IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 15, 2022, accepted March 9, 2022, date of publication March 14, 2022, date of current version March 30, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3159653

Modeling Neonatal EEG Using Multi-Output Gaussian Processes

VÍCTOR CARO[1,2], JOU-HUI HO[3], SCARLET WITTING[4,5], AND FELIPE TOBAR[1,6]

[1]Initiative for Data and Artificial Intelligence, Universidad de Chile, Santiago 8370456, Chile
[2]Department of Computer Science, Universidad de Chile, Santiago 8370456, Chile
[3]Department of Electrical Engineering, Universidad de Chile, Santiago 8370448, Chile
[4]Hospital Clínico San Borja Arriarán, Pediatric Neurology, Santiago 8360160, Chile
[5]Pediatric Department, Central Campus, Facultad de Medicina, Universidad de Chile, Santiago 8380453, Chile
[6]Center for Mathematical Modelling, Universidad de Chile, Santiago 8370456, Chile

Corresponding author: Felipe Tobar (ftobar@uchile.cl)

This work was supported in part by Google and in part by Agencia Nacional de Investigación y Desarrollo de Chile (ANID) under the following grants: Fondecyt-Regular 1210606, Center for Mathematical Modeling (FB210005), and Advanced Center for Electrical and Electronic Engineering (FB0008).

ABSTRACT Neonatal seizures are sudden events in brain activity with detrimental effects in neurological functions usually related to epileptic fits. Though neonatal seizures can be identified from electroencephalography (EEG), this is a challenging endeavour since expert visual inspection of EEG recordings is time consuming and prone to errors due the data's nonstationarity and low signal-to-noise ratio. Towards the greater aim of automatic clinical decision making and monitoring, we propose a multi-output Gaussian process (MOGP) framework for neonatal EEG modelling. In particular, our work builds on the multi-output spectral mixture (MOSM) covariance kernel and shows that MOSM outperforms other commonly-used covariance functions in the literature when it comes to data imputation and hyperparameter-based seizure detection. To the best of our knowledge, our work is the first attempt at modelling and classifying neonatal EEG using MOGPs. Our main contributions are: i) the development of an MOGP-based framework for neonatal EEG analysis; ii) the experimental validation of the MOSM covariance kernel on real-world neonatal EEG for data imputation; and iii) the design of features for EEG based on MOSM hyperparameters and their validation for seizure detection (classification) in a patient specific approach.

INDEX TERMS Electroencephalography, Gaussian processes, multi-output, data imputation, seizure detection, spectral mixture kernels.

I. INTRODUCTION

Detecting neonatal seizures [1], [2] is crucial, as they may affect the development of the brain during the first four weeks of a child's life [3]. Although electroencephalography (EEG) has been validated as a *de facto* resource for the diagnosis of neonatal brain seizures, its application faces clear challenges. First, long EEG recordings lasting from several hours to days are needed to detect neonatal seizures, which motivates the urgency of automatic seizure-detection methods to aid clinical decision making [4]. Second, EEG recordings are corrupted by observation artifacts related to, e.g., muscle movements, which need to be removed to identify brain activity. Third, several methods for EEG analysis operate as an ensemble of independent single-channel models [5]–[7], however, EEG channels aggregate and share multiple sources of neural information, thus cross-correlations between channels should be considered.

Motivated by the challenges outlined above and the hypothesis that EEG data should be modelled with a multi-channel approach, we conjecture that multi-output Gaussian processes (MOGP) can be instrumental for the EEG-based detection of neonatal seizures. As a first step towards this objective, in this article we show that MOGPs [8], and specifically the multi-output spectral mixture (MOSM) kernel [9], are suitable for modelling noise-corrupted EEG signals in terms of imputation of missing data, consistency of model selection and interpretation. Furthermore, we show that the MOSM kernel outperforms other multioutput kernels in the literature under different scenarios and, additionally,

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan-Pin Lin.

TABLE 1. Some advances on ML-based analysis of neonatal EEG and a summary of their contribution. To the best of the authors knowledge, no other works have implemented MOGP kernels directly on neonatal EEG for imputation and also for seizure detection by feeding the learnt hyperparameters as features to ML classifiers.

Ref.	Brief description
[27]	EEG signals are converted into images using the Morse wavelet and then processed via texture-analysis methods. The seizures are detected via an ensemble of classifiers.
[28]	The neonatal EEG recordings are used for prediction of cognitive outcomes emerging at two years of age. Features considered are derived from weighted phase-lag index (WPLI) and a number of entropy metrics
[29]	A deep residual neural network (NN) is implemented for artefact detection on neonatal EEG. Recordings were *clamped* and filtered prior entering the NN, also the output of the NN was averaged a specific time period for each artifact to ensure predictive consistency.
[30]	ECG artifacts are removed using EEG recording by mean of independent component analysis (ICA). The detection is performed based on monitoring the similarity of the independent components found.
[31]	The severity of injury in hypoxic-ischaemic encephalophathy (HIE) is quantified via detection of Trace alternant (TA). Time-series features are feed onto an SVM classifier.
[32]	Sleep stages are classified using linear/non-linear EEG features fed into a BiLSTM network. The output of the net is post-processed using a hidden Markov model.
[33]	By feeding time-frequency features of EEG into a convolutional NN, the severity of HIE is determined from background EEG.
[15]	Seizures are detected using GPs in two ways: i) assessing posterior variance, and ii) monitoring learnt hyperparameters. Features are evaluated using a NN. Critically, this method caters from multivariate features indirectly by using ICA and not modelling cross-correlations using GPs
[34]	A neonatal EEG background classifier is developed, it application ranges from visual background scoring to classifier design. Methods in the classifier include SVM, NN, RNN, which are fed with 98 features including amplitude, complexity and oscillatory behaviour of EEG.
[35]	MOGP-based EEG binary classification (not neonatal). The MOSM kernel is used and the discrimination is performed based on the likelihood score of the test signals only, no other feature-based classifiers are considered.

by feeding MOSM's hyperparameters as feature vectors into standard machine learning classifiers, we obtain promising results for automatic seizure detection.

The article is organised as follows. Section II presents the background and previous work on MOGP and EEG modelling. Section III describes the experimental settings of our model, while Section IV shows our experimental validation on neonatal EEG imputation and seizure detection. Then, Section V presents a discussion regarding the most relevant findings in our work, in particular, we conjecture about the feasibility of using MOSM features to represent different seizure types. Lastly, Section VI concludes our study and suggests future research steps.

II. BACKGROUND

A. EEG MODELLING

Previous approaches to model-based analysis of neonatal EEG have considered parametric, e.g., autoregressive (AR) [10] or nonlinear [11] models. For instance, [10] developed an autoregressive model for neonatal EEG, [11] built on a mechanical analogy to model EEG using concepts from nonlinear dynamic systems called Duffing oscillators [12], while [13], [14] characterised multi-channel neonatal EEG from a spectral perspective. However, these methods struggle to properly account for the dynamic features of seizure-related EEG comprising fast and repeating patterns [14]. This limitation is further evidenced when the heteroscedasticity and nonstationarity of EEG recordings are taken into account.

Improved performance over parametric approaches to EEG has been achieved by nonparametric models. In [15], the authors used Gaussian processes [16] to model neonatal EEG, which allowed them to classify seizure and nonseizure data using the magnitude of the learnt variance of the noise: seizure segments are known to be more repetitive and deterministic thus having a reduced noise variance than nonseizure ones.

B. BIOSIGNAL DISEASE CLASSIFICATION

Several attempts to biosignal disease identification have been proposed in the last years. For instance, [17] studied the classification of neurological states of a driver from their EEG recordings, and [18] discriminated EEG signals coming form stroke-derived and healthy brain activity.

In the line of seizure classification, the list of standard, off-the-shelf, machine learning approaches for EEG analysis is endless. Among them, [5], [19]–[22] used SVM for classification, [6] considered a GMM operating on hand-crafted EEG features, and [7] combined both in a mixture model. Also, [23] developed a *horse race* approach combining GMM, SVM, hybrid likelihood ratio and Gaussian *supervectors*. Other features that have also been evaluated for seizure classification include: i) cross-channel Fourier transforms [24], hyperparameters of Gaussian processes [25] and parameters of a normal inverse Gaussian model [26].

On the subfield of ML-based analysis of neonatal EEG, which is the particular focus of our work, there is a large and fast-growing literature. Table 1 provides a short presentation of recent articles (published in the last two years) with the aim to illustrate the various points of view on advancing neonatal EEG analysis using ML, with special emphasis on their difference with respect to our contribution.

It can be argued that the community of ML-based EEG analysis for seizure detection has converged to stacking more and more components in an ensemble, this is particularly clear with recent neural network methods. Though this can be advantageous in practice, our approach is towards interpretability and uncertainty modelling, which we argue is critical in medical applications.

C. MULTI-OUTPUT GAUSSIAN PROCESS

The Gaussian process (GP) is a Bayesian nonparametric generative model for scalar-valued functions $f : \mathbb{R}^D \to \mathbb{R}$. GPs are particularly suited for imputation of continuous data: for a set of (possibly unevenly-sampled) observations, the

GP computes the posterior distribution over the unobserved regions.

The extension of GPs to multiple channels is referred to as *multi-output GPs* (MOGP), which, akin to its scalar counterpart, are perfectly suited for unevenly-sampled or even missing data, and provide probabilistic predictions for multi-channel signals. The choice of the covariance function is fundamental for MOGPs. Besides the choice of independent Gaussian process (IGP), where no across-channel correlation is assumed, the most popular MOGP covariance functions include the cross-spectral mixture kernel (CSM) [36], the linear model of corregionalization (LMC) [37] and the multi-output spectral mixture (MOSM) [9].

In medical data modelling, [38] presented an MOGP framework for estimating temporal dependencies across multiple sparse and irregularly-sampled medical time series. Also, [39] proposed a hierarchical GP for spatio-temporal representation of EEG and source localization. Recently, [35] proposed a MOSM-based discriminative approach to binary classification of EEG with emotional content and hand movement prediction.

D. THE MULTI-OUTPUT SPECTRAL MIXTURE KERNEL

In this work, we pay particular attention to the MOSM covariance kernel, since we conjecture that this kernel will successfully extract relevant cross-channel information from neonatal EEG data based on i) its multivariate Fourier-based construction, and ii) its validated performance in other fields, in particular EEG.

For two channels $i, j \in \{1, \ldots, m\}$ and a temporal lag τ, the MOSM kernel is defined by

$$\kappa_{ij}(\tau) = \sum_{q=1}^{Q} \alpha_{ij}^{(q)} \exp\left(\frac{1}{2}(\tau + \theta_{ij}^{(q)})^{\top} \Sigma_{ij}^{(q)} (\tau + \theta_{ij}^{(q)})\right) \cdot \cos\left(\left(\tau + \theta_{ij}^{(q)}\right) \mu_{ij}^{(q)} + \phi_{ij}^{(q)}\right), \quad (1)$$

where the superscript $(\cdot)^{(q)}$ denotes the parameters corresponding to the q^{th} component of the spectral mixture.

The hyperparameters of the MOSM kernel have clear meaning regarding the interpretation of the cross-covariance between channel (electrodes) i and j. For the q^{th} component:

- $\alpha_{ij}^{(q)}$: covariance's magnitude (energy)
- $\theta_{ij}^{(q)}$: temporal delay between channels
- $\mu_{ij}^{(q)}$: fundamental oscillatory frequency
- $\phi_{ij}^{(q)}$: the phase shift between channels
- $\Sigma_{ij}^{(q)}$: inverse lenghtscales (usually diagonal).

Additionally, we assume that the i^{th} channel is contaminated with a zero-mean Gaussian noise of variance σ_i^2, $i \in \{1, \ldots, m\}$. Furthermore, we clarify that in our approach all the above hyperparemeters (including the noise variances) are unknown *a priori* and determined from the data via maximum likelihood.

Following [9], let us observe that MOSM is a generalisation of other MOGP kernels, that is, by restricting some of its parameters MOSM can replicate other kernels in the literature. Table 2 shows how some MOPG kernels can be recovered from MOSM by applying specific parametric constraints to MOSM. Therefore, when MOSM is trained appropriately it is expected to perform equal or better than other MOGP kernels. This ability to recover other models is the main reason MOSM is the principal model considered in our work.

