Received March 15, 2022, accepted May 4, 2022, date of publication May 18, 2022, date of current version June 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176367

An Interpretable Deep Learning Classifier for Epileptic Seizure Prediction Using EEG Data

IMENE JEMAL[1,2], NEILA MEZGHANI[2,3], LINA ABOU-ABBAS[2,3], AND AMAR MITICHE[1]
[1]INRS-EMT, Institut National de la Recherche Scientifique, Montréal, QC H2X 1E3, Canada
[2]Centre de Recherche LICEF, Université TÉLUQ, Montréal, QC G1K 9H6, Canada
[3]Laboratoire LIO, Centre de Recherche du CHUM, Montréal, QC H2X 0A9, Canada

Corresponding author: Imene Jemal (imene.jemal@inrs.ca)

This work was supported by the Canada Research Chair on Biomedical Data Mining under Grant 950-231214.

ABSTRACT Deep learning has served pattern classification in many applications, with a performance which often well exceeds that of other machine learning paradigms. Yet, in general, deep learning has used computational architectures built, albeit partially, by ad hoc means, and its classification decisions are not necessarily interpretable in terms of knowledge relevant to the application it serves. This is often referred to as the black box problem, which in certain applications, such as epileptic seizure prediction, can be a serious impediment. The purpose of this study is to investigate an interpretable deep learning classifier for epileptic EEG-driven seizure prediction. This neural network is interpretable because its layers can be visualized and interpreted as a result of a novel architecture where the learned weights follow from signal processing computations such as frequency sub-band and spatial filters. Consequently, the extracted features are no longer abstract as they correspond to the features commonly used for decoding EEG data. In addition, the network uses layer-wise relevance propagation to reveal pertinent features which can further explain the computations leading to the decisions. In seizure prediction experiments using the CHB-MIT data set, the method produced classification results which improved on the state-of-the art, with first network layer filters corresponding to clinically relevant frequency bands, and the input channels in the brain location in which the seizure originates contributing most significantly to the network predictions.

INDEX TERMS Epileptic seizure prediction, deep neural networks, interpretable decisions, EEG signal.

I. INTRODUCTION

Deep neural networks have extended considerably the ability of common neural networks to learn and classify patterns, with striking, unprecedented results in long standing applications, and in challenging new ones as well [1]. However, in many other important applications, such as EEG signal classification for epileptic seizure prediction, which is the subject of this study, pattern feature learning in deep neural networks, or deep learning (DL) [2], suffers from what is often referred to as the *black box* problem, where, in general, some prevalent network architecture is used, without explicit justification, to have its parameters learned from data by often adhoc trial and error experimentation. As a result, better classification is often missed. Moreover, even when the classification is accurate, the results come essentially with no interpretation of how the network reached its classification decisions. The ability to interpret these decisions may not be an issue in simple classification tasks where a wrong outcome is of little consequence. However, in general, interpretation aids in learning better network parameters, for instance by biasing the network structure to favor learning of application-relevant features. In domains such as healthcare, it may be essential to develop efficient applications relevant in clinical settings. In this study, we consider a neural network to be *interpretable* at two levels: First, by designing layers that bias learned filters toward common signal processing computations, such as frequency sub-band and spatial filtering, which are relevant to the application that the network serves and, second by explaining the influence of the various input variables, or of the network learned features, in reaching the classification decisions. For instance, the deep neural network we investigate in this study for epileptic seizure prediction is interpretable in that it uses a convolution layer similar to a filter bank to extract characteristic filters corresponding to clinically relevant frequency bands, and the input channels in

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Wang.

the brain location in which the seizure originates contribute most significantly to the network predictions.

The ability to interpret classifiers has been of general interest in Artificial intelligence and has first appeared in symbolic reasoning which supports decision making in expert systems, such as MYCIN [5] which sets to diagnose patients on the basis of reported symptoms and medical tests results, and also GUIDON [6] for knowledge based tutoring, and SOPHIE [7] for hazard identification. In machine learning as well, being able to explain the relationship between the input and output of predictive models, and interpret the outcomes of this relationship, has been a concern [8]. Notable contributions in this regard are decision trees and related classifying structures, considered interpretable from the flow of their successive decision making stages [9].

Interpretable neural networks are of relatively recent interest [11], although the need for these is quite plain because, as mentioned earlier, neural networks, particularly deep neural networks, are currently used mostly in a black box manner, to process a multitude of classification applications which would be more accurate and of more flexible and general usage were their internal architecture are interpretable. Noteworthy investigations include studies and applications of saliency maps to visualize and understand non-linearity in neural networks [10], and computer vision studies which related neural activity to image filtering [12]. The investigation in [12] was able to determine that the first-layer filters learned by a deep belief network for natural image patch recognition is analogous to location, orientation, and spatial frequency filters like Gabor filters used for edge detection. For digits recognition using the MNIST dataset, [13] observed that the network learned low-level features similar to those found in stroke detectors typically used for text localization. For accrued interpretation abiliy [14], an explanation producing model can be construed using an architecture designed so as to simplify interpretations of internal representations and corresponding processing.

A means to explain a deep neural network computations is the layer-wise relevance propagation (LRP) scheme [15], by which a network decision is decomposed into relevance scores for each neuron, starting from last layer and propagating back towards the input. LRP has been a useful explanation tool in many applications. For instance, [16] used LRP to explain digit recognition and gender classification using the AudioMNIST dataset, which contained spoken digit records. The spectrogram representation showed that different areas of the input were critical to each class for digit classification, and the low frequency range was a determinant for gender classification. Moreover, based on waveform data, large magnitude data were determined also important. In another study [17] LRP was used to explain the classification of subjects EEG recordings while imagining left and right hand movements. The relevance score was calculated for each channel at each time point. The channel relevance of a time point reveals a typical lateralized motor activation pattern, which, when averaged over all epochs, yields a similar pattern.

