Interpretable Epilepsy Detection in Routine, Interictal EEG Data using Deep Learning

Thomas Uyttenhove THOMAS.UYTTENHOVE@UGENT.BE
Ghent University

Aren Maes AREN.MAES@UGENT.BE
Tom Van Steenkiste TOM.VANSTEENKISTE@UGENT.BE
Dirk Deschrijver DIRK.DESCHRIJVER@UGENT.BE
Tom Dhaene TOM.DHAENE@UGENT.BE
Ghent University - imec

Editors: Emily Alsentzer[⊗], Matthew B. A. McDermott[⊗], Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy[‡], Stephanie L. Hyland[‡]

Abstract

Epilepsy, a common serious neurological disorder, is characterized by its frequently occurring seizures that cause its patients to be three times as likely to die prematurely. While the application of machine learning to EEG recordings has enabled the successful prediction of whether and when such seizures will occur, the reliable detection of epilepsy during seizure-free periods is lacking. As far as the authors are aware, this work proposes the first deep learning approach for the latter task – and the second machine learning approach altogether. Additionally, it does so in an interpretable fashion to validate the proposed method for a more wide-spread adoption in healthcare and to potentially unveil unknown epileptic biomarkers. The performance of the Tiny Visual Geometry Group (t-VGG) convolutional neural network is evaluated against Temple University Hospital's *EEG Epilepsy Corpus*, a data set of variable-length EEG recordings gathered during routine checkups. The t-VGG network predicted individual 10 second EEG windows with an Area Under the Precision-Recall Curve (AUPR) of 93.02% for epileptic predictions and 55.85% for healthy ones – a significant improvement of respectively 7.24pp and 18.6pp ($p < .001$) over the current state-of-the-art. Averaging window predictions per recording improved the t-VGG's respective AUPR performances further to 95.52% and 77.27%. The Gradient-weighted Class Activation Mapping method for interpretability confirmed that the model was able to learn sensible features with connections to well-known epilepsy markers.

Keywords: epilepsy, machine learning, interpretable, EEG, interictal

1. Introduction

Epilepsy affects over 50 million people around the world, making it a common serious neurological disorder. The number of epileptic people diagnosed each year is estimated at 5 million and is expected to increase as life expectancy is extended and people are aging on a global scale (World Health Organizition, 2019). It is infamously characterized by seizures, i.e., periods of excessive electrical discharges in the brain resulting in abnormal behavior, unusual sensations, or loss of awareness (Mayo Clinic, 2020). The risk of premature death is up to three times as high for epileptic patients due to the several complications that can occur during these seizures, such as falling, drowning, car accidents, etc. However, estimates by the World Health Organizition (2019) show that a proper diagnosis and treatment could allow as much as 70% of people suffering from epilepsy to live seizure-free.

1.1. Automated EEG Data Analysis

The most common test for diagnosing a patient with epilepsy and other neurological disorders is the analysis of an Electroencephalogram or EEG (Mayo Clinic, 2020), i.e., a non-invasive test where electrodes measure electrical impulses in the brain to monitor its activity. Automation using powerful machine learning techniques, that can grasp the broad spatiotemporal complexity of EEG signals, have become prominent in neurological research since recent advancements made such techniques more feasible and accessible than ever (Niedermeyer and da Silva, 2005; Castellano, 2016).

On one hand, the application of machine learning has enabled near-perfect accuracy when predicting imminent epileptic seizures or detecting them as they are ongoing. Conventional machine learning techniques have shown promising performance (Acharya et al., 2012; Agarwal et al., 2018), but the most recent breakthroughs were achieved at the hands of deep learning approaches revolving around Convolutional Neural Networks (CNN) (Daoud and Bayoumi, 2019; Daoud and Bayoumi, 2018).