TABLE 2. MOGP kernels utilized in this paper as particular cases of MOSM. For M channels, indices are denoted by $i, j \in 1, \ldots, M$, and δ_{ij} denotes the Kronecker delta between channels i and j. Notation: SM-IGP (Independent GP with spectral mixture kernels), SM-LMC (linear model of coregionalisation with spectral mixture kernels) and CSM (cross-spectral mixture). Table extracted from [40].

Model	Parametric relationship with MOSM				
SM-IGP	$\alpha_{ij}^{(q)} = \omega_i \delta_{ij}$	$\mu_{ij}^{(q)} = \mu_q$	$\Sigma_{ij}^{(q)} = \Sigma_q$	$\theta_{ij}^{(q)} = 0$	$\phi_{ij}^{(q)} = 0$
SM-LMC	$\alpha_{ij}^{(q)} = \omega_{ij}^{(q)}$	$\mu_{ij}^{(q)} = \mu_q$	$\Sigma_{ij}^{(q)} = \Sigma_q$	$\theta_{ij}^{(q)} = 0$	$\phi_{ij}^{(q)} = 0$
CSM	$\alpha_{ij}^{(q)} = \sqrt{\omega_{ij}^{(q)}}$	$\mu_{ij}^{(q)} = \mu_q$	$\Sigma_{ij}^{(q)} = \Sigma_q$	$\theta_{ij}^{(q)} = 0$	–

III. METHODOLOGY

A. NEONATAL SEIZURE DATASET

We considered a public dataset of neonatal EEG recordings with seizure annotations [1]. The dataset contained multi-channel EEG recordings from 79 pre-term neonates admitted to the Neonatal Intensive Care Unit (NICU) at the Helsinki University Hospital; each recording lasted 74 minutes in average. The EEG signals were recorded at 256 Hz with 19 electrodes positioned as per the international 10-20 standard. In particular, we considered the following bipolar montage in our study: Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5 and T5-O1—see Fig. 1. These recordings were annotated by three experts for the presence or absence of seizures on each second of recording (henceforth 1-s segments), annotations for artifacts were not available in the dataset considered. In order to ensure label consistency for training the proposed model, we only considered a subset of the Helsinki NICU dataset for which all three expert labels agreed; this way, the findings reported in this work correspond exclusively to 1-s segments where the three experts agreed unanimously. Out of the 79 EEG recordings, 39 had seizures by expert consensus and thus they were considered here.

Fig. 2 shows the proportion of 1-s segments labelled as seizures by full expert agreement on the considered pool of 39 patients. Observe that the proportion of seizure to nonseizure 1-s segments vary greatly among patients. This uneven distribution of classes has to be taken into account when designing and evaluating (classification) experiments, since the considered dataset is heavily unbalanced towards samples with low proportions of seizure labels. For more detailed information regarding every patient considered in this work, we refer the reader to Table 6 (Appendix), which lists the proportion of seizures of every EEG alongside their primary seizure localization.

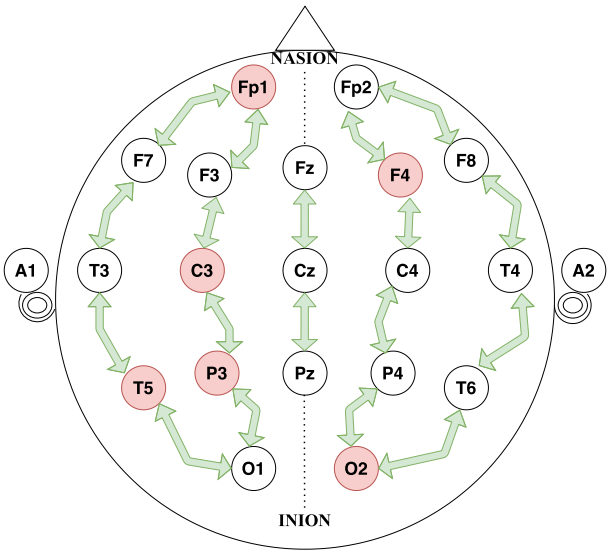

FIGURE 1. Electrode layout in the neonatal seizure EEG dataset. Red electrodes are removed in the sensor failure simulation.

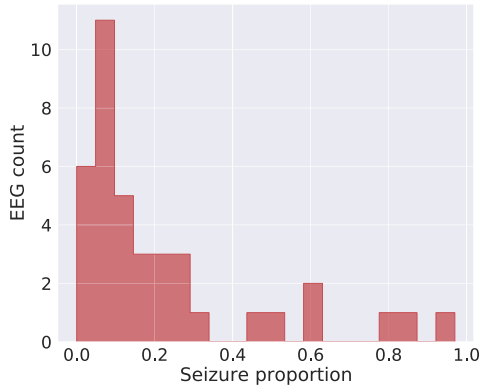

FIGURE 2. Histogram of proportion of 1-s segments labelled as seizures on the pool of 39 patients affected by seizures. It can be seen that the proportion of seizures vary significantly across the dataset.

B. DATA PRE-PROCESSING

Most MOGP kernels assume the data to be stationary. The choice of 1-second segments is also motivated for consistency with this assumption: we used 1-second EEG segments since intervals of that length can be safely considered as stationary for EEG [41]. This choice was also supported by a preliminary analysis of the spectrogram and rolling-window computation of mean and variance of the signal. Then, the data were i) filtered with a 6-order Butterworth IIR notch-filter at 50 Hz to discard powerline artifacts, ii) filtered with a 6-order Butterworth IIR high-pass filter with a cut-off frequency of 0.25 Hz so as to eliminate possible trends, iii) downsampled to 128 Hz, iv) standardised and lastly v) partitioned into 1-s EEG segments. The diagram in Figure 3 summarises the proposed methodology including the main aspects of the chosen dataset, the preprocessing stages and the two applications considered.

C. EXPERIMENTAL SETTING

We considered two practical tasks: imputation and classification of EEG data based on the dataset described above. In addition to the noise-corrupted nature of real-world data, with the aim of simulating challenging realistic environments we eliminated parts of the training data as follows:

- a randomly (uniformly) chosen 20% in each channel,
- the last 35% of channels **Fp1**, **F4**, **C3**, **T5**, **P3** and **O2** (red electrodes in Fig. 1). This configuration is used only on the data imputation task.

The second data elimination procedure is referred to as *sensor failure*, as it simulates the unwanted disconnection of an electrode or other reasons that may lead to the deletion of such recording. The aim of this recreation is to test our models in the reconstruction of complete missing EEG channels from the observed ones, a task of particular interest in wearable EEG devices [42], [43].

For the data imputation task, we considered all 39 eligible (full expert agreement) patients. For each patient, we randomly chose 15 seizure and 15 non-seizure 1-s EEG segments. This way, by fitting one model to each 30-segment set acquired from each patient, we trained 1170 CSM, SM-LMC, SM-IGP and MOSM models over 300 iterations and compared their performance under the same experimental settings.

For the classification task, we adopted a patient specific approach, that is, we trained, tested and evaluated models for each patient separately (each model only saw data from a single patient). We excluded seven patients from our analysis as the proportion of the minority class in their recordings was lower than 2% (see Table 6 and Figure 3), thus, our analysis considered only 32 (out of the 39) patients to guarantee an appropriate number of samples available for training, evaluating and testing for the patients studied. Then, we sampled a total of 16, 800 1-s segments (10, 160 nonseizure and 6, 640 seizure). In particular, we sampled 525 1-s segments from each patient according their own distribution of classes (both for validation and test), see Table 6.

For each considered 1-s segment, we trained a MOSM model over 100 iterations in order to control training times. Then, we constructed a set of features from each model as described in Section III-E. Although the fixed and limited number of iterations in training these models may result in some of them not reaching convergence, we decided to extract features from them nonetheless. The reason for this was to replicate a realistic scenario when the optimiser cannot reach the desired minima due to, e.g., unavailability of the required training time or even poorly conditioned data. The complete specification of feature and model selection are described on Sections III-E and III-F.

Training in both tasks was achieved using MOGPTK [44], a PyTorch toolkit for training MOGPs via maximum

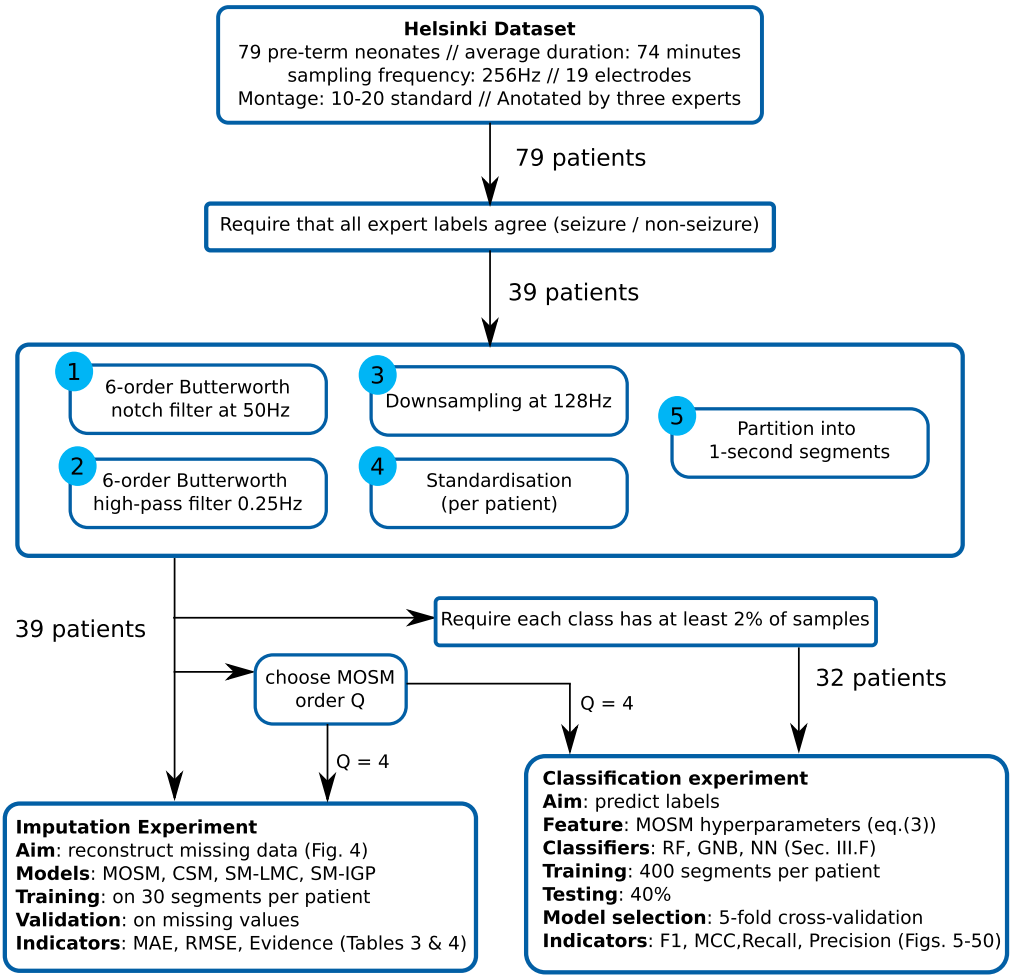

FIGURE 3. Diagram of the data processing pipeline: From dataset subselection to preprocessing of segments and the imputation / classification experiments.

likelihood. All experiments were run on an 8GB NVIDIA GeForce GTX 1080.

D. SELECTING NUMBER OF COMPONENTS Q

The number of MOSM components Q and the number of iterations I were set by grid search and cross-validation. The trained models were evaluated via the mean average error (MAE) and the normalized model evidence over $I \in \{50, 100, 500, 1000\}$ iterations and $Q \in \{1, 2, 3, 4, 5\}$ components using Adam [45]. These performance indicators, averaged over three nonseizure EEG segments, are shown in Table 3. As expected, the larger the Q the better the performance due to flexibility of more mixtures components. However, for $Q > 4$ we witnessed diminishing returns in terms of MAE, model evidence and training time, thus we chose $Q = 4$ for our experiments. Regarding the number of iterations, based on preliminary results we chose $I = 300$ for the data imputation and $I = 100$ for seizure detection in order to keep computational costs at bay.