In spite of the evident progress in building interpretable pattern classifiers for several important image-based applications, the subject remains actual and challenging for waveform data. Little work has been done in EEG-based applications, and none in epilepsy seizure prediction, the subject of this study. This study takes up the problem of developing an interpretable deep learning neural network applicable to epileptic seizure prediction in EEG recordings. Epilepsy seizure prediction is a subject worthy of investigation because epilepsy affects about 2.4 million people of all ages worldwide each year [18] and involve seizures with the risk of periodic disruptions in cognitive and behavioral functions. Predicting seizures would obviously benefits patients significantly, and also lighten physicians workload. Electroencephalography (EEG), which involves recording brain activity with electrodes placed on the scalp, has proven to be a reliable non-invasive clinical approach for epilepsy diagnosis. Predicting the eventual occurrence of seizures relies on identifying the pre-ictal period prior to the onset of a seizure [19], during which EEG recordings show different patterns from the patterns of the seizure and also from the earlier periods, so-called inter-ictal periods. As a result, the classification of inter-ictal and pre-ictal states simplifies the prediction of seizures. This study provides three contributions to the field: 1. Design of an architecture following the filter bank common spatial pattern (FBCSP) paradigm and build an explanation-producing model that biases learned filters toward relevant common sub-band frequency and spatial filters, 2. interpretation of the network abstract features encoding, by learned filters visualization and, 3. explanation of the network model decisions by layer relevance propagation. We tested the model on the CHB-MIT dataset for epilepsy prediction and its results outperformed those of the current state of the art. The model architecture showed a fair interpretability. Indeed, we found that the first layer trained filters gather data from specific frequency bands. Explanation of the model's decisions for several trials of patients with focal seizures reveals that the input channels in the brain location from which the seizure originates contribute most to the model's prediction.

The remainder of this paper is organized as follows: Section II describes the data set and the proposed architecture; Section III details the experimental results, and Section IV contains a discussion.

II. MATERIALS AND METHODS

The functional diagram of the seizure prediction task is illustrated in Figure 1. The proposed framework consists of three main steps: the first step consists of pre-processing and segmentation of the data(Section II-A). This is followed by training and evaluation of the neural network (Section II-B). The resulting models are interpreted by visualizing the learned filters and explaining the model decision for several trials (Section III).

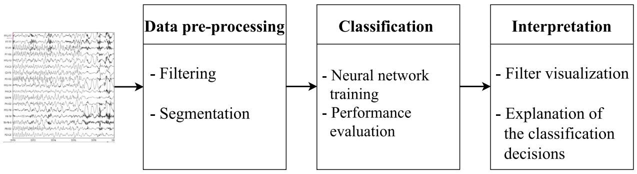

FIGURE 1. The functional diagram of the seizure prediction task.

A. DATASET DESCRIPTION AND PRE-PROCESSING

The dataset used for this study is the publicly available CHB-MIT dataset collected at the Boston Children's Hospital [20]. CHB-MIT contains 940 hours of long-term continuous multi-channel scalp EEG recordings collected from 23 pediatric patients aged 1.5 to 19 years as shown in Table 1. A minimum of 17 electrodes was used in all trials distributed according to the international standard 10/20 system. The sampling rate was set to 256Hz. Using a notch filter with an upper cutoff frequency of 50Hz and a band-pass filter with a bandwidth of 0.5-70Hz, we eliminated noise and artifacts and focused on relevant frequencies. Based on published literature, we set the pre-ictal period to be 30 minutes before the onset of the seizure, as outlined in [21], [22], and eliminated 30 minutes after the end of the seizure to exclude effects of the post-ictal periods. Subsequently, we divided the recordings into non-overlapping 5-second- windows yielding 529,415 and 66,782 samples of inter-ictal and pre-ictal activity respectively.

B. NEURAL NETWORK ARCHITECTURE

The deep neural network architecture uses the Filter Bank Common Spatial Pattern (FBCSP) algorithm [23] as follows.

FBCSP aims at finding spatial filters that map the raw data into additive components that are capable of discriminating between the sources more efficiently. FBSCP is widely used for decoding EEG data in different applications, such as brain-computer interfaces experiments [24], mental workload estimation [24], major depression detection [25], and epilepsy prediction [26], [27]. The algorithm is composed of two main components: (1) A filter bank and (2) Spatial filtering using the Common Spatial Pattern (CSP) algorithm.

- The Filter bank consists of a set of band-pass filters that separates the input signal into multiple signals, each corresponding to a unique frequency sub-band of the original input.
- The spatial filters are linear transformations which project raw channel data into a spatial space known as ‘‘source space’’ to separate sources of activity [28]. The CSP algorithm computes a transformation matrix which maximize the variance of the output signal for one class and minimize it for the other.

The outputs of the algorithm are generally used to extract features such as the log-variance of each sub-component in all sub-frequency bands. These features are then used for the classification task.

In this study, we followed the steps of this algorithm to design a neural network architecture that simplifies the interpretation of its layers. The use of FBSCP-inspired architectures for different applications has been very little studied before. The study by [24] suggested a convolutional neural network (CNN) with convolutional layers similar to the bandpass and spatial filters used to decode and visualize task-related information from EEG recordings. An alternative similar compact architecture [17] has been used to classify EEG signals from different brain-computer interface paradigms. Neither of these studies considered long, continuous EEG data. Instead, they used the simpler data of event-related potentials (ERPs), which are brain responses to a specific sensory, cognitive, or motor events. [30] have investigated a convolutional neural network architecture whose first layer function is equivalent to spatial filtering for sleep-stage classification using multivariate, multimodal continuous time-series data. However, these studies did not include classification decisions explanations or interpretations. In this paper, we present an architecture that is designed to take into account the type of long continuous EEG data. Moreover, a number of interpretations and explanations have been provided to delve further into the architecture.

Figure 2 and Table 2 show the overall diagram and the full detailed description of the proposed architecture.