On the other hand, the diagnosis of epilepsy during seizure-free, i.e., interictal, periods has received less attention. Buettner et al. (2019) presented a random forest classifier that uses the frequency band powers measured in the 1.5-2Hz and 11-12.5Hz frequency sub-bands – to the authors' knowledge the only machine learning approach so far. Their setup achieved a near-perfect 99% accuracy, 100% sensitivity, and 98% specificity when tested on the short, single-channel intracranial, i.e., recorded within the skull, EEG segments of the Bonn University department of Epileptology data set (Andrzejak et al., 2001). However, a more modest 75.6% accuracy was documented when applied to less invasively and routinely recorded regular EEG segments (Rieg et al., 2020).

1.2. Interpretable Machine Learning

For the successful and wide-spread adoption of machine learning techniques in healthcare, a higher standard in terms of reliable, accurate, and explainable predictions is required - a unique combination of challenges for machine learning models (Ahmad et al., 2018). Explainable diagnosis of any neurological disorders could offer valuable new insights for future work, consequently improving the accuracy of automated diagnoses and thus working towards a higher quality of life for people suffering from them.

Despite having the ability to be very powerful, the black-box nature of deep learning models inherently offers no insight into how decisions are made exactly (The Lancet Respiratory Medicine, 2018). To this end, several techniques have found their way into the EEG analysis domain to analyze predictions made by such models. Often, these rely on creating heat-maps of the contribution of each input segment to making the final decision (Lawhern et al., 2018; Zhou et al., 2016; Selvaraju et al., 2017; Bach et al., 2015).

1.3. Objective and Outline

This paper presents a deep learning solution for automatically diagnosing epilepsy in routine EEG data, collected during routine recordings of mostly interictal periods. The Tiny-Visual Geometry Group (t-VGG) architecture, proposed by Jonas et al. (2019), is applied to tackle this task. It is particularly interesting as it lends itself to visualized and explainable predictions: the Gradient-weighted Class Activation Mapping (Grad-CAM) technique is used to offer insight into how exactly predictions are being made.

The outline for this work is as follows. In Section 2, the proposed method is presented in greater detail, including the data set of choice, t-VGG model architecture, and Grad-CAM technique. Section 3 explains how the proposed method will be objectively evaluated. Section 4 presents and discusses the model's performance and analyzes its interpretability. Finally, Section 5 summarizes the presented work and proposes some guidelines for future work.

2. Materials and Methods

2.1. Data

The data set used for this experiment is the Temple University Hospital (TUH) *EEG Epilepsy Corpus* (TUEP) (Veloso et al., 2017). It is part of a large collection of over 30,000 clinical EEG signals recorded at TUH, Philadelphia. The TUEP data set contains over 1,600 variable-length EEG recordings from over 100 epileptic and 100 healthy male and female subjects of varying ages, ranging from 17 up to 88 years old. An overview of the data distribution is shown in Table 1. With over 400 hours of EEG recordings, TUEP is possibly the largest and most varied public collection of routinely recorded epileptic EEG signals. EEG recordings were acquired using at least the 19 electrodes defined by the 10-20 standard montage (Morley et al., 2016). The majority of recordings were average referenced and sampled at 256Hz. The epileptic recordings took place in interictal periods mostly, but patients may have experienced one or more epileptic discharges or seizures at the time of recording.

Table 1: Overview of Temple University Hospital's *EEG Epilepsy Corpus* data set.

Diagnosis	Patients	Sessions	Files
Epilepsy	100	436	1,360
No epilepsy	100	134	288
All	200	561	1,648

To avoid discrepancies in the recordings, only the EEG signals that were average referenced are used. Channels other than those of the 10-20 montage are omitted and the signals of all 19 channels are resampled to 256Hz. Frequency components are limited to the 0.5-128Hz range, i.e., including the brain's slowest delta brain waves up to brain waves as fast as half the sampling rate and just short of the highest frequency 140Hz gamma brain waves. The signals are also notch filtered at multiples of 60Hz to exclude any line noise due to medical equipment that may have occurred during recording (Leske and Dalal, 2019). To explore if the proposed model can be learnt to deal with artifacts, just like trained physicians would during manual assessment, artifact removal is not considered. Finally, to deal with the variable recording length each recording is cut into 10 second windows to match the window length employed by Jonas et al. (2019). Though they decided on an overlap of 75%, this is decided against as the experiment would otherwise become too computationally expensive. As a result, recordings are reduced to multiple windows of 2,560 samples over 19 channels; the total amount of preprocessed EEG windows is 149,529 of which 123,815 are epileptic and 25,714 were recorded from healthy patients.