E. FEATURE SELECTION FOR SEIZURE DETECTION

We posed the detection of seizures as a binary classification problem, where 1-s segments are classified into seizure and non-seizure ones. Inspired by EEG classification approaches using scalar GPs (e.g., [15]), we built our seizure detector using MOSM hyperparameters as features.

Based on the original construction of MOSM proposed by [9, Sec. 3.1], we considered the natural reparametrisation of the (scalar) hyperparameters defined in Sec. II-D given by

- $\Sigma_{ij} = 2\Sigma_i(\Sigma_i + \Sigma_j)^{-1}\Sigma_j$
- $\mu_{ij} = (\Sigma_i + \Sigma_j)^{-1}(\Sigma_i\mu_j + \Sigma_j\mu_i)$
- $\alpha_{ij} = \alpha_i\alpha_j \exp\left(-\frac{1}{4}(\mu_i - \mu_j)^\top(\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)\right)$
- $\theta_{ij} = \theta_i - \theta_j$
- $\phi_{ij} = \phi_i - \phi_j$,

where $i, j = 1, \ldots, m$. This way, we can use the unconstrained hyperparameters which, for the Q-component m-channel MOSM case, can be aggregated into the hyperparameter matrix

$$\Theta = \left[\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2\right] \in \mathbb{R}^{m\times(5Q+1)}, \tag{2}$$

TABLE 3. Mean absolute error (MAE) and model evidence (log marginal likelihood) for trained MOSM models with different number of components Q and iterations I. These results are averages over three nonseizure segments.

Q	Iterations	MAE	Model Evidence
Q = 1	100	0.31 ± 0.08	-0.07 ± 0.17
	300	0.32 ± 0.09	-0.04 ± 0.18
	500	0.31 ± 0.09	-0.03 ± 0.17
	1000	0.31 ± 0.09	-0.03 ± 0.17
Q = 2	100	0.17 ± 0.04	0.28 ± 0.11
	300	0.17 ± 0.03	0.37 ± 0.12
	500	0.17 ± 0.04	0.40 ± 0.17
	1000	0.16 ± 0.02	0.41 ± 0.13
Q = 3	100	0.13 ± 0.02	0.35 ± 0.07
	300	0.12 ± 0.01	0.53 ± 0.11
	500	0.13 ± 0.01	0.48 ± 0.11
	1000	0.13 ± 0.01	0.46 ± 0.13
Q = 4	50	0.14 ± 0.01	0.00 ± 0.03
	100	0.12 ± 0.01	0.37 ± 0.07
	300	0.11 ± 0.01	0.58 ± 0.10
	500	0.12 ± 0.01	0.59 ± 0.11
	1000	0.16 ± 0.06	0.41 ± 0.07
Q = 5	100	0.12 ± 0.01	0.36 ± 0.06
	300	0.17 ± 0.09	0.34 ± 0.29
	500	$\mathbf{0.10 \pm 0.01}$	$\mathbf{0.60 \pm 0.08}$
	1000	0.17 ± 0.07	0.43 ± 0.33

where we define the notation

$$\begin{aligned}
\boldsymbol{\alpha} &= \left[\alpha_i^{(q)}\right]_{i=1,q=1}^{m,\ Q} \in \mathbb{R}^{m\times Q} \\
\boldsymbol{\mu} &= \left[\mu_i^{(q)}\right]_{i=1,q=1}^{m,\ Q} \in \mathbb{R}^{m\times Q} \\
\boldsymbol{\Sigma} &= \left[\Sigma_i^{(q)}\right]_{i=1,q=1}^{m,\ Q} \in \mathbb{R}^{m\times Q} \\
\boldsymbol{\theta} &= \left[\theta_i^{(q)}\right]_{i=1,q=1}^{m,\ Q} \in \mathbb{R}^{m\times Q} \\
\boldsymbol{\phi} &= \left[\phi_i^{(q)}\right]_{i=1,q=1}^{m,\ Q} \in \mathbb{R}^{m\times Q} \\
\boldsymbol{\sigma}^2 &= \left[\sigma_i^2\right]_{i=1}^{m} \in \mathbb{R}^{m}.
\end{aligned}$$

Notice that the dimension of the parameters follows from the fact that for each one of the m channels there are $5Q$ hyperparameters corresponding to the spectral mixture and one to the noise variance.

When selecting the features, we considered all the above defined matrices flattened into one-dimensional arrays and concatenated with the trained model mean absolute error (MAE, denoted $\boldsymbol{e}$ for convenience). Therefore, following from the choice of $Q = 4$ components and $m = 16$ channels, and denoting the *flattening operator* by $\underline{\cdot}$, the considered features were:

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}} &= \begin{bmatrix}\underline{\boldsymbol{\alpha}} \\ \boldsymbol{e}\end{bmatrix} \in \mathbb{R}^{65}, \quad \tilde{\boldsymbol{\phi}} = \begin{bmatrix}\underline{\boldsymbol{\phi}} \\ \boldsymbol{e}\end{bmatrix} \in \mathbb{R}^{65}, \\
\tilde{\boldsymbol{\mu}} &= \begin{bmatrix}\underline{\boldsymbol{\mu}} \\ \boldsymbol{e}\end{bmatrix} \in \mathbb{R}^{65}, \quad \tilde{\boldsymbol{\sigma}}^2 = \begin{bmatrix}\underline{\boldsymbol{\sigma}^2} \\ \boldsymbol{e}\end{bmatrix} \in \mathbb{R}^{17}, \\
\tilde{\boldsymbol{\Sigma}} &= \begin{bmatrix}\underline{\boldsymbol{\Sigma}} \\ \boldsymbol{e}\end{bmatrix} \in \mathbb{R}^{65}, \quad \tilde{\boldsymbol{\theta}} = \begin{bmatrix}\underline{\theta} \\ \boldsymbol{e}\end{bmatrix} \in \mathbb{R}^{65},
\end{aligned}$$

$$\tilde{\boldsymbol{\Theta}} = \begin{bmatrix}\underline{\Theta} \\ \boldsymbol{e}\end{bmatrix} \in \mathbb{R}^{337}. \tag{3}$$

The inclusion of the MAE ($\boldsymbol{e}$) as a feature was considered after preliminary results on data imputation indicated that MOSM performs consistently better on seizure segments (as opposed to nonseizure ones) because they are more deterministic. Therefore, the MAE of a trained model can be considered as a proxy for determining the seizure/nonseizure label of the segment.

F. CLASSIFIERS

Our choice of hyperparameter-based features for seizure detection was validated on three standard classifiers under the aforementioned patient specific approach: Random Forest (RF), Gaussian Naive Bayes (GNB) and fully connected Neural Networks (NN), which were implemented to operate on every feature in eq. (3). Both RF and GNB were implemented via scikit-learn v1.0.2 [46] and trained as follows: we allocated 40% of the data (balanced labels according to each patient) for testing and performed model selection using 5-fold cross-validation on the remaining data.

Training and architecture selection for the NN classifier was, however, more involved. We relied on KerasTuner [47] to set the hyperparameters of the NN, while the architectures were found using the Hyperband search algorithm [48], which is also implemented on KerasTuner. We performed 4 iterations of Hyperband in each case, which proved successfully in terms of the number of NN architectures evaluated. Furthermore, in order to control the computational effort and time related to cross-validation for every possible architecture, we adopted the following procedure for every set of candidate NN architecture and feature vector in eq. (3):

1) Define an acceptance criterion based on the model's F_1-score and Matthews Correlation Coefficient (MCC).
2) Reject models that do not meet the acceptance criterion
3) Train via 5-fold cross-validation every accepted model
4) Select the model with the best average performance indicator (on cross-validation test sets).

To deal with the label imbalance in each patient, we considered class weighting in each classifier. For a patient with *n_class* 1-s segments of each *class* out of *n_total* samples, we implemented three different variants:

1) no class weighting;
2) weights adjusted to the proportion of class frequencies in training data: $\frac{n\_class}{(n\_total)}$; and
3) a scaled weight defined for each *class* as

$$\frac{1}{n\_class} * \frac{n\_total}{2}.$$

Additionally, for MLP classifiers we introduced a bias to the output layer given by $b_0 = \log(\frac{n\_seizure}{n\_nonseizure})$. Following this pipeline, we selected one model for each feature vector in eq. (3) and report their performance on both the held-out test set and average cross-validation sets in Secs. IV and VII.

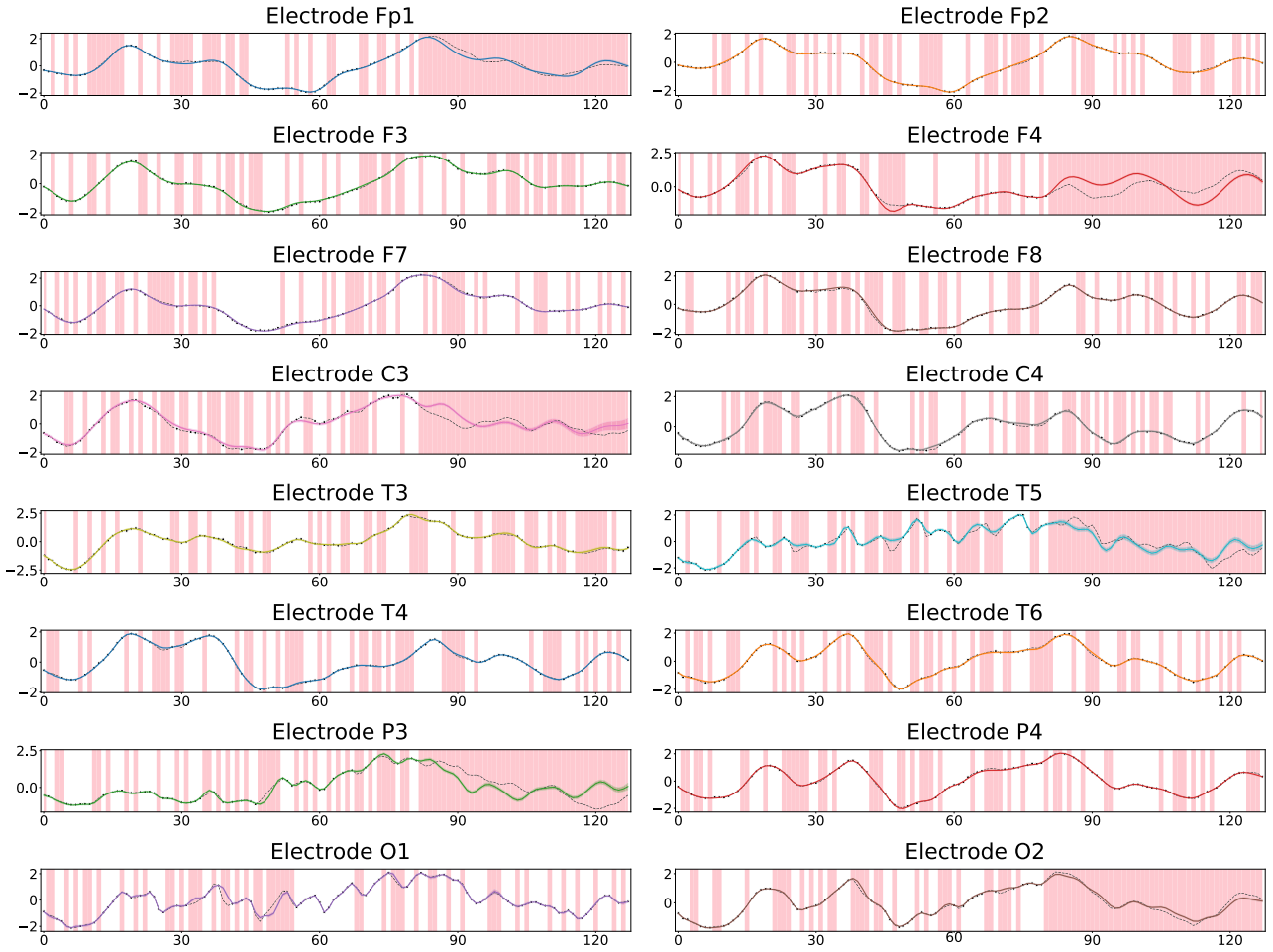

FIGURE 4. Signal reconstruction over a 1-second EEG segment using MOGP with MOSM kernel. Light-red shaded areas correspond to missing data (removed from training), training points are shown in black dots and the ground truth as black dashed lines. The coloured bands show the 95% error bars of the posterior GP prediction.