- The network first layer performs a standard temporal 2D convolution that learns a set of band-pass filters to output multiple components, each representing a frequency band within the original signal. This step is equivalent to the filter bank stage of the FBCSP algorithm. With a temporal kernel that is half the sampling frequency, it is possible to capture frequency information starting at 2 Hz. Following this operation, the batch normalization is used to stabilize the training.
- The subsequent component of the architecture starts with a depth-wise convolution, which is the application of convolution filters to each feature map (output from the previous layer) independently from the other maps. To learn spatial filters, this type of convolution is implemented using kernels of shape (C, 1) where C is the number of channels. Subsequently, batch normalization, non-linear activation and average pooling were consecutively applied. A dropout layer is added to regularize the model.
- For the feature extraction from the activity source signals learned in the previous layers, we applied a combination of 2D convolutional layers, a nonlinear activation layer and an averaging layer. The outputs are finally passed through a dropout layer.
- The last stage of the architecture is a fully connected layer that flattens the features into a one-dimensional vector that is fed to a Softmax classifier.

C. NEURAL NETWORK INTERPRETATION

1) FILTER VISUALIZATION

In neural network terminology, the learned filters are simply the weights of the convolutional kernels of the network. Visualizing the learned filters allows us to see how each layer

TABLE 1. An overview of the CHB-MIT dataset.

Patient	Sex	Age	Seizure type*	Origin	# Seizure
1	F	11	SP, CP, GTC	Temporal	6
2	M	11	SP, CP	Frontal	3
3	F	14	SP, CP, GTC	Temporal	7
4	M	22	SP, CP, GTC	Temporal, Occipital	4
5	F	7	SP, CP	Frontal	4
6	F	1.5	SP, CP, GTC	Temporal	7
7	F	14.5	SP, CP, GTC	Temporal	3
8	M	3.5	SP, CP, GTC	Temporal	5
9	F	10	SP, CP	Frontal	4
10	M	3	SP, CP, GTC	Temporal	7
11	F	12	SP, CP	Frontal	3
12	F	2	SP, CP	Frontal	24
13	F	3	CP, GTC	Temporal, Occipital	12
14	F	9	SP, CP, GTC	Temporal	8
15	F	16	CP, GTC	Frontal, Temporal	20
16	M	7	SP, CP, GTC	Temporal	10
17	F	12	SP, CP, GTC	Temporal	3
18	F	18	CP, GTC	Temporal, Occipital	5
19	F	19	SP, CP	Frontal	3
20	F	6	SP, CP, GTC	Temporal	8
21	F	13	SP, CP, GTC	Temporal	4
22	F	9	CP, GTC	Temporal, Occipital	5
23	F	6	SP, CP	Frontal	8

*SP: simple partial, CP: complex partial, GTC:generalized tonic-clonic.

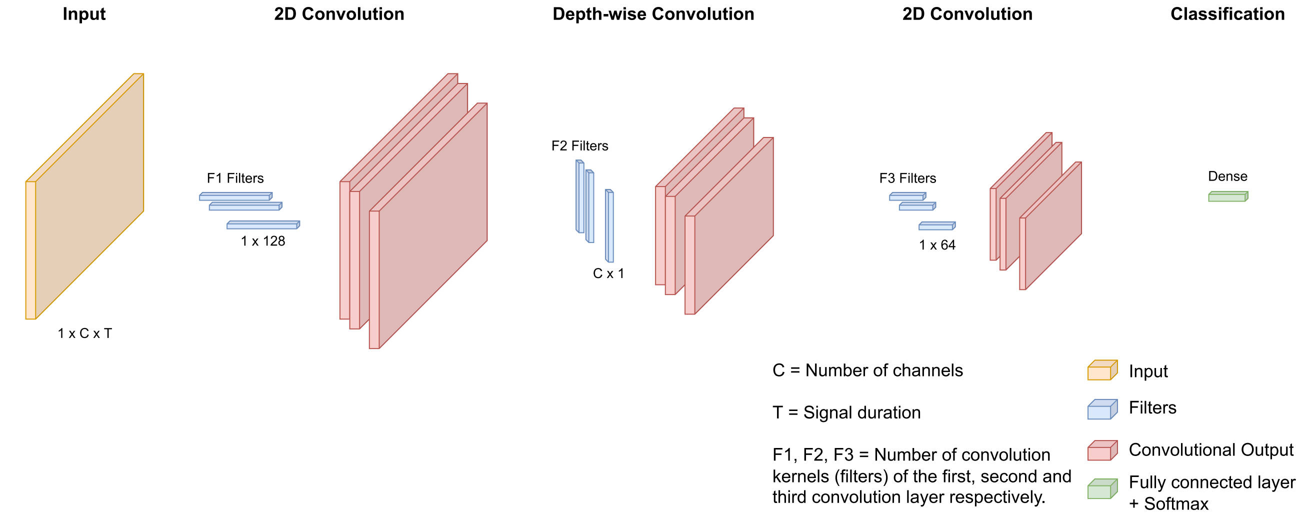

FIGURE 2. Diagram of the proposed architecture. The network processes EEG inputs with standard convolution with filters of shape (1,128) allowing learning frequency filters. It uses depth-wise convolution to learn spatial filters for each feature-map output of the previous layer separately. Finally 2D convolution is used to extract features. The outputs of the pipeline are finally fed to a fully connected layer.

extracts information from the input. Generally, for standard convolution layers, interpreting the filters proves challenging since it performs both an in-channel and in-space computation at the same time. Using the 2D convolutional filter of size (1, 128) and the depth-wise convolution to learn filters for each input channel separately, it is possible to interpret time convolution as band-pass frequency filters and depth convolution as spatial filters.

2) LAYER-WISE RELEVANCE PROPAGATION

The RLP technique explains neural network decisions by assigning a score to each input point (e.g., pixels) related to its relevance to the classification decision. RLP is based on back-propagating the prediction score $f(x)$ according to specific propagation rules (see Figure 3). The prediction score is redistributed from the output layer down to the neurons of the lower layer and so forth until it reaches the input layer. For each point in the input layer, the relevance score corresponds to its contribution to the decision. A high relevance score indicates a relevant pattern. On the other hand, parts of input with a low relevance score are considered irrelevant.