2.2. Machine learning modeling

The Tiny Visual Geometry Group (t-VGG) CNN architecture proposed by Jonas et al. (2019) for coma outcome prediction has shown promising performance on the task of automated EEG analysis. It is based on VGGNet, a complex and powerful deep CNN architecture (Simonyan and Zisserman, 2015) that has shown powerful performance in several image recognition domains (Russakovsky et al., 2015; Ke et al., 2018). Because of its complexity, it often suffers from overfitting and requires tedious hyperparameter tuning – especially with a limited amount of data. The t-VGG is a simplified, single dimensional version of this architecture that reduces depth and width both of blocks and of layers of blocks. Each convolutional layer goes accompanied by a Batch Normalization (BN) layer (Ioffe and Szegedy, 2015) and a Rectified Linear Unit (ReLU) activation function (Nair and Hinton, 2010). The fully connected part of the network consists of only one hidden layer consisting of 16 nodes, is regularized by the dropout of 50% of its neurons during training, and large weights are penalized using L2-regularization with a λ parameter of 0.01. The architecture of the model is presented in Figure 1.

The t-VGG network is the main model being tested in this research. Data windows of recordings are shuffled and split in groups of 128 windows to achieve batches of independent EEG windows that can be loaded in one at a time. During training, the Adaptive Moment estimating (ADAM) (Kingma and Ba, 2014) optimizer is used with its standard settings ($\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) to minimize the binary cross-entropy loss, as is customary for binary classification problems.

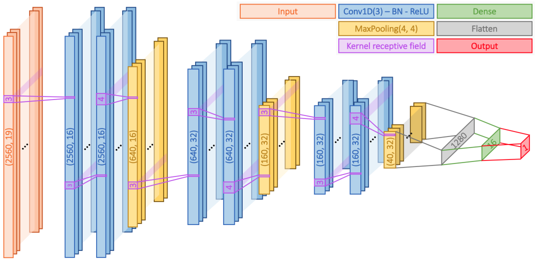


Figure 1: The t-VGG architecture. A kernel length of 3 was used for the convolutional layers, and both a pool size and stride of 4 for the max pooling layers.

2.3. Model Interpretability

The importance of interpretable machine learning, especially for black-box deep learning models, was already established in Section 1. One technique for introducing transparency specifically in CNN models is Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017), put forward by Jonas et al. (2019) as a companion interpretability technique for their t-VGG network. The technique relies on convolutional layers' property of preserving spatial information that was present in the input data. The last convolutional layer is expected to be far enough up the feature learning hierarchy to translate into sensible and high-level features while still offering insight into the temporal information, that would otherwise get lost in the fully-connected layers (Jonas et al., 2019). The gradient flowing from the target class, i.e., an epileptic or healthy diagnosis, into the last convolutional layer is then used to weight the importance of each neuron. The result is a rough heat-map that, when run through the model's activation function, highlights which segments of the input data had positive influences on it being classified as the class of interest.

Fully-connected layers at the end of a CNN cause part of the decision to happen after the last convolutional layer, resulting in the weights of the latter not accurately representing the model's learned features. A solution offered by Zhou et al. (2016) proposes to end the CNN with a Global Average Pooling (GAP) layer instead. This layer then resembles the features detected by each filter of the last convolutional layer for the model to base its classification decision on directly.

Following the suggestion by Jonas et al. (2019), a GAP variant of the t-VGG model is also considered in this work. The diagnoses of the test data, made by the t-VGG GAP model, is inspected with the Grad-CAM procedure and activation maps are visualized to see if systematic decision patterns can be identified from the trained network for both epileptic and healthy diagnoses.