IV. RESULTS

A. DATA IMPUTATION

Fig. 4 shows the results of MOSM-based imputation in a single 1-s segment of multichannel EEG signal. Notably, the MOGP posterior recovered the ground truth signals in the regions with missing data to a fairly acceptable degree of accuracy. In general, MOSM reconstructed the EEG signal even for the cases of sensor failure, as shown by the figure for electrodes **Fp1** and **O2**.

Table 4 summarises the performance of the models averaged over all channels both for seizure and non-seizure segments in terms of MAE, root mean square error (RMSE) and model evidence. Observe that MOSM outperformed the rest of the kernels in all metrics considered both for seizure and nonseizure segments. In line with the underlying theory, all models performed worse for nonseizure data, since seizure signals are more repetitive, deterministic and thus more predictable [15].

Table 5 presents the performance (via the MAE) of data imputation for the channels affected by the simulated sensor failure. The best overall performance is once again obtained by MOSM, followed by SM, with the exception for C3 where both MOSM and SM achieved a similar performance. This was expected, since MOSM can be regarded to be the extension of SM to multiple channels, but it also highlights the representation capacity of MOSM and its ability to leverage information from multiple sources to impute missing data in any given channel.

Recall, from Fig. 1, that out of all electrodes affected by sensor failure, **C3**, **P3** and **T5** are adjacent, thus they can be considered to form a cluster of missing data for which signal reconstruction is more challeging. Electrodes **F4** or **O2**, conversely, are isolated from other *faulty* electrodes and thus they are expected to be easily recovered by their neighbouring electrodes. This explains why the performance is worse in those channels, as evidenced visually by Fig. 4 and quantitatively by Table 5.

These indicators validate the ability of MOSM to reconstruct the electrical activity on unobserved channels using observations from other channels, where the quality of the

TABLE 4. Average performance of MOGPs: mean absolute error (MAE), root mean square error (RMSE) and model evidence for different models. Each model was trained for 300 iterations. The model evidence is normalized by the total number of points. The best performing kernel is denoted by **bold font** and * indicates the second best kernel.

Model	Seizure			Nonseizure		
	MAE ↓	RMSE ↓	Model evidence ↑	MAE ↓	RMSE ↓	Model evidence ↑
MOSM	**0.20 ± 0.10**	**0.26 ± 0.13**	**0.57 ± 0.37**	**0.25 ± 0.14**	**0.33 ± 0.18**	**0.40 ± 0.40**
CSM	0.40 ± 0.17	0.50 ± 0.22	-0.34 ± 0.52	0.42 ± 0.19	0.53 ± 0.24	-0.34 ± 0.52
SM-LMC	0.34 ± 0.15	0.43 ± 0.20	0.02 ± 0.25	0.37 ± 0.16	0.47 ± 0.21	-0.07 ± 0.50
SM-IGP	0.24 ± 0.09*	0.32 ± 0.12*	0.26 ± 0.35*	0.27 ± 0.13*	0.36 ± 0.17*	0.15 ± 0.43*

TABLE 5. Model performance for the sensor failure scenario: MAE for electrodes with their last 35% data removed. For each channel, the best performing kernel is indicated in **bold font** and * indicates the second best kernel.

Channel	Seizure				Nonseizure			
	MOSM	CSM	SM-LMC	SM-IGP	MOSM	CSM	SM-LMC	SM-IGP
FP1	**0.37 ± 0.34**	0.61 ± 0.61	0.47 ± 0.54*	0.52 ± 0.32	**0.48 ± 0.42**	0.70 ± 0.63	0.57 ± 0.79	0.57 ± 0.35*
F4	**0.28 ± 0.20**	0.49 ± 0.40	0.44 ± 0.56	0.37 ± 0.23*	**0.29 ± 0.18**	0.50 ± 0.44	0.43 ± 0.40	0.35 ± 0.19*
C3	0.36 ± 0.29*	0.57 ± 0.51	0.59 ± 0.75	**0.35 ± 0.22**	0.33 ± 0.21*	0.58 ± 0.59	0.56 ± 0.68	**0.33 ± 0.19**
T5	**0.40 ± 0.38**	0.83 ± 0.74	0.68 ± 0.74	0.52 ± 0.30*	**0.46 ± 0.44**	0.80 ± 0.71	0.75 ± 1.20	0.55 ± 0.33*
P3	**0.34 ± 0.30**	0.69 ± 0.70	0.58 ± 0.59	0.38± 0.23*	**0.33 ± 0.24**	0.63 ± 0.65	0.54 ± 0.57	0.36 ± 0.23*
O2	**0.32 ± 0.27**	0.73 ± 0.75	0.48 ± 0.44*	0.52 ± 0.29	**0.39 ± 0.36**	0.75 ± 0.71	0.55 ± 0.58	0.50 ± 0.29*

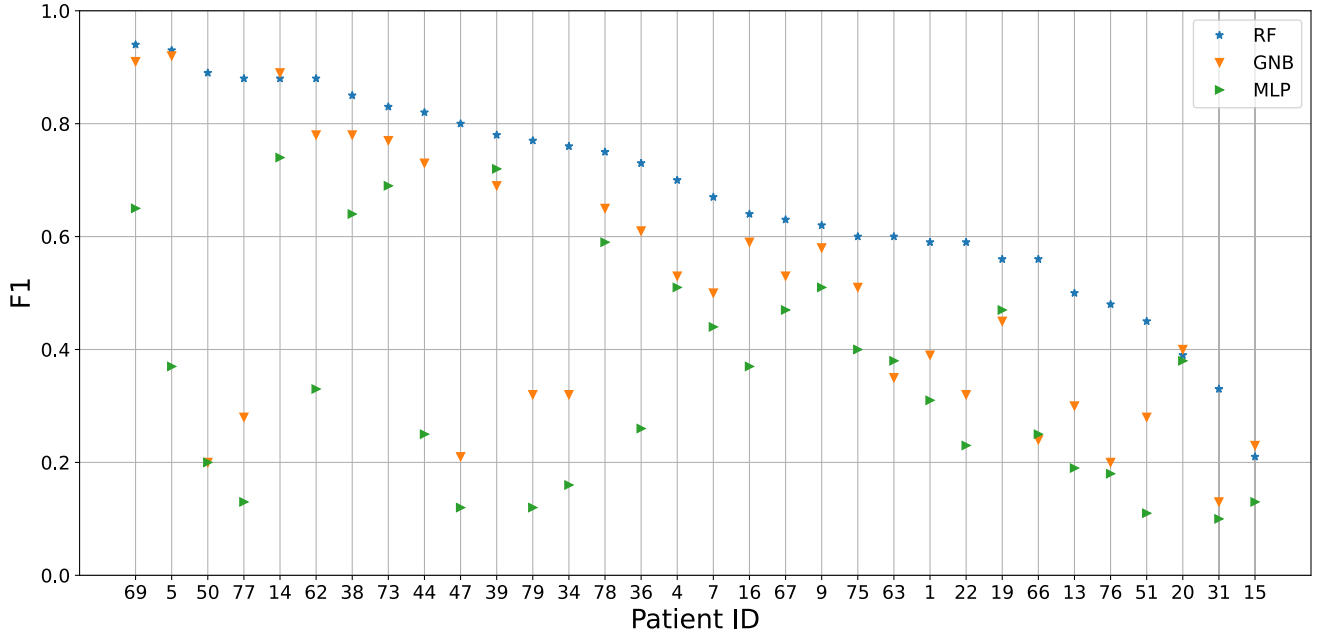

FIGURE 5. F_1-score on held-out test sets of RF, GNB, and MLP trained on MOSM feature $\tilde{\Theta}$ for each patient. The patient IDs are sorted according to the performance of RF.

reconstruction certainly depends on the proximity between the unobserved region and the observations. Additionally, this encouraging result also shows the robustness of the approach to the chosen montage.

B. SEIZURE DETECTION VIA CLASSIFICATION

For each one of the 32 patients considered, we trained GNB, NN and RF classifiers on the proposed MOSM features described in Section III-E. We chose the best GNB, RF and NN classifiers based on the maximisation of their F_1-score and MCC, and minimisation of their variance. Figure 5 compares the F_1-score of the aforementioned classifiers on the test set of each patient on MOSM feature $\tilde{\boldsymbol{\Theta}}$—see eq. (3). From the same figure, it can be seen that the best-performing family of classifiers corresponds to Random Forests (RF). Notice that both RF and GNB exhibited a similar pattern of performance suggesting that, for some patients, it was easier to automatically detect seizure events. In contrast, MLP classifiers fail to attain the same performance as RF, which could be a consequence of the restrictive size of the datasets used for training.

For RF, Figure 6 shows the performance of each proposed MOSM feature on every patient in terms of its achieved MCC on their respectives test sets. Overall, observe that the

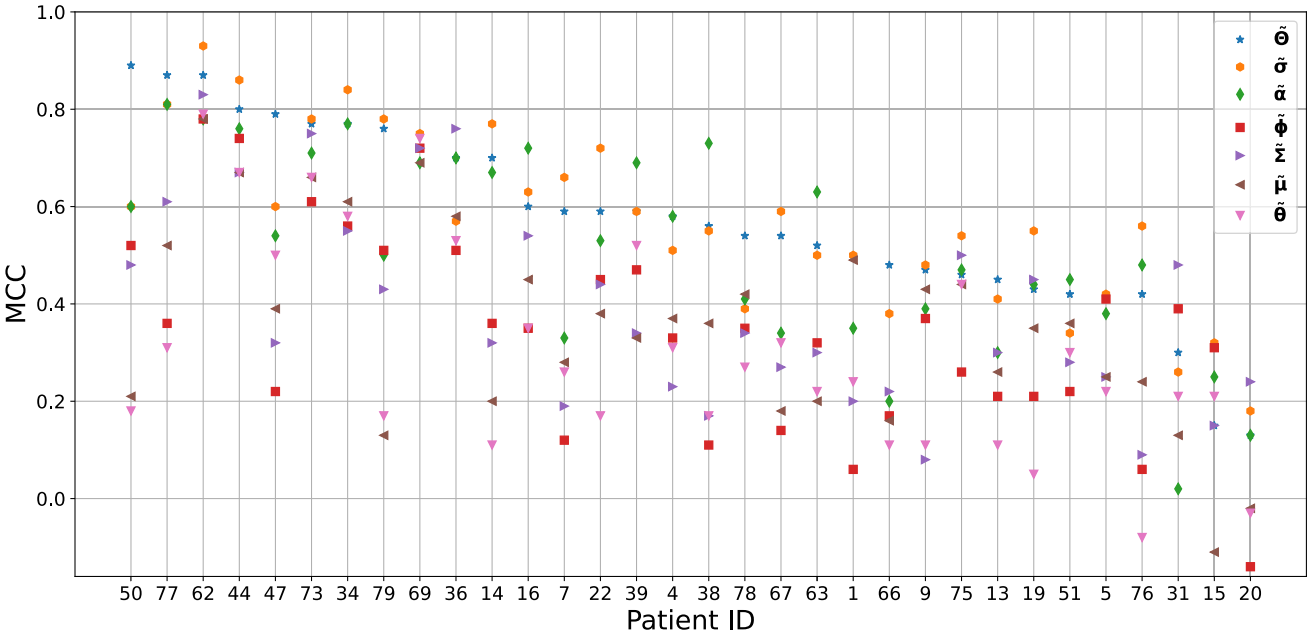

FIGURE 6. MCC on held-out test sets of RF classifiers trained on every proposed feature for each patient. The patient IDs are sorted according to the performance of RFs trained on $\tilde{\Theta}$.