Let $f(x)$ and $R_j^{(l)}$ be the prediction score and relevance score of the neuron j in layer l respectively. Any propagation rule satisfying the following properties could be used for PRL. First, the relevance must be preserved between layers, so that

TABLE 2. The detailed architecture of the network, where C = number of channels, T = signal duration, F1 = number of convolution kernels filters to learn frequency filters, F2 = number of convolution kernels to learn spatial filters, F3 = number of convolution kernels for feature extraction, N = number of classes, respectively.

Layer	#Filter	Filter size	#Parameters	Output	Activation
Input				(1,C,T)	
2D Convolution	F1	(1,128)	128 * F1	(F1,C,T)	Linear
Batch normalization	2*F1			(F1,C,T)	
Depth-wise convolution	F2*F1	(C,1)	C*F2*F1	(F2*F1,1,T)	
Batch normalization			2*F2*F1	(F2*F1,1,T)	
Activation				(F2*F1,1,T)	Relu
Average-pooling		(1,16)		(F2*F1,1,T//16)	
Dropout				(F2*F1,1,T//16)	
2D Convolution	F3	(1,64)	64*F3	(F3,F2*F1,T')	Linear
Batch normalization			2*F3	(F3,F2*F1,T')	
Activation				(F3,F2*F1,T')	Relu
Average-pooling		(1,16)		(F3,F2*F1,T'//16)	
Dropout				(F3,F2*F1,T'//16)	
Linear (flatten)				(F3*F2*F1*(T'//16))	
Dense				N = 2	Softmax

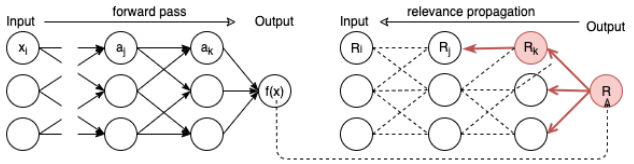

FIGURE 3. Diagram of the LRP technique. The prediction score $f(x)$ is first computed through the forward pass. Then, it is back-propagated from the output layer to the input layer according to specific rules. The scores obtained on the input layer indicate the contribution of each feature in the classification decision.

the following equation is verified.

$$f(x) = \ldots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} \tag{1}$$

Moreover, the relevance score of a node should be equal to the sum of the relevance score coming from nodes in upper layers and redistributed in same amount to nodes in lower layers as indicated in Equation 2.

$$R_j^{(l)} = \sum_k R_{j \leftarrow k}^{(l,l+1)} \text{ and } R_k^{(l+1)} = \sum_i R_{j \leftarrow k}^{(l,l+1)} \tag{2}$$

where $R_{j \leftarrow k}^{(l,l+1)}$ is the relevance score sent from the neuron k in layer l to the neuron i in the next layer $l+1$. Finally, the propagation rule must ensure that the relevance scores are related to the neuron activation or inhibition; a positive score corresponds to the existence of a feature whereas a negative or null score indicates to the absence of a pattern. There are several propagation rules that have proven effective in practice that satisfy the constraints listed above.

a: BASIC RULE (LRP − 0)

This intuitive rule redistributes the relevance score in proportion to the contribution of each input to the neuron's pre-activation.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_j a_j w_{jk}} R_k. \tag{3}$$

where a_j is the neuron activation from the previous layer and w_{jk} is the weight of the connection from unit j to unit k.

b: EPSILON RULE (LRP − ϵ)

To ensure that R_j does not take unbounded values for small or null values of neuron activation, a positive term ϵ is added to the denominator.

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_j a_j w_{jk}} R_k. \tag{4}$$

where a_j is the neuron activation from the previous layer and w_{jk} is the weight of the connection from unit j to unit k.

c: GAMMA RULE (LRP − γ)

The $LRP - \gamma$ rule denoted by the equation 5 is used to highlight the positive contributions over the negative contributions. The parameter γ controls the importance of positive evidence.

$$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^{+})}{\sum_j a_j (w_{jk} + \gamma w_{jk}^{+})} R_k. \tag{5}$$

where a_j the neuron activation from the previous layer, w_{jk} is the weight of the connection from unit j to unit k and w_{jk}^{+} is the positive part of the weight.

As suggested in [31] we used $LRP - 0$ rule for the upper layers, the Epsilon rule for the middle layer and the Gamma rule for the lower layers.

D. CLASSIFICATION AND IMPLEMENTATION DETAILS

As mentioned earlier, an inter-ictal and pre-ictal segments classification could simplify seizure prediction. However, epileptic patterns vary widely from seizure to seizure as well as from patient to patient, which makes binary classification challenging. Seizure prediction can be performed through general cross-subject models applicable to all patients or by patient-specific modeling applicable to each patient individually. Models that are patient-specific are generally impractical since it requires recording a sufficient number of seizures for each patient. Cross-subject modeling does not require treating each patient separately, but it faces the major challenge of adapting the prediction algorithm to unseen data from new patients, mainly due to the high variability of cross-subject EEG patterns [32].

In this study, we focused mostly on the patient-specific modelling to evaluate the proposed architecture. Accordingly, a single architecture was designed and trained for each subject separately. Pytorch [33] was used to implement the proposed architecture. For data pre-processing The MNE-Python package [34] was utilized. To ensure reliable generalization performance, a 5-fold stratified cross-validation test setup was used. A holdout validation is nested within a cross-validation procedure in order to further divide the training set of each fold into a validation set and a training set so that the early stopping criteria can be enforced to prevent over-fitting. In fact, the training runs up to 500 epochs, or until the validation loss remains constant for at least 20 epochs. Across all tasks, we use the gradient-based ADAM optimizer with coefficients β_1, and β_2 of 0.9 and 0.999 respectively because it is fast and reliable for reaching a global minimum. We use a learning rate of 0.005 and a dropout regularization value of 0.25.

III. RESULTS

In the following, we applied the proposed architecture to the classification of inter-ictal and pre-ictal brain states for seizure prediction. After the model training, the next step is to visualize the learning filters and explain the decisions made by the neural network using LRP-based technique.