3. Experimental Setup

3.1. Models for Testing

To objectively measure the performance of the proposed method, it is compared to the performance of a collection of reference models. On top of the proposed t-VGG architecture, these reference models are also trained following the same method as described in Section 2. What follows is an overview of the reference models considered in this experiment. For completeness' sake, each method will be accompanied by a number of potential methods for explaining their predictions. For more in-depth information and additional techniques, the reader is referred to specialized literature such as Molnar (2020).

The **Dummy Classifier (DC)** is a baseline model that performs a weighted guess that should reveal biases in the data distribution. Since this model merely provides a weighted guess, it is inherently explainable. The **Support Vector Machine (SVM)** (Evgeniou and Pontil, 2001) is often the simplest well-performing model for a variety of machine learning tasks. This model returns a decision boundary which can be inspected, though more advanced techniques may be desired, e.g., extracting the most influential support vectors or features for the classification of a sample (Barbella et al., 2009). The **Random Forest (RF)** (Breiman, 2001), configured as proposed by Buettner et al. (2019), is the current state-of-the-art for automated epilepsy diagnosis. Its underlying structure can be inspected to explain predictions, e.g., the most important features for distinguishing between target classes (Buettner et al., 2019) or the decision trees that constitute the RF (Saeed, 2020). **EEGNet** (Lawhern et al., 2018) is a compact CNN model that offers well-known feature extraction concepts and has shown robust performance in Brain-Computer Interface related tasks. With this model, four interpretability methods were presented: the Grad-CAM technique described in Section 2.3, averaging the outputs of hidden unit activations that can be linked to specific temporal features to localize them spatially, an average wavelet time-differencing method that allows the creation of visualized topologies per filter, and the visualization of convolutional filter weights. The **Tiny-Visual Geometry Group (t-VGG)** and its **t-VGG Global Average Pooling (t-VGG GAP)** variant (Jonas et al., 2019; Zhou et al., 2016) are the CNN models being tested and compared to the above models for their performances. Beside the Grad-CAM technique, these models are also eligible for the application of the aforementioned EEGNet interpretability methods.

To allow for a fair comparison, a selection of features is extracted to train and test the non-neural network models. The band powers for the fine frequency sub-bands proposed by Buettner et al. (2019), i.e., 1.5-2Hz, 10.5-11Hz, 11-11.5Hz, and 11.5-12Hz, are extracted for each channel of the EEG windows and serve as the input features for the DC, SVM, and RF models. The neural network classifiers, i.e., EEGNet and t-VGG (GAP), are trained on the preprocessed EEG window data.

3.2. Training and Testing Approach

Since different EEG recordings of the same patient are inherently correlated, it is of utmost importance to create train, validation, and test sets according to a group split on the level of the patient to avoid date leakage. The train and test sets are constructed by first calculating the cumulative number of recording windows per patient and then splitting off a number of patients that constitute an approximate 20% of the data set for testing. This procedure is repeated to split off an approximate 20% of the training data for validation purposes. To confirm the proposed model performs well across all patients rather than just those present in the validation data, the experimental setup is extended with 5-fold cross-validation (Kohavi, 1995). This too is applied as a group split on a patient level to prevent data leakage.

Keeping the class imbalances present in the data of the resulting 5 folds could inhibit the model from learning important features that allow it to accurately distinguish between the target classes. To balance the data while still providing the model with as many different samples to learn from as possible, all epileptic EEG windows are preserved by oversampling the minority healthy class. Note that this is only done for the training and validation data, as it is important to test the model's performance against the original distribution to estimate how it would perform in a real-world setting.

The test split and 5-fold cross-validation split are done once and reused for all models being trained. The validation data serves specifically for tuning the hyper-parameters of the machine learning models until their best-performing configurations are found. Finally, all models are trained on the combination of the training and validation data, that make up the complete training set. Predictions are made for the test data and are evaluated to estimate the models' ability to generalize on unseen data.