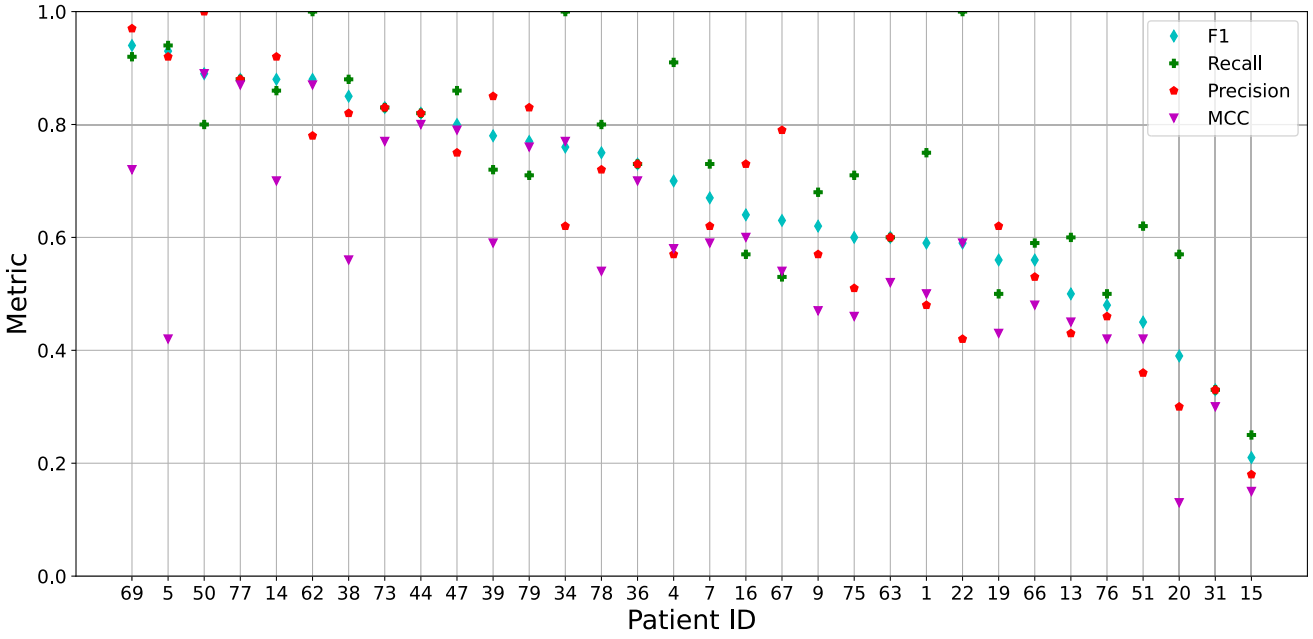

FIGURE 7. Detailed average performance on held-out test sets of RF classifiers trained on $\tilde{\Theta}$ for each patient. The patient IDs are sorted according to the F_1-score of RFs trained on $\tilde{\Theta}$.

three most successful features are $\tilde{\sigma}^2$, $\tilde{\boldsymbol{\Theta}}$ and $\tilde{\boldsymbol{\alpha}}$. Of particular interest are the results for $\tilde{\boldsymbol{\Theta}}$, as it corresponds to the concatenation of all MOSM parameters. These results suggest that the combination of different MOSM hyper-parameters effectively captures relevant information for the detection of seizures on several patients. For instance, for patients 47, 50 and 78, $\tilde{\boldsymbol{\Theta}}$ far surpasses the other proposed features. On the other hand, patients number 31 and 20, which could be considered difficult cases, performed better by training on features $\tilde{\boldsymbol{\Sigma}}$. Considering that $\tilde{\boldsymbol{\Theta}}$ contains every other proposed feature, this behavior could be a consequence of the size of the dimension of $\tilde{\boldsymbol{\Theta}}$, or perhaps the type of seizure exhibited in those patients is more adequately captured by $\tilde{\boldsymbol{\Sigma}}$, while the other features do not contribute to the appropriate detection of their seizures. An extreme case of this phenomenon is exhibited on patients 15, 20 and 76, where features $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\Sigma}}$

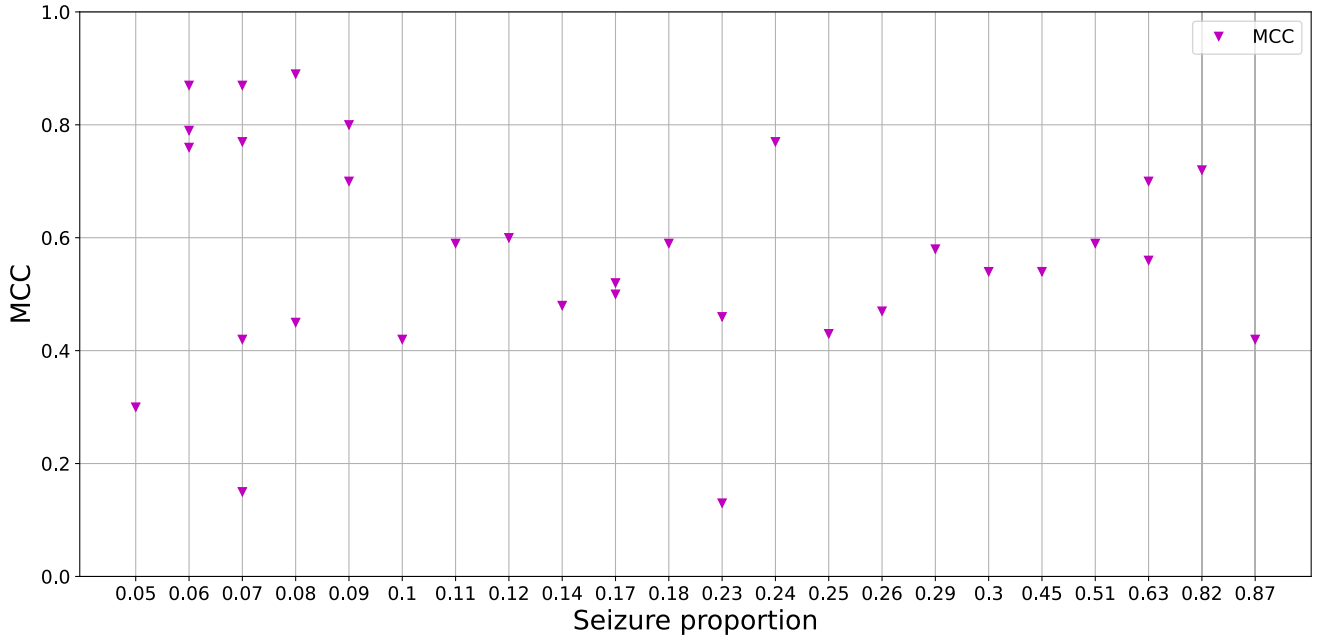

FIGURE 8. MCC on held-out test sets of RF classifiers trained on $\tilde{\Theta}$ vs seizure proportion present in each EEG recording.

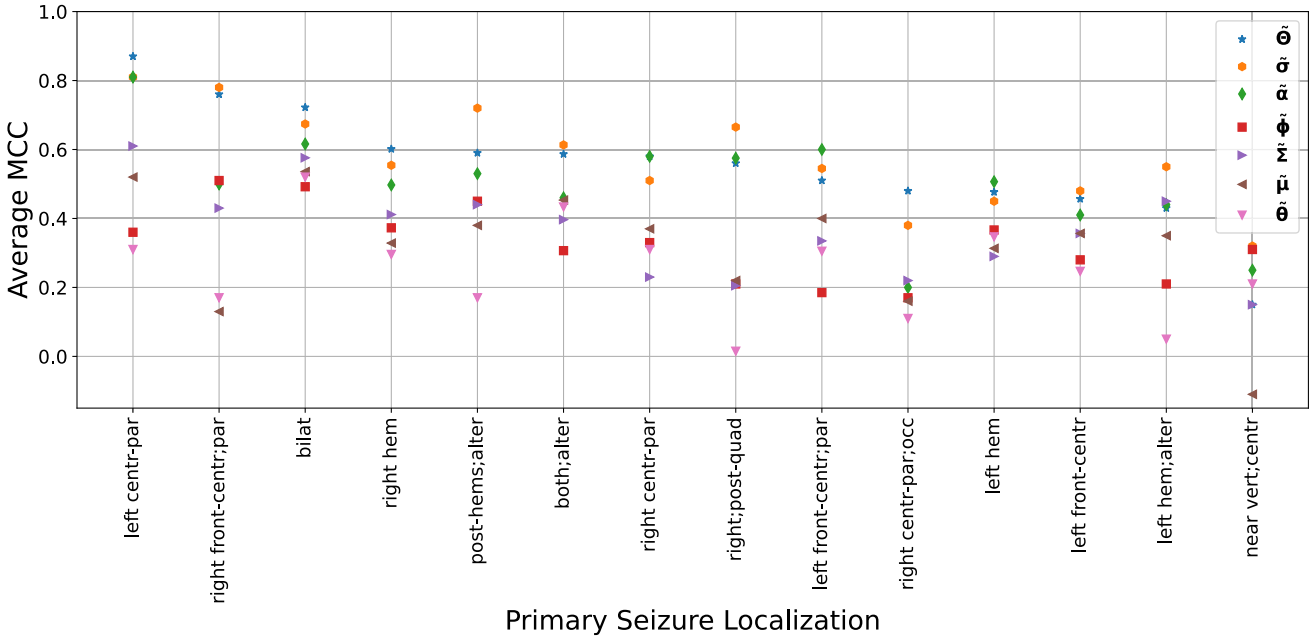

FIGURE 9. MCC of RF classifiers trained on every proposed MOSM feature aggregated according to the primary seizure localization of each EEG recording.

obtained a performance that can be considered worse than that of a baseline that predicts randomly irrespective of its input. Thus, not only do these features not contribute to the detection of their seizures, but act akin to random noise. These results suggest that, for most cases, it was possible to characterize different types of seizures via different MOSM parameters. It is, however, not possible to determine beforehand which features will be useful for an arbitrary patient, so perhaps more sophisticated classifiers could leverage the relevant information conveyed by $\tilde{\boldsymbol{\Theta}}$ with larger training datasets.

Figure 7 shows a more detailed analysis of the performance of RF on the MOSM feature $\tilde{\boldsymbol{\Theta}}$. It can be seen that, in general, the RF classifiers exhibit a balanced behavior in terms of their precision and recall values. This is particularly relevant in the context of automatic seizure detection, where it is needed to both maximize the detection of seizure episodes

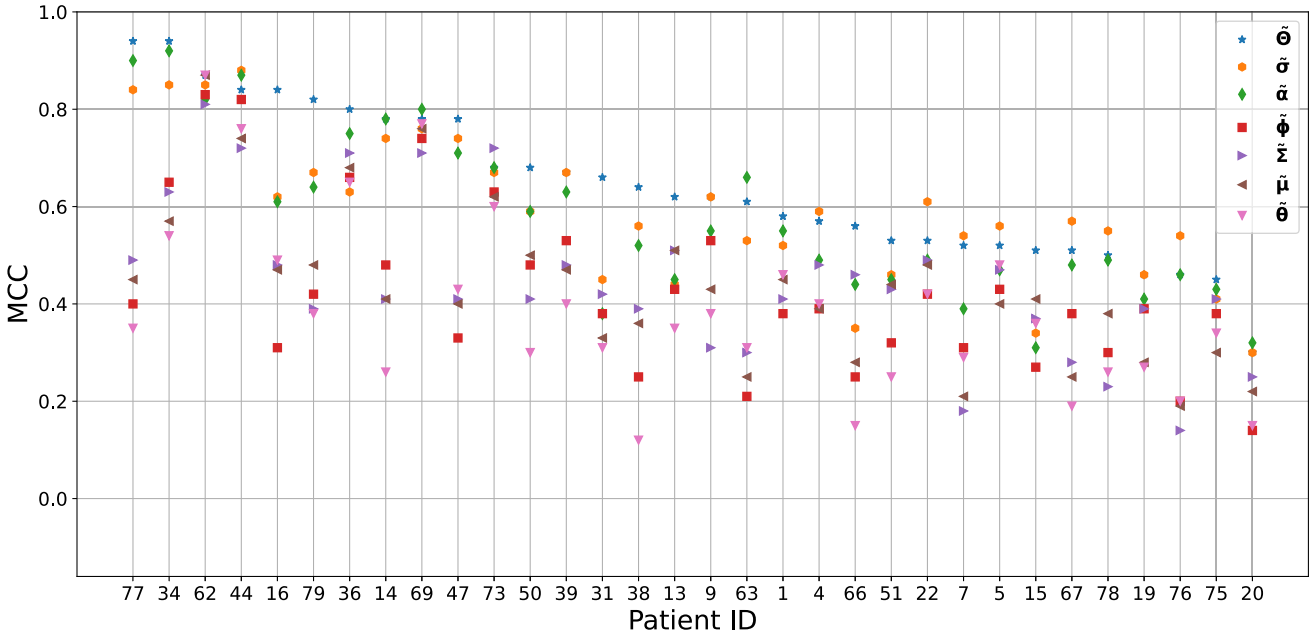

FIGURE 10. MCC on cross-validation test sets of RF classifiers trained on every proposed feature for each patient. The patient IDs are sorted according to the performance of RFs trained on $\tilde{\Theta}$.