A. PATIENT-SPECIFIC SEIZURE PREDICTION

Using the CHB-MIT data, we evaluated the proposed architecture for specific-subject seizure prediction on 23 patients. Table 3 shows some performance measures such as prediction accuracy, sensitivity, specificity, precision, F1-score, Area under the ROC Curve(AUC) and false alarm per hour for each patient model. Across all patients, the overall averaged accuracy, sensitivity, specificity and F1-score across all patients are 90.9%, 96.1%, 84.6%, and 91.9%, respectively. An averaged area under the ROC curve of 0.918% was achieved. The models have a reasonably low averaged false prediction rate per hour(FPR/h) of 0.041 indicating good predictive power.

For further evaluation of the proposed architecture, we compared our results to earlier publications that employed the same dataset, as given in Table 4. In the previous works considered, CNN architectures with different numbers of layers were used, such as the single-layer architecture as in [35], the three-layer architecture as in [36] and the five-layers architecture as in [37]. The authors of [27] used the FBCSP algorithm followed by a CNN classifier. The proposed architecture achieved the highest sensitivity with the lowest false alarm rate.

B. FILTER VISUALIZATION

Following model training, the interpretation of the learned data representation was conducted. As previously stated, the proposed architecture's first layer is supposed to be equivalent to a filter-bank. As a result, we are especially interested in seeing if the model was able to learn band-pass frequency filters. Therefore, for all subjects we visualized the filters learned on the first layer of the various patient-specific models. The convolution filter for the first layer of patient 20's model, as well as the frequency domain representation derived using the Fast Fourier Transform (FFT) are shown in Figure 4. The frequency bandwidths were calculated using the FFT. Figure 5 shows the frequency bandwidths of the seven learnt first layer filters in each of the 23 patients' subject-specific models.

C. EXPLAINING MODEL DECISION

The LRP technique is used to explain the model decision for many samples, which is the third level of interpretability explored in this work. As outlined in Section II, LRP computes, on a sample basis, the relevance scores for individual features related to their contribution to the ultimate classification decision. Positive relevance values suggest features that support the classification decision, whilst negative values indicate features that are irrelevant to the prediction. The relevance scores of individual features for successfully and inaccurately detected pre-ictal EEG samples were determined in this study. To display the topographic map, the relevance scores were averaged across time. Figure 6 shows the topographic representation of the relevance scores for various pre-ictal samples from seven patients with focal frontal seizures.

D. CROSS-SUBJECT SEIZURE PREDICTION

To evaluate the patient-independent model we used all of the data from the 23 patients at CHB-MIT. Thus, we divided the dataset into three stratified sets with the same proportions of classes: the training set, the validation set, and the test set. The proposed architecture yields satisfactory results. With a false prediction rate of 0.6/h, we were able to achieve a sensitivity of 67.17%. An F1 score of 65.84% was achieved.

IV. DISCUSSION AND CONCLUSION

This study investigated an interpretable deep learning model for seizure prediction using EEG signals. Its evaluation was conducted in three steps.

As a first step, we created an interpretable deep learning architecture whose earlier layers act according to the FBSCP scheme. The architecture was tested with a patient-specific seizure prediction task using the CHB-MIT dataset. The proposed architecture achieved a reasonably high level of prediction accuracy. Table 4 shows the benchmark of recent seizure prediction methods. Because these methods have been evaluated according to different metrics, the proposed classifier has been evaluated using several metrics commonly used in seizure prediction. From a clinical perspective, it is desirable to have a high sensitivity and a low false alarm rate. Authors of [36] proposed a three-layer CNN architecture that yielded a sensitivity of 81.% and FPR of 0.16/h as tested with 13 patients from the CHB-MIT dataset. The study in [35] adopted a more compact single-layer CNN which performs

TABLE 3. **The performance of the proposed architecture on the 23 patients of the CHB-MIT dataset.**

Patient ID	F1	Accuracy	Sensitivity	Specificity	Precision	FPR(h^{-1})	AUC
1	97.4	97.1	99.9	93.6	95.0	0.010	0.974
2	95.9	95.7	99.3	91.8	92.8	0.013	0.960
3	98.8	98.8	99.7	97.9	98.0	0.003	0.988
4	92.0	91.5	96.4	86.3	88.0	0.005	0.920
5	86.2	85.2	90.0	80.0	82.7	0.032	0.855
6	88.8	86.2	95.9	73.6	82.7	0.109	0.879
7	94.1	93.7	98.0	89.2	90.6	0.009	0.941
8	86.9	84.4	91.8	75.0	82.5	0.089	0.851
9	89.6	88.2	96.0	79.5	84.1	0.018	0.894
10	91.7	90.9	96.0	85.4	87.8	0.017	0.915
11	99.3	99.3	100.0	98.6	98.7	0.002	0.993
12	79.9	78.8	87.0	65.5	82.3	0.181	0.855
13	94.7	94.0	98.1	89.3	91.4	0.020	0.945
14	77.9	75.2	82.0	67.4	74.3	0.086	0.755
15	82.9	79.9	90.1	68.2	76.9	0.078	0.813
16	88.8	86.2	95.9	73.6	82.7	0.109	0.879
17	97.3	97.1	99.2	94.9	95.4	0.013	0.973
18	96.4	96.2	98.7	93.5	94.2	0.011	0.964
19	99.5	99.5	100.0	99.0	99.1	0.002	0.995
20	99.0	98.9	99.8	98.0	98.1	0.005	0.990
21	90.4	89.3	98.4	79.8	83.7	0.035	0.908
22	90.5	89.0	98.5	78.2	83.7	0.038	0.908
23	95.6	95.1	99.8	89.7	91.7	0.024	0.957
Average	91.9	90.9	96.1	84.7	88.5	0.040	0.918

TABLE 4. **Comparison to prior works on epileptic seizure prediction using the CHB-MIT dataset.**