3.3. Model Evaluation

To grasp the model's ability to distinguish epileptic from non-epileptic EEG recordings, the most important metric is the Area Under the Precision-Recall curve (AUPR), i.e., a combination of the recall and precision metrics that is high when both recall and precision are high, corresponding to a low False Negative Rate (FNR) and a low False Positive Rate (FPR) respectively. However, other conventional metrics, e.g., accuracy, recall, specificity, Area Under the ROC Curve (AUC), etc., are kept track of for when a more nuanced comparison is in order. A statistical paired t-test (Hsu and Lachenbruch, 2005) of all tracked metrics quantifies the difference in performance between the proposed model and the described reference models. A change in performance is considered significant if the paired t-test results in a p-value value smaller than 0.05.

To evaluate the proposed model's interpretability, the predictions made by the final trained t-VGG model are looked into using the Grad-CAM procedure. The activations within the model are visualized and mapped back on the original input data. Manual inspection of multiple samples of each type of prediction, i.e., TP, FP, FN, and TN, have to determine if the network makes systematic, interpretable decisions that can be connected to findings made by Buettner et al. (2019) or potentially reveal new insights into how the classification is done by the t-VGG model. To present as much of the proposed method and interpretability method, the regular interpretability methods mentioned in Section 3.1 are be explored further.

4. Results

4.1. Model performance

A summary of the proposed models' most important performance metrics and how they stack up against the reference models is shown in Table 2.

The proposed t-VGG model predicted window diagnoses with an accuracy, recall, and specificity of respectively 70.38%, 66.92%, and 82.98%, resulting in the highest positive and negative AUPR scores of 93.02% and 55.85% respectively. This performance significantly outperforms the SVM (p = .0014), RF (p < .001), and EEGNet (p < .001) classifiers. When averaged across tokens, the positive and negative AUPRs increased to 95.52% and 77.27% – again the highest recorded performance so far. The

Table 2: A summary of the most important performance metrics, with the two best scores for each metric indicated in a bold typeface.

	Window predictions				Token-averaged predictions			
Classifier	AUPR	Acc.	Rec.	Sp.	AUPR	Acc.	Rec.	Sp.
DC	0.7858	0.5019	0.5003	0.5078	0.7735	0.5191	0.5319	0.4762
SVM	0.8651	0.6174	0.6109	0.6410	0.8945	0.7432	0.7518	0.7143
RF	0.8578	0.6633	**0.7014**	0.5246	0.8867	**0.8361**	**0.9291**	0.5238
EEGNet	0.8834	0.5553	0.4886	**0.7984**	0.9011	0.7541	0.7589	0.7381
t-VGG	**0.9302**	**0.7038**	0.6692	**0.8298**	**0.9552**	0.7650	0.7589	**0.7857**
t-VGG GAP	**0.9242**	**0.7552**	**0.7501**	0.7737	**0.9315**	**0.8142**	**0.8156**	**0.8095**

corresponding achieved token-averaged accuracy and recall are lower than the SVM ($p < .001$) and RF ($p = .73$) classifier while the specificity is higher; all three are better than reported for the EEGNet classifier ($p = .0028$). Note that the high p-value for RF was distorted by a trade-off between recall and specificity and that the t-VGG actually improved the RF's positive and negative AUPR with 6.85pp and 8.02pp respectively. The t-VGG, RF, and EEGNet Precision-Recall (PR) curves for their token-averaged predictions are shown in Figure 2(a).

The GAP version of the t-VGG classifier showed similar performance as the original version. A slight, but insignificant decrease in AUPR scores for both window and token-averaged predictions can be observed, corresponding to slightly higher accuracy and recall and lower specificity.