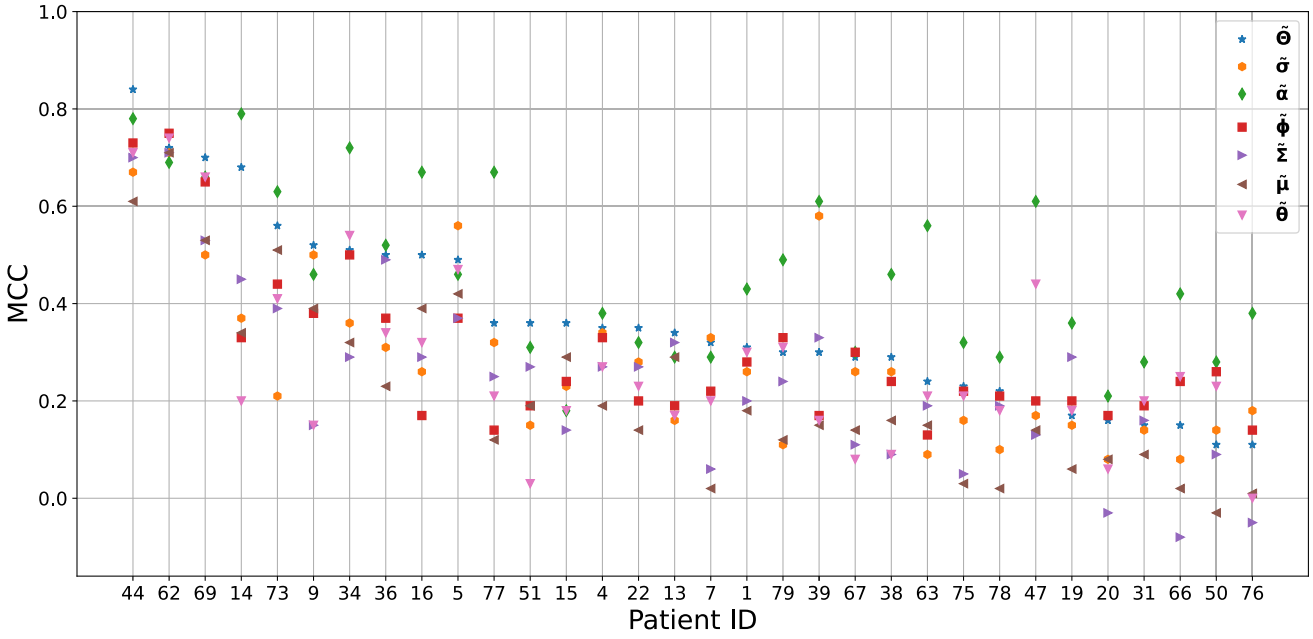

FIGURE 11. MCC on cross-validation test sets of GNB classifiers trained on every proposed feature for each patient. The patient IDs are sorted according to the performance of GNBs trained on $\tilde{\Theta}$.

and minimize the number of false positives reported for further analysis by professionals in the NICU.

Regarding the class imbalance, Table 7 (Appendix) lists the hyperparameters chosen for each RF classifier. For most cases, a *balanced* approach was beneficial to deal with the class imbalance challenge, as described in Section III-F. However, in some cases, the best result was achieved by using no class weights whatsoever, even for patients with highly imbalanced classes such as patients (compare with Table 6). Thus, there is seemingly no clear relationship between the class imbalance and the optimal weights to train RF classifiers. Furthermore, Figure 8 shows the performance of RF trained on $\tilde{\boldsymbol{\Theta}}$ vs the proportion of seizure 1-s segments in each EEG. These results, from where no apparent relationship between MCC and seizure proportion can be identified, suggest that the RF classifiers can be well trained with $\tilde{\boldsymbol{\Theta}}$

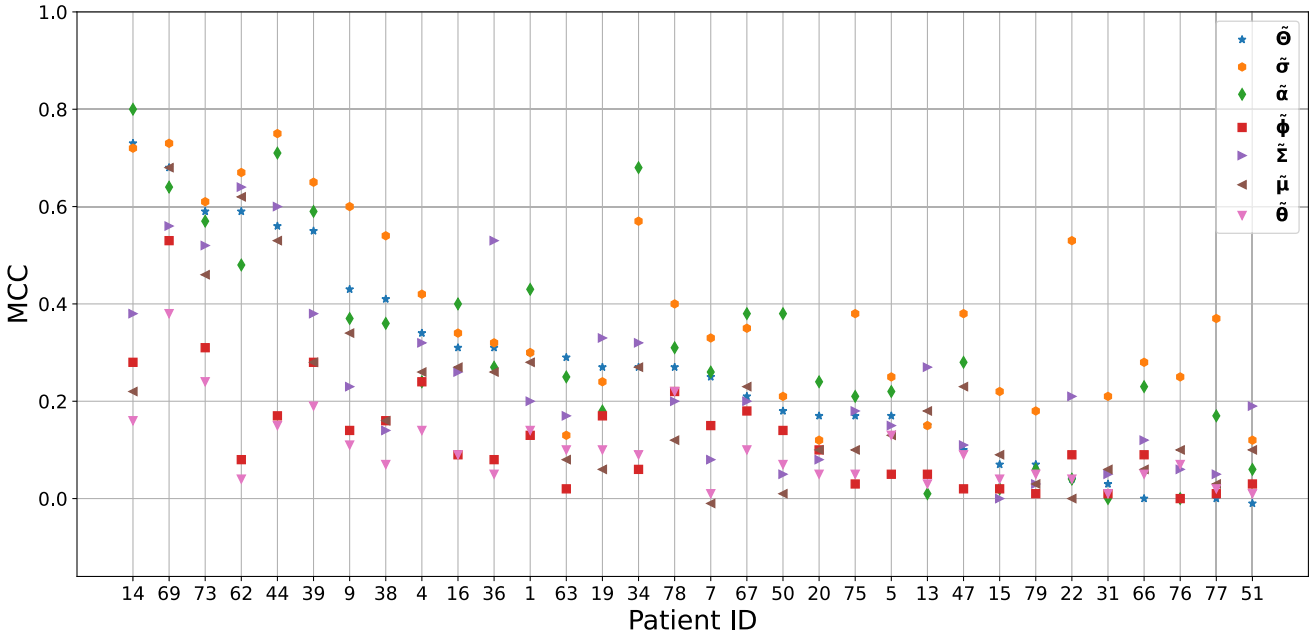

FIGURE 12. MCC on cross-validation test sets of MLP classifiers trained on every proposed feature for each patient. The patient IDs are sorted according to the performance of MLPs trained on $\tilde{\Theta}$.

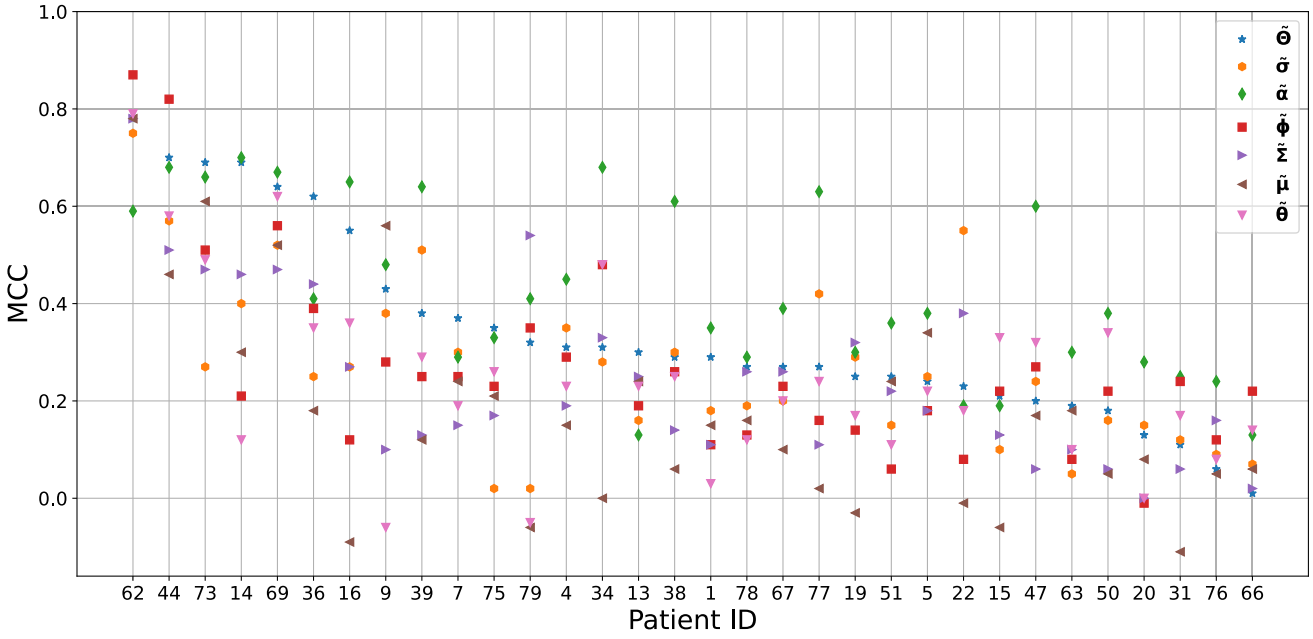

FIGURE 13. MCC on held-out test sets of GNB classifiers trained on every proposed feature for each patient. The patient IDs are sorted according to the performance of GNBs trained on $\tilde{\Theta}$.

features irrespective of the seizure proportion in the EEG recordings. In particular, notice that for patients with 7% of their 1-s segments labelled as seizures, there is an ample variance of performance in terms of MCC. This difference on performance could be a consequence of other factors such as number of artifacts present in the data (which were not labeled in the dataset) or seizure types.

Figure 9 shows the average MCC of RFs trained on each proposed feature aggregated by seizure primary localization. It can be seen that irrespective of the seizure primary localization, the best features are $\tilde{\boldsymbol{\sigma}}^2$, $\tilde{\boldsymbol{\Theta}}$ and $\tilde{\boldsymbol{\alpha}}$. There is no clear relationship between seizure primary localization and performance of MOSM features, thus validating the applicability of this approach to seizures present on arbitrary locations.