Article	Data	Method	Accuracy	Sensitivity	Specificity	FPR(h^{-1})	AUC
[36]	CHB-MIT (13 patients)	CNN(3 conv)	-	81.2		0.16	-
[35]	CHB-MIT	CNN(1 conv layer)	95.6	94.2	96.9	-	-
[27]	CHB-MIT	CSP+ CNN	90	92	92	0.12	0.90
[37]	CHB-MIT (5 patients)	Bi-CNN(5 conv layer)	-	94.69	-	0.095	0.97
This work	**CHB-MIT**	**CNN inspired by FBCSP**	**90.9**	**96.1**	**84.7**	**0.040**	**0.918**

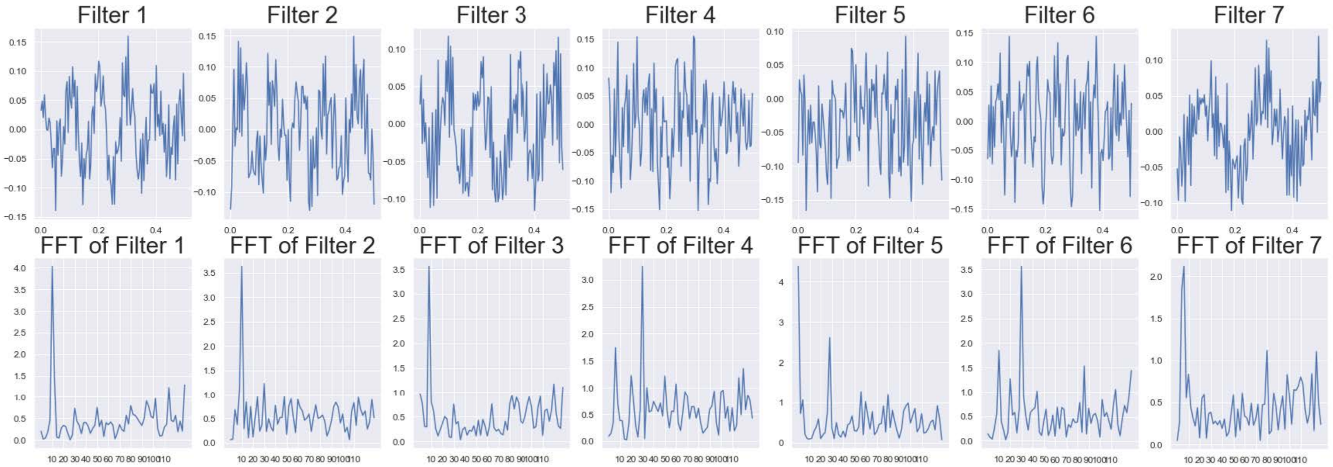

FIGURE 4. **Visualization of the learned convolution filters of the first layer of the patient 20's model. Top row shows the temporal kernels of shape (1,128) for a 0.5 window. Bottom row display the FFT calculated for each filter to determine the frequency bandwidths.**

much better, giving a better sensitivity of 94.2% a high accuracy of 95.6%, and a specificity of 96.9%. However, the evaluation is incomplete because no false alarm rate was reported. Another patient-specific CNN classifier has been described in [27]. The FBCSP algorithm was applied prior to the feature extraction step. The authors reported accuracy, specificity and sensitivity values of 90%, 92%, and 92%, respectively, with a relatively low false alarm rate of 0.12/h. A more advanced approach [36] used a five-layer one-dimensional binary convolutional neural network. They tested the model on only

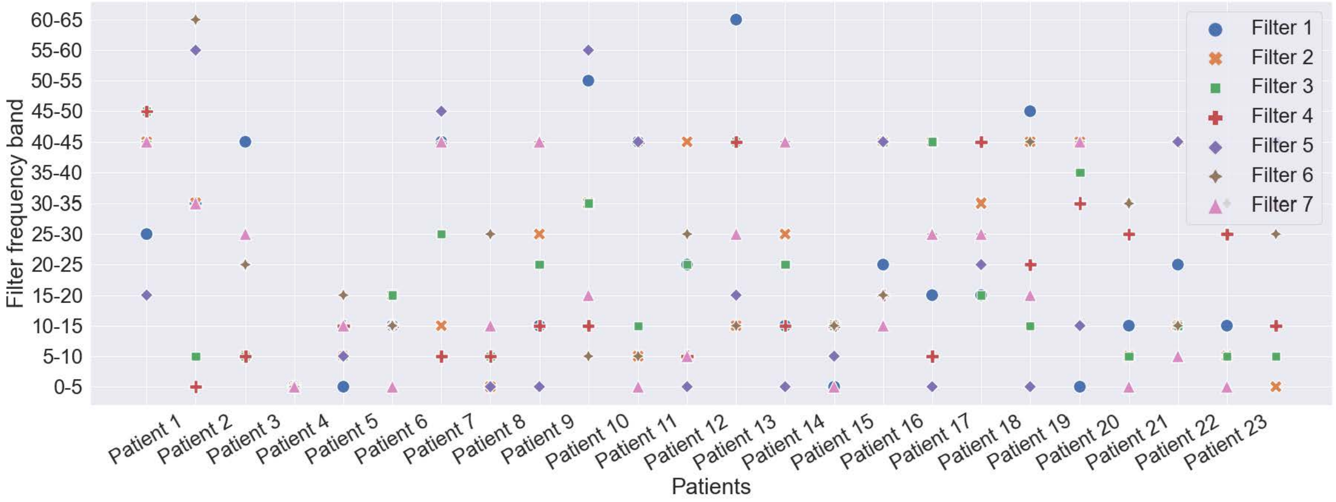

FIGURE 5. The frequency bandwidths of the seven learned filters of the first layer in all subject-specific models of the 23 patients.