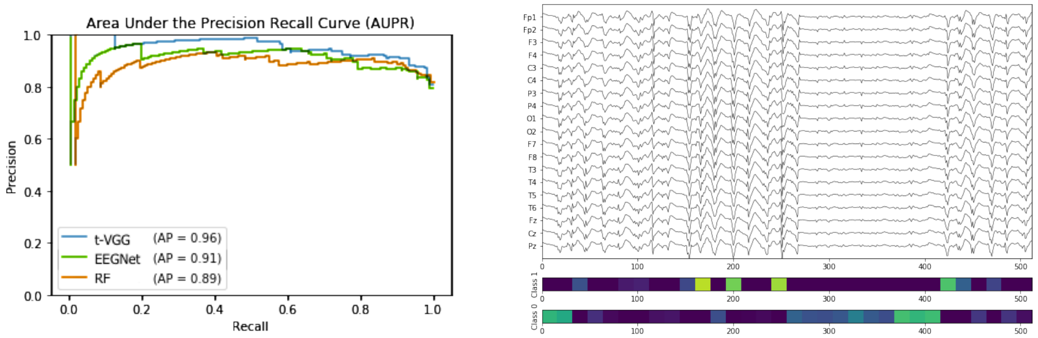
4.2. Interpretability

The Grad-CAM technique was used to inspect predictions made by the t-VGG GAP model to find features that enabled the model to make correct or incorrect decisions. The heat-maps resulting from the Grad-CAM algorithm were then overlaid on the original input and reoccurring patterns were identified. The model presented the ability to recognize interictal spikes and associate them with a positive diagnosis. Additionally, it managed to indicate other known wave patterns that are discriminative for epilepsy, such as seizure discharges, triphasic waves, burst suppressions, and slowing waves (Seneviratne et al., 2016; Trinka and Leitinger, 2015). Furthermore, it mainly associated segments resembling regular waking EEG recording with a negative diagnosis. Figure 2(b) displays such CAM for an epileptic segment, which shows interictal epileptic spikes being correctly associated with an epileptic outcome and regular waking EEG rhythms corresponding with a healthy class activation.

4.3. Discussion

The RF model's previously reported near-perfect performance on the Bonn University department of Epileptology (Andrzejak et al., 2001) could not be reproduced, but it still managed to achieve a token-averaged accuracy of 83.61% which surpasses the results from the retrospective study by Rieg et al. (2020). This high accuracy, obtained using only the band power of the presented fine frequency sub-bands, seems to confirm that these features are discriminatory for this task to an extent. However, the large discrepancies between performances based on the data set of choice seem to suggest that it is not a golden rule that can be followed; further dedicated studies are in order to refine the feature range to a general discriminator for epilepsy in EEG signals.

The performance of the proposed t-VGG architecture is the highest of all tested models. As shown in Table 2, for each metric on both the window and token-averaged level,



(a) (b)

Figure 2: The results of the experiment; Figure 2(a) shows the Precision-Recall (PR) curve for the window predictions made by the three best performing classifiers and Figure 2(b) shows the Grad-CAM of a recording containing interictal epileptic discharges; *Class* 1 and *Class* 0 respectively indicate the epileptic and healthy class.

either one of the t-VGG models is among the two best performing models and both are always among the three best performing models. The high token-averaged AUPR scores reveal that the model managed to distinguish between epileptic and healthy patients' EEG recordings quite well. It performs better than the current state-of-the-art RF model. Additionally, it significantly outperforms the EEGNet classifier at only a fraction of the model parameters and less training iterations.

Despite the t-VGG GAP classifier sacrificing some of its classification power by omitting its fully-connected layer, its narrow and heavily regularized dense layer shifted the majority of the complexity for learning features to the convolutional layers. Consequently, replacing the fully-connected layer with a Global Average Pooling layer did not influence the model architecture nor predictive power much. Furthermore, the lack of performance deterioration is beneficial for the Grad-CAM technique, as the visualizations can reveal more sensible features that are almost as discriminatory as the ones learned by the regular t-VGG model.

