, pp. 1124–1131, jun 2015.

[7] M. do Carmo Vilas-Boas and J. P. S. Cunha, "Movement Quantification in Neurological Diseases: Methods and Applications," *IEEE Reviews in Biomedical Engineering*, vol. 9, pp. 15–31, 2016.

[8] C. Hoppe, A. Poepel, and C. E. Elger, "Epilepsy: accuracy of patient seizure counts," *Archives of Neurology*, vol. 64, no. 11, pp. 1595, nov 2007.

[9] D. E. Blum, J. Eskola, J. J. Bortz, and R. S. Fisher, "Patient awareness of seizures," *Neurology*, vol. 47, no. 1, pp. 260–264, jul 1996.

[10] F. Kerling, S. Mueller, E. Pauli, and H. Stefan, "When do patients forget their seizures? An electroclinical study," *Epilepsy & Behavior*, vol. 9, no. 2, pp. 281–285, sep 2006.

[11] A. V. Kurada, T. Srinivasan, S. Hammond, A. Ulate-Campos, and J. Bidwell, "Seizure detection devices for use in antiseizure medication clinical trials: A systematic review," *Seizure*, vol. 66, pp. 61–69, mar 2019.

[12] J. P. S. Cunha, H. M. P. Choupina, A. P. Rocha, J. M. Fernandes, F. Achilles, A. M. Loesch, C. Vollmar, E. Hartl, and S. Noachtar, "NeuroKinect: A novel low-cost 3dvideo-EEG system for epileptic seizure motion quantification," *PLOS ONE*, vol. 11, no. 1, pp. e0145669, jan 2016.

[13] H. M. Pereira Choupina, A. P. Rocha, J. M. Fernandes, C. Vollmar, S. Noachtar, and J. P. Silva Cunha, "NeuroKinect 3.0: Multi-Bed 3Dvideo-EEG System for Epilepsy Clinical Motion Monitoring.," *Studies in health technology and informatics*, vol. 247, pp. 46–50, 2018.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. jun 2009, IEEE.

[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *arXiv*, 2017.

[16] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. jul 2017, IEEE.

[17] F. Achilles, F. Tombari, V. Belagiannis, A. M. Loesch, S. Noachtar, and N. Navab, "Convolutional neural networks for real-time epileptic seizure detection," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 264–269, jul 2016.

[18] F. Achilles, V. Belagiannis, F. Tombari, A. Loesch, J. Cunha, N. Navab, and S. Noachtar, "Deep convolutional neural networks for automatic identification of epileptic seizures in infrared and depth images," *Journal of the Neurological Sciences*, vol. 357, pp. e436, oct 2015.

[19] D. Ahmedt-Aristizabal, C. Fookes, S. Denman, K. Nguyen, T. Fernando, S. Sridharan, and S. Dionisio, "A hierarchical multimodal system for motion analysis in patients with epilepsy," *Epilepsy & Behavior*, vol. 87, pp. 46–58, oct 2018.

[20] D. Ahmedt-Aristizabal, C. Fookes, S. Dionisio, K. Nguyen, J. P. S. Cunha, and S. Sridharan, "Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey," *Epilepsia*, vol. 58, no. 11, pp. 1817–1831, oct 2017.

[21] P. Maia, E. Hartl, C. Vollmar, S. Noachtar, and J. P. S. Cunha, "Epileptic seizure classification using the NeuroMov database," in *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*. feb 2019, IEEE.

[22] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*. oct 2017, IEEE.

[23] F. Chollet et al., "Keras," `https://keras.io`, 2015.

[24] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow," `https://github.com/matterport/Mask_RCNN`, 2017.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, pp. 740–755. Springer International Publishing, 2014.

[26] W. Abdulla, "Keras implementation of Inflated 3D from Quo Vardis paper + weights," `https://github.com/dlpbc/keras-kinetics-i3d`, 2018.

[27] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. 2015, ICML'15, pp. 448–456, JMLR.org.

[28] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*. dec 2015, IEEE.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR 2015*, 2014.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.