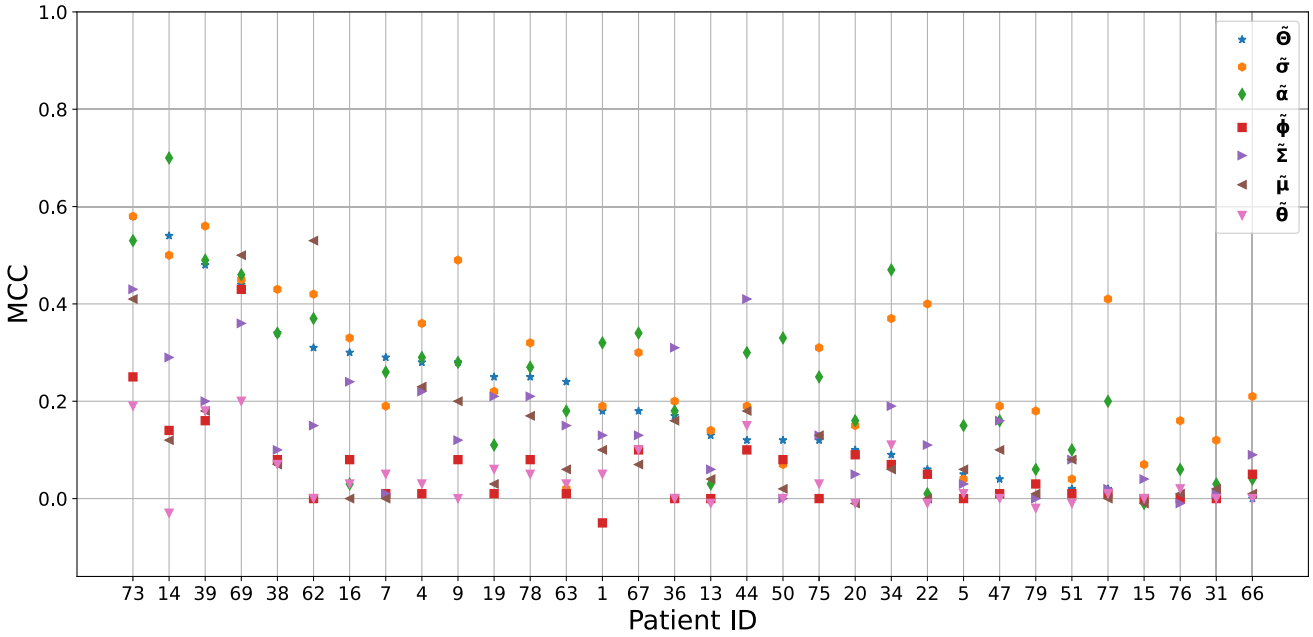

FIGURE 14. MCC on held-out test sets of MLP classifiers trained on every proposed feature for each patient. The patient IDs are sorted according to the performance of MLPs trained on $\tilde{\Theta}$.

Just as we have done in this section for the RF classifier, the performance of all three considered classifiers on every feature vector in eq. (3) is shown in the Appendix over cross-validation sets and held-out test sets. In general, there is performance consistency for validation and test sets across all performance indicators and features, this is indicative of a correct training procedure for the proposed classifiers as they exhibit a reliable performance for out of sample data.

V. DISCUSSION

Taking a patient specific approach, we investigated the ability of MOSM features to represent 1-s EEG segments within two tasks: data imputation and automatic seizure detection.

In the first task, MOSM not only was validated to perform data reconstruction in time but also across channels. This is supported by the results from the sensor failure scenario, where the missing segments were imputed using the rest of the available channels. This reveals the robustness of MOGP, and MOSM in particular, to the chosen montage, but also suggests its strength for the classification under for the cases where electrodes are faulty or unreliable.

In the seizure detection task, we used binary labels, that is, the presence or absence of seizures (and not their type or location), to empirically validate the capacity of MOSM features for EEG analysis. In our study, we found that different MOSM parameters exhibited varying degrees of relevance for different patients (see Figure 6).

In the light of these results, our conjecture is that MOSM features are capable of adequately representing different seizure types, by extending our binary treatment to the multiclass setting in a straightforward manner. In fact, as some seizure types could be well represented by $\tilde{\boldsymbol{\alpha}}$, while others by $\tilde{\boldsymbol{\sigma}}^2$ or $\tilde{\boldsymbol{\Sigma}}$, we can argue that they are of seizures of different types. Furthermore, since $\tilde{\boldsymbol{\Theta}}$ is comprised of all MOSM hyperparameters, training with larger and more thorough datasets may improve the results presented in this work, as the classifiers will be able to leverage the information present in $\tilde{\boldsymbol{\Theta}}$ more competently.

Another aspect worth of discussion is that our proposed methodology only considered short recordings with no temporal information, i.e., the time index of the segment relative to the entire recording. Though there is evidence supporting the importance of temporal information for automatic EEG graphoelement detection (e.g. [49], [50]), we chose not to include it because our focus is to test the proposed features as a proof of concept instead of deploying a fully fledged model without some initial empirical results. Thus, we believe that an interesting line of further research is to explore the effect of temporal information on the performance of more sophisticated classifiers, such as recurrent neural networks, as well as more detailed information of EEG in general.

Lastly, our main aim has been towards interpretability, since the predictions provided by an automatic seizure detector need to be fully understood from healthcare experts in practical scenarios in order to justify the presence (or absence) of a seizure from the underlying neurological theory. In that sense, though our approach does not aim to beat the state of the art in terms of classification performance, the fact that MOSM hyperparameters are interpretable in terms of signal power, correlations and delays as explained in Sec. II-D, allow field experts to understand the decisions made by particular detectors (as a matter of fact, that is precisely what our RFs are doing).

TABLE 6. Label proportion and primary seizure localization of each EEG recording used in this work. Ids marked with * were excluded from the automatic seizure classification task. The informed proportions were calculated with respect to 1-s segments under the full expert agreement criterion described in Section III-A.

EEG id	Seizure proportion	Seizure Primary Localization
1	17%	Both hemispheres; Alternating
4	29%	Right Centro-Parietal
5	87%	Left Hemisphere
7	18%	Right Hemisphere
9	26%	Left Fronto-Central
11*	1%	Right posterior quadrant
13	8%	Bilateral
14	63%	Right; Posterior quadrant
15	7%	Near Vertex; Central
16	12%	Right Hemisphere
17*	1%	Near Vertex; Central
19	25%	Left Hemisphere; Alternating
20	23%	Left Fronto-Central
21*	1%	Right Hemisphere
22	11%	Both Posterior Hemispheres; Alternating
25*	1%	Both Hemispheres; Alternating
31	5%	Right Hemisphere
34	7%	Right Hemisphere
36	9%	Bilateral
38	63%	Left Fronto-Central; Parietal
39	51%	Left Hemisphere
40*	2%	Left Parietal-Occipital
41*	98%	Right Hemisphere
44	9%	Bilateral
47	6%	Bilateral
50	8%	Right Hemisphere
51	7%	Left Hemisphere
52*	2%	Bilateral
62	6%	Bilateral
63	17%	Right Hemisphere
66	14%	Right Centro-Parietal; Occipital
67	30%	Both Hemispheres; Alternating
69	82%	Both Hemispheres; Alternating
73	24%	Left Fronto-Central
75	23%	Left Fronto-Central; Parietal
76	10%	Right Hemisphere; Posterior quadrant
77	7%	Left Fronto-Central
78	45%	Right Hemisphere
79	6%	Right Fronto-Central; Parietal

TABLE 7. Hyperparameters selected for RF classifiers trained on $\tilde{\Theta}$ according to their performance on cross-validation test sets for each patient. The nomenclature for Class Weight is as follows: (1) *None* means no class weighting was used; (2) *Balanced* means that class weights were adjusted to the proportion of class frequencies in training data; (3) *Scaled* means that class weights were scaled as indicated in Section III-F. For max depth, *None* means that the nodes were expanded until all leaves were pure or until all leaves contain less than 2 samples. Recall that the Random Forest classifiers were implemented on scikit-learn version 1.0.2, and thus we refer the reader to its documentation for further clarification of the considered hyperparameters.

EEG id	Number of Trees	Max depth of the tree	Class Weight
1	50	*None*	*Scaled*
4	200	2	*Balanced*
5	300	*None*	*Balanced*
7	300	6	*None*
9	1500	2	*Balanced*
13	100	*None*	*Balanced*
14	500	*None*	*None*
15	1500	10	*Scaled*
16	200	6	*None*
19	500	20	*None*
20	300	2	*Balanced*
22	100	4	*None*
31	100	4	*None*
34	50	6	*Balanced*
36	300	*None*	*Balanced*
38	1500	2	*None*
39	300	20	*Balanced*
44	500	20	*Balanced*
47	100	10	*Balanced*
50	200	4	*Balanced*
51	200	20	*Scaled*
62	100	6	*None*
63	200	*None*	*Scaled*
66	500	20	*None*
67	200	20	*Balanced*
69	100	6	*Balanced*
73	300	20	*Balanced*
75	100	6	*Balanced*
76	300	2	*None*
77	200	6	*None*
78	100	4	*Scaled*
79	1000	*None*	*Balanced*

VI. CONCLUSION

We have validated the hypothesis that EEG analysis can benefit from multivariate models, rather than a ensemble of single-channel ones. In particular, we have proposed an MOGP framework for analysing neonatal EEG with application to data imputation, even for large missing regions arising from sensor failure. This is a crucial point, specially on neonatal EEG, where artifacts, and thus loss of data, are abundant. We have also empirically demonstrated the superiority of the MOSM covariance kernel in this task, which has outperformed standard MOGP kernels owing to its ability for cross-signal covariance modelling.

In addition to the data-imputation task, we have obtained supporting evidence as to the suitability of hyperparameter-based seizure detection using MOSM models; we implemented standard ML classifiers on the proposed MOSM features and show extensive quantitative validation of their performance. The superiority of the proposed model is its interpretability through the designed MOSM features, and though there are NN-based methods in the state of the art that have exhibited a unique classification performance, our contribution is towards informed and reliable clinical decision making.

To the best of our knowledge, the work presented here is the first attempt at modelling and classifying neonatal EEG with MOGPs and, therefore, the first validation of the MOSM kernel on such setting. In the same fashion as those of [35], our results are auspicious and promising as a jumping off point for reliable MOGP-based seizure detection mechanisms deployed at real-world clinical environments. Although in this work we only considered seizure/nonseizure classification, the proposed method can be easily extended to seizure type classification with suitable datasets by means of multilabel classifiers.

In the authors perspective, further research work for a fully-automated seizure detection framework should include the following aspects. First, sparse GPs [51]–[53] specially

suited for MOGPs to control the computational costs. Second, flexible kernels [54], [55] and non-Gaussian trainable likelihoods to replicate the known features of EEG signals [56], [57]. Third, an automatic identification of (quasi) stationary segments possibly based on the generalized likelihood ratio [58], stationary subspace analysis [59], or non-stationary MOGPs [60]. Fourth, further studies into hybrid approaches combining interpretable probabilistic generative models as the MOGP considered in this work together with deep neural networks that sacrifice interpretation for performance; combining these two approaches can lead to models that combine are, at the same time, sufficiently interpretable an high performing.

Lastly, some key open questions that stem from our study include: 1) is successful data imputation with MOSM related to improved performance in the automatic classification task?; 2) is it possible to leverage the proposed approach for automatic artifact detection in order to recover data lost to such artifacts. This could allow for MOSM to be established as a multi-purpose tool within the EEG analysis framework.

APPENDIX

A. TABLES

B. ADDITIONAL PERFORMANCE INDICATORS

The following figures show the MCC for all three classifiers both for 5-fold cross-validation (CV) and held-out test sets:

- Fig. 10: Random Forest in 5-fold CV
- Fig. 11: Gaussian Naive Bayes in 5-fold CV
- Fig. 12: Multilayer Perceptron in 5-fold CV
- Fig. 13: Gaussian Naive Bayes in held-out set
- Fig. 14: Multilayer Perceptron in held-out set

Recall that the MCC of Random Forest in held-out set is shown in Fig. 6.