The good performance by the three DL approaches tested in this experiment indicates that these advanced ML techniques are also suited for the automated diagnosis of epilepsy. Despite not being further optimized for the task at hand, the EEG-Net classifier managed to achieve high performance that is on par with the RF classifier. On top of that, considering the t-VGG model's performance in the coma outcome prediction task and the results of the Grad-CAM algorithm, the model architecture seems to also enable the encapsulation of learning sensible features that can be connected to known EEG abnormalities and in some cases even to interictal or ictal biomarkers for epilepsy (Seneviratne et al., 2016; Trinka and Leitinger, 2015). Specifically, the model was able to detect epileptiform discharges, triphasic waves with slowing waves, and periods of burst suppressions as discriminant features for the epileptic class. Mostly waking EEG signals without much variations in frequency and amplitude were as indicating a healthy EEG window. Besides its ability to correctly nuance the presence of slight epileptic activations, the model also seems to distinguish between epileptic abnormalities and other irregularities quite well. However, some features, such as spikes due to artifacts or less prominent epileptic-seeming patterns also occasionally caused the model to make incorrect epileptic and healthy predictions respectively. Note that close clinical collaboration and input is in order to obtain event-annotated data for a complete and thorough analysis of the model's interpretability.

Though both t-VGG variants have shown to be a great match for the problem at hand, some remarks should be made on the limitations of the described experiment. For one, the lack of significant token-averaged performance improvement of the t-VGG model compared to the RF model can be explained by the observation that a reduction of accuracy and recall scores comes with a rather big increase in specificity. To allow for a better and fairer comparison, prediction thresholds should be looked into for all models to reach an optimum recall and specificity trade-off. Secondly, when further averaging predictions on a session or patient level from the token level, a counter-intuitive drop in performance is observed. This can be resolved using two techniques. On one hand, more advanced techniques, such as recurrent neural networks, could account for mistakes made by the model and the absence of discriminatory wave patterns in some of the EEG windows to come to an improved conclusion. On the other hand, more advanced data splitting techniques could resolve the changing class imbalance ratios across levels and thus mismatches in performance bias. Thirdly, the Grad-CAM techniques confirmed that the overall performance of the proposed method could be improved by removing abnormal EEG segments caused by artifacts. Fourthly, as this work aimed to present as much of the proposed method as possible, a comparison with the analyses of the reference models' interpretability was not included in this paper. Lastly, while the impact of several alterations to the t-VGG architecture was looked into, e.g., kernel size, convolutional block depth, network depth, network width, regularization, etc., no significant improvements in performance could be noted. However, there is a lot of room for further experimentation still. Rigorous testing of variations of the t-VGG model in a grid-search-like fashion or approaches with vastly different DL architectures, inspired by state-of-the-art DL models in neurology or new specially designed architectures, should be looked into. Specific attention should go to recurrent techniques to exploit temporal correlations in electrode channels and 2D-CNNs to exploit topology-related spatial correlations between the different channels.

5. Conclusion

In this work, the second machine learning approach for the task at hand and the first deep learning approach – to the authors' knowledge – was proposed to diagnose epilepsy from EEG signals recorded routinely during seizure-free periods. The t-VGG network architecture was compared to an assortment of reference models and was shown to achieve state-of-the-art performance for this task. Further, its predictions were inspected with the Grad-CAM technique by creating heatmaps of input segments that were positively influenced the classification. This revealed that deep learning is also capable of learning the complex features from EEG data that allow it to distinguish between epileptic and healthy patients with high accuracy. This research thus opened the door for other, more advanced deep learning techniques to tackle the challenge of improving the accuracy by which epileptic patients can be properly and reliably diagnosed.

Acknowledgments

This research received funding from the Flemish Government (AI Research Program).

References

U. R. Acharya, S. V. Sree, P. C. A. Ang, R. Yanti, and J. S. Suri. Application of non-linear and wavelet based features for the automated identification of epileptic EEG signals. *International journal of neural systems*, 22(02):1250002, 2012.

P. Agarwal, H. Wang, and K. Srinivasan. Epileptic Seizure Prediction over EEG Data using Hybrid CNN-SVM Model with Edge Computing Services. In *MATEC Web of Conferences*, volume 210, page 03016. EDP Sciences, 2018.

M. A. Ahmad, A. Teredesai, and C. Eckert. Interpretable Machine Learning in Healthcare. *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018. doi: 10.1109/ichi.2018.00095.