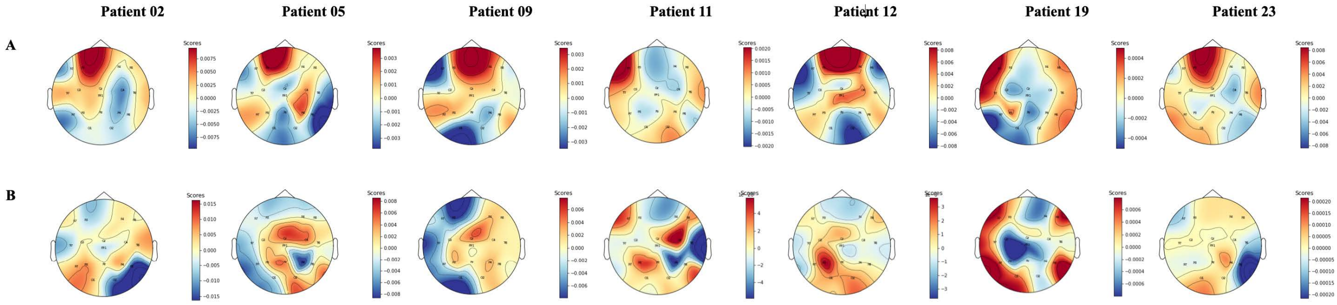

FIGURE 6. Topographical representation of relevance scores for various pre-ictal samples from seven patients with focal frontal seizures. A high relevance score implies a relevant feature, whereas a low score indicates an irrelevant input. The top row A shows the relevance scores of the correctly classified samples. The bottom B row shows the relevance scores of the miss-classified segments.

5 patients from the CHB-MIT database and their seizure prediction sensitivity averaged 94.96% with an FPR of 0.096/h. Based on the results of evaluation on all 23 patients of the CHB-mit dataset, our model reaches the highest sensitivity of 96.1 with the lowest false alarm rate of 0.041. The overall average of the areas under the receiver operating curve was 0.91. The model not only improves the results of the recent others but also enhances and simplifies its interpretation.

Our next focus was on the interpretation of the learned filters. We found that the first layer's filters were found to be similar to band-pass frequency filters and moreover, each patient-specific model has its own set of filters. Additionally, similar filters, such as those in ranges 0-5, 5-20, and 40-45 appear frequently in almost all subjects as shown in Figure 5. The model learns low- and high-frequency filters in the range 0 to 60 Hz range for each subject, which is critical for the epilepsy prediction task, since abnormal seizure discharge is primarily observed in the 5 to 50 Hz frequency range. Likewise, we found that, the models recover essential features of each patient by learning filters with a frequency of 25 Hz or higher, which is consistent with the fact that epilepsy prediction relies more on characteristics in the gamma band (30-140 Hz) that are more relevant than other bands for epilepsy prediction [29].

Finally, LRP enabled us to interpret several of the classification decisions. We found that features extracted from the channels in the region of the seizure origin were shown to be the most relevant features for pre-ictal segments classification. Hence, we determined that well classified samples with a high prediction value (Figure 6 top row) possess high relevance scores in the frontal regions where the seizure will occur, while misclassified samples (6 bottom row) displayed a distribution of relevance scores more broadly distributed throughout the scalp.

Since EEG data vary greatly between subjects and only a few patients are available, developing patient-independent models is a complex task. Therefore, most researchers simplified the problem to develop models that are patient-specific. To our knowledge, this is the first study to examine between-subjects modeling. The proposed architecture yields satisfactory results when tested on the entire dataset of the 23 patients of the CHB-MIT dataset. However, due to the substantial variability of EEG data between patients, cross-subject seizure prediction performed somewhat worse than patient-specific modelling, but the model appears to have potential applicability to data from unknown subjects.

In summary, we introduced a novel interpretable neural network architecture to simplify its opaque representation of data. The proposed architecture is based on the common FBCSP paradigm where its layers where correspond to known signal processing calculations, such as sub-frequency band and spatial filtering. The architecture performance was evaluated using the CHB-MIT dataset for the patient-specific prediction task. The proposed architecture has achieved a reasonably high predictive accuracy compared to other deep learning methods. Next, the model was interpreted by visualizing the learned filters, showing that the first-layer filters are similar to the band-pass filters. Finally, using the LRP, we were able to explain several model decisions. We observed that for the pre-ictal segments, the channels in the seizure origin region were the most relevant characteristics for classification.

The study could be strengthened by using larger amounts of data and using different EEG databases. Furthermore, it is highly useful to study how to transfer learning for cross-patient modelling, which could help to learn new representations shared between-data subjects that would transfer knowledge gained from multiple patients to new unseen patients. Finally, the proposed architecture and explanation scheme can be applied to other EEG-based classification tasks, such as seizure diagnosis and seizure type categorization, as well as autism and Alzheimer's disease detection.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1, no. 2. Cambridge, MA, USA: MIT Press, 2016.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[4] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, Dec. 2018.

[5] B. G. Buchanan and E. H. Shortliffe, "Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project," Tech. Rep., 1984.

[6] W. J. Clancey, *Knowledge-Based Tutoring: The GUIDON Program*. Cambridge, MA, USA: MIT Press, 1987.

[7] J. S. Brown, "Pedagogical, natural language, and knowlege engineering techniques in SOPHIE I, II, and III," in *Intelligent Tutoring Systems*. New York, NY, USA: Academic, 1982.

[8] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, Aug. 2001.

[9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[10] N. J. S. Morch, U. Kjems, L. K. Hansen, C. Svarer, I. Law, B. Lautrup, S. Strother, and K. Rehm, "Visualization of neural networks using saliency maps," in *Proc. Int. Conf. Neural Netw. (ICNN)*, vol. 4, 1995, pp. 2085–2090.

[11] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[12] G. Hinton, S. Osindero, M. Welling, and Y.-W. Teh, "Unsupervised discovery of nonlinear structure using contrastive backpropagation," *Cogn. Sci.*, vol. 30, no. 4, pp. 725–731, 2006.

[13] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.*, vol. 10, no. 1, 2009.

[14] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 80–89.

[15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[16] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," 2018, *arXiv:1807.03418*.

[17] 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," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[18] World Health Organization, Geneva, Switzerland. (2019). *Epilepsy*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy

[19] F. Mormann, T. Kreuz, C. Rieke, R. Andrzejak, A. Kraskov, P. David, C. Elger, and K. Lehnertz, "On the predictability of epileptic seizures," *Clin. Neurophysiol., Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 116, pp. 569–587, Mar. 2005.