REFERENCES

[1] N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal EEG recordings with seizure annotations," *Sci. Data*, vol. 6, no. 1, pp. 1–8, Mar. 2019.
[2] J. J. P. Alix, A. Ponnusamy, E. Pilling, and A. R. Hart, "An introduction to neonatal EEG," *Paediatrics Child Health*, vol. 27, no. 3, pp. 135–142, Mar. 2017.
[3] C. P. Panayiotopoulos, *The Epilepsies: Seizures, Syndromes and Management*. Oxfordshire, U.K.: Bladon Medical, 2005.
[4] A. Temko and G. Lightbody, "Detecting neonatal seizures with computer algorithms," *J. Clin. Neurophysiol.*, vol. 33, no. 5, pp. 394–402, Oct. 2016.
[5] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 73–464, 2011.
[6] E. Thomas, A. Temko, G. Lightbody, W. Marnane, and G. Boylan, "Gaussian mixture models for classification of neonatal seizures using EEG," *Physiol. Meas.*, vol. 31, p. 1047, Jun. 2010.
[7] A. Temko, G. Lightbody, E. M. Thomas, G. B. Boylan, and W. Marnane, "Instantaneous measure of EEG channel importance for improved patient-adaptive neonatal seizure detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 717–727, Mar. 2012.
[8] M. A. Alvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," *Found. Trends Mach. Learn.*, vol. 4, no. 3, pp. 195–266, 2012.
[9] G. Parra and F. Tobar, "Spectral mixture kernels for multi-output Gaussian processes 30," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2017.
[10] G. Mohammadi, P. Shoushtari, B. M. Ardekani, and M. B. Shamsollahi, "Person identification by using AR model for EEG signals," in *Proc. World Acad. Sci., Eng. Technol.*, vol. 11, 2006, pp. 281–285.
[11] N. J. Stevenson, M. Mesbah, G. B. Boylan, P. B. Colditz, and B. Boashash, "A nonlinear model of newborn EEG with nonstationary inputs," *Ann. Biomed. Eng.*, vol. 38, no. 9, pp. 3010–3021, Sep. 2010.
[12] R. Srebro, "The duffing oscillator: A model for the dynamics of the neuronal groups comprising the transient evoked potential," *Electroencephalogr. Clin. Neurophysiol./Evoked Potentials Sect.*, vol. 96, no. 6, pp. 561–573, Nov. 1995.
[13] L. Rankine, N. Stevenson, M. Mesbah, and B. Boashash, "A nonstationary model of newborn EEG," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 1, pp. 19–28, Jan. 2007.
[14] M. F. Al-Sa'd and B. Boashash, "Design and implementation of a multi-sensor newborn EEG seizure and background model with inter-channel field characterization," *Digit. Signal Process.*, vol. 90, pp. 71–99, Jul. 2019.
[15] S. Faul, G. Gregorcic, G. Boylan, W. Marnane, G. Lightbody, and S. Connolly, "Gaussian process modeling of EEG for the detection of neonatal seizures," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 12, pp. 2151–2162, Dec. 2007.
[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[17] I. Hussain, S. Young, and S.-J. Park, "Driving-induced neurological biomarkers in an advanced driver-assistance system," *Sensors*, vol. 21, no. 21, p. 6985, Oct. 2021.
[18] I. Hussain and S.-J. Park, "Quantitative evaluation of task-induced neurological outcome after stroke," *Brain Sci.*, vol. 11, no. 7, p. 900, Jul. 2021.
[19] R. Sharma and K. Chopra, "EEG-based epileptic seizure detection using GPLV model and multi support vector machine," *J. Inf. Optim. Sci.*, vol. 41, no. 1, pp. 143–161, Jan. 2020.
[20] K. T. Tapani, S. Vanhatalo, and N. J. Stevenson, "Time-varying EEG correlations improve automated neonatal seizure detection," *Int. J. Neural Syst.*, vol. 29, no. 4, May 2019, Art. no. 1850030.
[21] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. B. Boylan, "Performance assessment for EEG-based neonatal seizure detectors," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 474–482, Mar. 2011.
[22] A. Temko, G. Boylan, W. Marnane, and G. Lightbody, "Robust neonatal EEG seizure detection through adaptive background modeling," *Int. J. Neural Syst.*, vol. 23, no. 4, p. 1350018, Aug. 2013.
[23] A. Temko, A. Sarkar, and G. Lightbody, "Detection of seizures in intracranial EEG: UPenn and mayo Clinic's seizure detection challenge," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 6582–6585.
[24] W. Bomela, S. Wang, C.-A. Chou, and J.-S. Li, "Real-time inference and detection of disruptive EEG networks for epileptic seizures," *Sci. Rep.*, vol. 10, no. 1, p. 8653, May 2020.
[25] M. Zhong, F. Lotte, M. Girolami, and A. Lécuyer, "Classifying EEG for brain computer interfaces using Gaussian processes," *Pattern Recognit. Lett.*, vol. 29, no. 3, pp. 354–359, Feb. 2008.
[26] A. B. Das, M. I. H. Bhuiyan, and S. M. S. Alam, "Classification of EEG signals using normal inverse Gaussian parameters in the dual-tree complex wavelet transform domain for seizure detection," *Signal, Image Video Process.*, vol. 10, no. 2, pp. 259–266, Feb. 2016.
[27] M. Diykh, F. S. Miften, S. Abdulla, R. C. Deo, S. Siuly, J. H. Green, and A. Y. Oudahb, "Texture analysis based graph approach for automatic detection of neonatal seizure from multi-channel EEG signals," *Measurement*, vol. 190, Feb. 2022, Art. no. 110731.
[28] N. Alotaibi, D. Bakheet, D. Konn, B. Vollmer, and K. Maharatna, "Cognitive outcome prediction in infants with neonatal hypoxic-ischemic encephalopathy based on functional connectivity and complexity of the electroencephalography signal," *Frontiers Hum. Neurosci.*, vol. 15, Jan. 2022, doi: 10.3389/fnhum.2021.795006.
[29] L. Webb, M. Kauppila, J. A. Roberts, S. Vanhatalo, and N. J. Stevenson, "Automated detection of artefacts in neonatal EEG with residual neural networks," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106194.
[30] G. Tamburro, F. Croce, F. Zappasodi, and S. Comani, "Automated detection and removal of cardiac and pulse interferences from neonatal EEG signals," *Sensors*, vol. 21, no. 19, p. 6364, Sep. 2021.
[31] S. A. Raurale, G. B. Boylan, S. R. Mathieson, W. P. Marnane, G. Lightbody, and M. J. O'Toole, "Tracé alternant detector for grading hypoxic-ischemic encephalopathy in neonatal EEG," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 1177–1181.

[32] H. Ghimatgar, K. Kazemi, M. S. Helfroush, K. Pillay, A. Dereymaker, K. Jansen, M. D. Vos, and A. Aarabi, "Neonatal EEG sleep stage classification based on deep learning and HMM," *J. Neural Eng.*, vol. 17, no. 3, Jun. 2020, Art. no. 036031.

[33] S. A. Raurale, G. B. Boylan, S. R. Mathieson, W. P. Marnane, G. Lightbody, and J. M. O'Toole, "Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time–frequency distributions," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 046007.

[34] S. M. Moghadam, E. Pinchefsky, I. Tse, V. Marchi, J. Kohonen, M. Kauppila, M. Airaksinen, K. Tapani, P. Nevalainen, C. Hahn, E. W. Y. Tam, N. J. Stevenson, and S. Vanhatalo, "Building an open source classifier for the neonatal EEG background: A systematic feature-based approach from expert scoring to clinical visualization," *Frontiers Hum. Neurosci.*, vol. 15, May 2021, doi: 10.3389/fnhum.2021.675154.

[35] C. Torres-Valencia, Á. Orozco, D. Cárdenas-Peña, A. Álvarez-Meza, and M. Álvarez, "A discriminative multi-output Gaussian processes scheme for brain electrical activity analysis," *Appl. Sci.*, vol. 10, no. 19, p. 6765, Sep. 2020.

[36] K. R. Ulrich, D. E. Carlson, K. Dzirasa, and L. Carin, "GP kernels for cross-spectrum analysis," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2015, pp. 1999–2007.

[37] P. Goovaerts, *Geostatistics for Natural Resource Evaluation*, vol. 42. Oxford, U.K.: Oxford Univ. Press, 1997.

[38] L.-F. Cheng, B. Dumitrascu, G. Darnell, C. Chivers, M. Draugelis, K. Li, and B. E. Engelhardt, "Sparse multi-output Gaussian processes for online medical time series prediction," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–23, Dec. 2020.

[39] D. Kuzin, O. Isupova, and L. Mihaylova, "Spatio-temporal structured sparse regression with hierarchical Gaussian process priors," *IEEE Trans. Signal Process.*, vol. 66, no. 17, pp. 4598–4611, Sep. 2018.

[40] T. D. Wolff, A. Cuevas, and F. Tobar, "Gaussian process imputation of multiple financial series," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8444–8448.

[41] F. Gonen and G. Tcheslavski, "Techniques to assess stationarity and Gaussianity of EEG: An overview," *Int. J. Bioautomation*, vol. 16, no. 2, p. 135, 2012.

[42] K. B. Mikkelsen, S. L. Kappel, D. P. Mandic, and P. Kidmose, "EEG recorded from the ear: Characterizing the ear-EEG method," *Front Neurosci.*, vol. 9, p. 438, Nov. 2015.

[43] D. Looney, P. Kidmose, C. Park, M. Ungstrup, M. L. Rank, K. Rosenkranz, and P. D. Mandic, "The in-the-ear recording concept: User-centered and wearable brain monitoring," *IEEE Pulse*, vol. 3, no. 6, pp. 32–42, Nov. 2012.

[44] T. de Wolff, A. Cuevas, and F. Tobar, "MOGPTK: The multi-output Gaussian process toolkit," *Neurocomputing*, vol. 424, pp. 49–53, Feb. 2021.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[46] 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," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[47] T. O'Malley *et al.* (2019). *KerasTuner*. [Online]. Available: https://github.com/keras-team/keras-tuner

[48] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *J. Mach. Learn. Res.*, vol. 18, no. 185, pp. 1–52, 2018. [Online]. Available: http://jmlr.org/papers/v18/16-558.html

[49] A. Shoeibi, M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, M. Panahiazar, F. Khozeimeh, A. Zare, H. Hosseini-Nejad, A. Khosravi, A. F. Atiya, D. Aminshahidi, S. Hussain, M. Rouhani, S. Nahavandi, and U. R. Acharya, "Epileptic seizures detection using deep learning techniques: A review," *Int. J. Environ. Res. Public Health*, vol. 18, no. 11, p. 5780, May 2021.

[50] P. Nejedly, V. Kremen, V. Sladky, J. Cimbalnik, P. Klimes, F. Plesinger, I. Viscor, M. Pail, J. Halamek, B. H. Brinkmann, M. Brazdil, P. Jurak, and G. Worrell, "Exploiting graphoelements and convolutional neural networks with long short term memory for classification of the human electroencephalogram," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.

[51] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing System*. Cambridge, MA, USA: MIT Press, 2006, pp. 1257–1264.

[52] M. Titsias, "Variational learning of inducing variables in sparse Gaussian processes," in *Proc. AISTATS*, vol. 5, 2009, pp. 567–574.

[53] A. Matthews, J. Hensman, R. Turner, and Z. Ghahramani, "On sparse variational methods and the Kullback–Leibler divergence between stochastic processes," in *Proc. AISTATS*, vol. 51, 2016, pp. 231–239.

[54] F. Tobar, "Band-limited Gaussian processes: The sinc kernel," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2019, pp. 12749–12759.

[55] F. Tobar, T. Bui, and R. Turner, "Learning stationary time series using Gaussian processes with nonparametric kernels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3483–3491.

[56] G. Rios and F. Tobar, "Compositionally-warped Gaussian processes," *Neural Netw.*, vol. 118, pp. 235–246, Oct. 2019.

[57] E. Snelson, Z. Ghahramani, and C. E. Rasmussen, "Warped Gaussian processes," in *Proc. Adv. neural Inf. Process. Syst.*, 2004, pp. 337–344.

[58] L. Wong and W. Abdulla, "Time-frequency evaluation of segmentation methods for neonatal EEG signals," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2006, pp. 1303–1306.

[59] P. von Bünau, F. C. Meinecke, S. Scholler, and K.-R. Müller, "Finding stationary brain sources in EEG data," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 2810–2813.

[60] M. Altamirano and F. Tobar, "Nonstationary multi-output Gaussian processes via harmonizable spectral mixtures," in *Int. Conf. Artif. Intell. Statist.*, 2022, pp. 1–14.