R.G. Andrzejak, K. Lehnertz, C. Rieke, F. Mormann, P. David, and C.E. Elger. Indications of nmiscar deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, page 061907, 2001.

S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

D. Barbella, S. Benzaid, J. M. Christensen, B. Jackson, X. V. Qin, and D. R. Musicant. Understanding Support Vector Machine Classifications via a Recommender System-Like Approach. In *DMIN*, pages 305–311. Las Vegas, NV, 2009.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

R. Buettner, J. Frick, and T. Rieg. High-performance detection of epilepsy in seizure-free EEG recordings: A novel machine learning approach using very specific epileptic EEG sub-bands. 2019.

M. Castellano. The rise and decline of eeg, where are we now?, Feb 2016. URL https://www.neuroelectrics.com/blog/2016/03/20/the-rise-and-decline-of-eeg-where-are-we-now/. Accessed: 2020-05-26.

H. Daoud and M. Bayoumi. Deep learning based Reliable Early Epileptic Seizure Predictor. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2018.

H. Daoud and M. A. Bayoumi. Efficient Epileptic Seizure Prediction Based on Deep Learning. *IEEE Transactions on Biomedical Circuits and Systems*, 13(5): 804–813, 2019.

T. Evgeniou and M. Pontil. Support Vector Machines: Theory and Applications. volume 2049, pages 249–257, 01 2001. doi: 10.1007/3-540-44673-7_12.

H. Hsu and P. A. Lachenbruch. Paired t test. *Encyclopedia of Biostatistics*, 6, 2005.

Se. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

S. Jonas, A. O. Rossetti, M. Oddo, S. Jenni, P. Favaro, and F. Zubler. Eeg-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. *Human Brain Mapping*, 40 (16):4606–4617, 2019.

H. Ke, D. Chen, X. Li, Y. Tang, T. Shah, and R. Ranjan. Towards brain big data classification: Epileptic eeg identification with a lightweight vggnet on global mic. *IEEE Access*, 6:14722–14733, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

R. Kohavi. A study of cross-validation and bo otstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

S. Leske and S. S. Dalal. Reducing power line noise in EEG and MEG data via spectrum interpolation. *NeuroImage*, 189:763 – 776, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.01.026.

Mayo Clinic. Epilepsy, May 2020. URL https://www.mayoclinic.org/diseases-conditions/epilepsy/symptoms-causes/syc-20350093. Accessed: 2020-05-26.

C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

A. Morley, L. Hill, and A.G. Kaditis. 10-20 system EEG Placement. *European Respiratory Society, European Respiratory Society*, 2016.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

E. Niedermeyer and F.H. L. da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

T. Rieg, J. Frick, and R. Buettner. Machine learning-based diagnosis of epilepsy in clinical routine: Lessons learned from a retrospective pilot study. *Information Systems and Neuroscience, LNISO*, 2020.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

A. Saeed. Improving the interpretability of the Random Forest classifier, 2020. URL https://towardsdatascience.com/representing-a-random-forest-with-few-decision-trees-11283e433dbb. Accessed: 2020-09-25.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

U. Seneviratne, G. Hepworth, M. Cook, and W. D'Souza. Atypical EEG abnormalities in genetic generalized epilepsies. *Clinical Neurophysiology*, 127(1):214–220, 2016.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015.

The Lancet Respiratory Medicine. Opening the black box of machine learning. *The Lancet Respiratory Medicine*, 6(11):801, 2018. doi: 10.1016/s2213-2600(18)30425-9.

E. Trinka and M. Leitinger. Which EEG patterns in coma are nonconvulsive status epilepticus? *Epilepsy & behavior*, 49:203–222, 2015.

L. Veloso, J. R. McHugh, E. von Weltin, I. Obeid, and J. Picone. Big Data Resources for EEGs: Enabling Deep Learning Research. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, page 1, 2017.

World Health Organizition. Epilepsy, 2019. URL https://www.who.int/news-room/fact-sheets/detail/epilepsy. Accessed: 2020-05-26.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.