[20] A. H. Shoeb and J. V. Guttag, "Application of machine learning to epileptic seizure detection," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 975–982.

[21] C. A. Teixeira, B. Direito, M. Bandarabadi, M. Le Van Quyen, M. Valderrama, B. Schelter, A. Schulze-Bonhage, V. Navarro, F. Sales, and A. Dourado, "Epileptic seizure predictors based on computational intelligence techniques: A comparative study with 278 patients," *Comput. Methods Programs Biomed.*, vol. 114, no. 3, pp. 324–336, May 2014.

[22] M. Bandarabadi, C. A. Teixeira, J. Rasekhi, and A. Dourado, "Epileptic seizure prediction using relative spectral power features," *Clin. Neurophysiol.*, vol. 126, no. 2, pp. 237–248, Feb. 2015.

[23] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 2390–2397.

[24] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[25] S.-C. Liao, C.-T. Wu, H.-C. Huang, W.-T. Cheng, and Y.-H. Liu, "Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns," *Sensors*, vol. 17, no. 6, p. 1385, Jun. 2017.

[26] T. N. Alotaiby, S. A. Alshebeili, F. M. Alotaibi, and S. R. Alrshoud, "Epileptic seizure prediction using CSP and LDA for scalp EEG signals," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–11, Oct. 2017.

[27] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 465–474, Feb. 2020.

[28] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background EEG," *Brain Topogr.*, vol. 2, no. 4, pp. 275–284, 1990.

[29] Y. Park, L. Luo, K. K. Parhi, and T. Netoff, "Seizure prediction with spectral power of EEG using cost-sensitive support vector machines," *Epilepsia*, vol. 52, no. 10, pp. 1761–1770, Oct. 2011.

[30] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018.

[31] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. 2019, pp. 193–209.

[32] I. Jemal, A. Mitiche, and N. Mezghani, "A study of EEG feature complexity in epileptic seizure prediction," *Appl. Sci.*, vol. 11, no. 4, p. 1579, Feb. 2021.

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," Tech. Rep., 2017.

[34] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers Neurosci.*, vol. 7, p. 267, Dec. 2013.

[35] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, and J. Xiang, "Epileptic seizure detection based on EEG signals and CNN," *Frontiers Neuroinform.*, vol. 12, p. 95, Dec. 2018.

[36] N. D. Truong, A. D. Nguyen, L. Kuhlmann, M. R. Bonyadi, J. Yang, S. Ippolito, and O. Kavehei, "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Netw.*, vol. 105, pp. 104–111, Sep. 2018.

[37] S. Zhao, J. Yang, Y. Xu, and M. Sawan, "Binary single-dimensional convolutional neural network for seizure prediction," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020, pp. 1–5.

IMENE JEMAL received the degree in computer science engineering from the National School of Engineering of Sfax (Enis). She is currently pursuing the Ph.D. degree in telecommunication with the Institut National de Recherche Scientifique—Centre Énergie Matériaux et Télécommunications de Montréal. In 2017, she joined the Computer Science Laboratory, University of Pau and Pays de l'Adour, as a Research Intern. Her research interests include pattern recognition, machine learning, deep learning, and explainable AI and their application to biomedical data.

NEILA MEZGHANI received the degree in telecommunications engineering from the Higher School of Telecommunications of Tunis (Sup'Com), the master's degree in information technology from the National School of Engineers of Tunis, and the Ph.D. degree from the Institut National de Recherche Scientifique—Centre Énergie Matériaux et Télécommunications, Montréal. She is currently a Data Scientist Professor with Université TELUQ (Quebec University) and a Researcher with the Centre de recherche du Centre Hospitalier de l'Université de Montréal (CR-CHUM). She is the author of two patents and about 100 peer-reviewed publications in renowned scientific journals and international conferences. Her research interests include biomedical data mining and classification, artificial intelligence, decision support systems in the medical field, and mobile health. She was the Canada Research Chair in biomedical data mining.

LINA ABOU-ABBAS received the Ph.D. degree in electrical engineering from the Ecole de Technologie Superieure, Montreal, QC, Canada, in December 2016. After her Ph.D. degree and until May 2020, she worked as a Postdoctoral Fellow with the Faculty of Medicine, Department of Neurology and Neurosurgery, McGill University. Her position was cross-appointed at the Douglas University Mental Health Institute and the McGill University Health Center. She contributed to the longitudinal behavioral data collection and organization by designing a web-based data capturing and archiving platform that allows multisite collaboration (Halifax, McMaster, and McGill Universities). In August 2020, she joined as a Researcher at LICEF Institute and the Centre de recherche du Centre hospitalier de l'Université de Montréal (CR-CHUM). Her research interest was focused on understanding the brain basis of behavioral disorders. Her current research interests include machine learning, deep learning, pattern recognition, and signal processing. She was a Core Member of the Transforming Autism Care Consortium and the Canadian Autism Neuroinformatics platform funded by Brain Canada and grouping a national and international multi-disciplinary researcher. She was awarded the Quebec Autism Research Training (QART) Program Fellowship, in 2019. She was a recipient of the Fonds du Quebec—Nature et Technologie Postdoctoral Fellowship, from 2019 to 2022.

AMAR MITICHE received the Licence Es Sciences degree in mathematics from the University of Algiers and the Ph.D. degree in computer science from The University of Texas at Austin. He is currently a Professor with the Department of Telecommunications (INRS-EMT), Institut National de la Recherche Scientique (INRS), Montreal, QC, Canada. His research interests include computer vision and pattern recognition. He has written several articles on the subjects and three books: *Computational Analysis of Visual Motion* (Plenum Press, 1994), *Variational and Level Set Methods in Image Segmentation* (Springer, 2011), with Ismail Ben Ayed, and *Computer Vision Analysis of Image Motion by Variational Methods* (Springer, 2014), with J. K. Aggarwal. His current interests include image segmentation, image motion analysis, and pattern classification by neural networks.