Driver Drowsiness Detection

An Approach Based on Intelligent Brain–Computer Interfaces

by Tharun Kumar Reddy and Laxmidhar Behera

Estimating reaction times (RTs) and drowsiness states from brain signals is a notable step in creating passive brain–computer interfaces (BCIs). Prior to the deep learning era, estimating RTs and drowsiness from electroencephalogram (EEG) signals was feasible only with moderate accuracy, which led to unreliability for neuro-engineering applications. However, recent developments in machine learning algorithms, notably stationarity-based approaches and deep convolutional neural networks (CNNs), have demonstrated promising results for a class of BCI systems, e.g., motor imagery BCIs, and affective state classification. These methods have not been systematically analyzed for EEG-based driver drowsiness detection and RT prediction.

This article studies the approaches, proposes new variants, and compares them with classical baselines to predict RTs from EEG data. We assess performance within subject-specific and subject-independent calibration settings, helping to reduce the need for session and subject calibration in BCI systems. Our results show that a stationarity incorporating the information theoretic joint approximate diagonalization method with fuzzy divergence (F-DivIT-JAD), when combined with a least absolute shrinkage and selection operator (LASSO) regressor, showed superior performance, recording the lowest root-mean-square error (RMSE), followed by other stationarity methods and CNNs, such as EEGNet regression networks. Our results motivate researchers to improve EEG driver drowsiness detection via deep learning and stationarity-based methods. In addition, guidelines are provided for using specific machine learning approaches.

©SHUTTERSTOCK.COM/LIGHTSPRING

Overview

BCIs facilitate communication between a brain and a device that enables neurological signals to direct an

Digital Object Identifier 10.1109/MSMC.2021.3069145
Date of current version: 14 January 2022

2333-942X/22©2022IEEE

external act [1], [2]. The earliest research into BCIs was conducted during the 1970s. In [1], all the elements to build a BCI are outlined. In the past few decades, human BCIs have generated ample interest. With the goal of appending cognitive monitoring to BCI systems, for instance, in driver assistance applications, recent approaches have employed BCI systems in driving simulators [3]–[6] to assess operator performance and inattention.

Driver Drowsiness Detection

Cognitive fatigue is a neurological state arising from extended exhaustive mental work [7]. Fatigue marks the arrival of, and generally coexists with, drowsiness, an intermediate state between waking and sleep [8]. Recent research indicates that drowsiness can be correctly diagnosed by effective decoding brain dynamics [9]. Several studies concluded that there are significant differences in the EEG power spectrum across fatigue and alert states [10]. Thus, a large number of EEG-based drowsiness tracking and detection systems have been introduced for real-world driving.

Drowsiness estimation and EEG-based RT prediction [3], [4] are regression problems. After capturing a signal, the regression problem involves multiple blocks, which are outlined in the following:

1) *Signal processing to enhance the signal-to-noise ratio and frequency realm filters.* These include bandpass filters and notch filters [11], [12] and spatial filters analogous to the common spatial pattern (CSP), fuzzy common spatial patterns regression one versus reset (CSPROVR), the fuzzy time delay common spectrospatial pattern (FTDCSSP) [5], and stationarity optimizing methods, such as fuzzy divergence common spatial patterns one versus reset (F-DivCSP-WS) [13].
2) *Feature processing to identify semantic predictors.* For example, Riemannian geometry (RG) [14] and EEG power band features [15], [16].
3) *Regression routines to project an analog output level.* For example, ordinary linear regression [11], [12], ridge regression [16], [17], the LASSO [18], transfer learning [16], dual task learning [4], and optimal learning [15] with deep CNNs.

In this work, we analyze the performance of the fuzzy CSPROVR, FTDCSSP, and F-DivCSP-WS and their novel variants. We benchmark the results of the proposed approach to RT prediction from EEG signals measured in a lane-keeping, sustained-attention psychomotor vigilance task [3], which collected 32-channel EEGs from 27 subjects while they drove on a four-lane road. CNNs that have shown promising results for multiple BCI systems are discussed. The state of the art (including the baseline and proposed methods) is detailed in Figure 1.

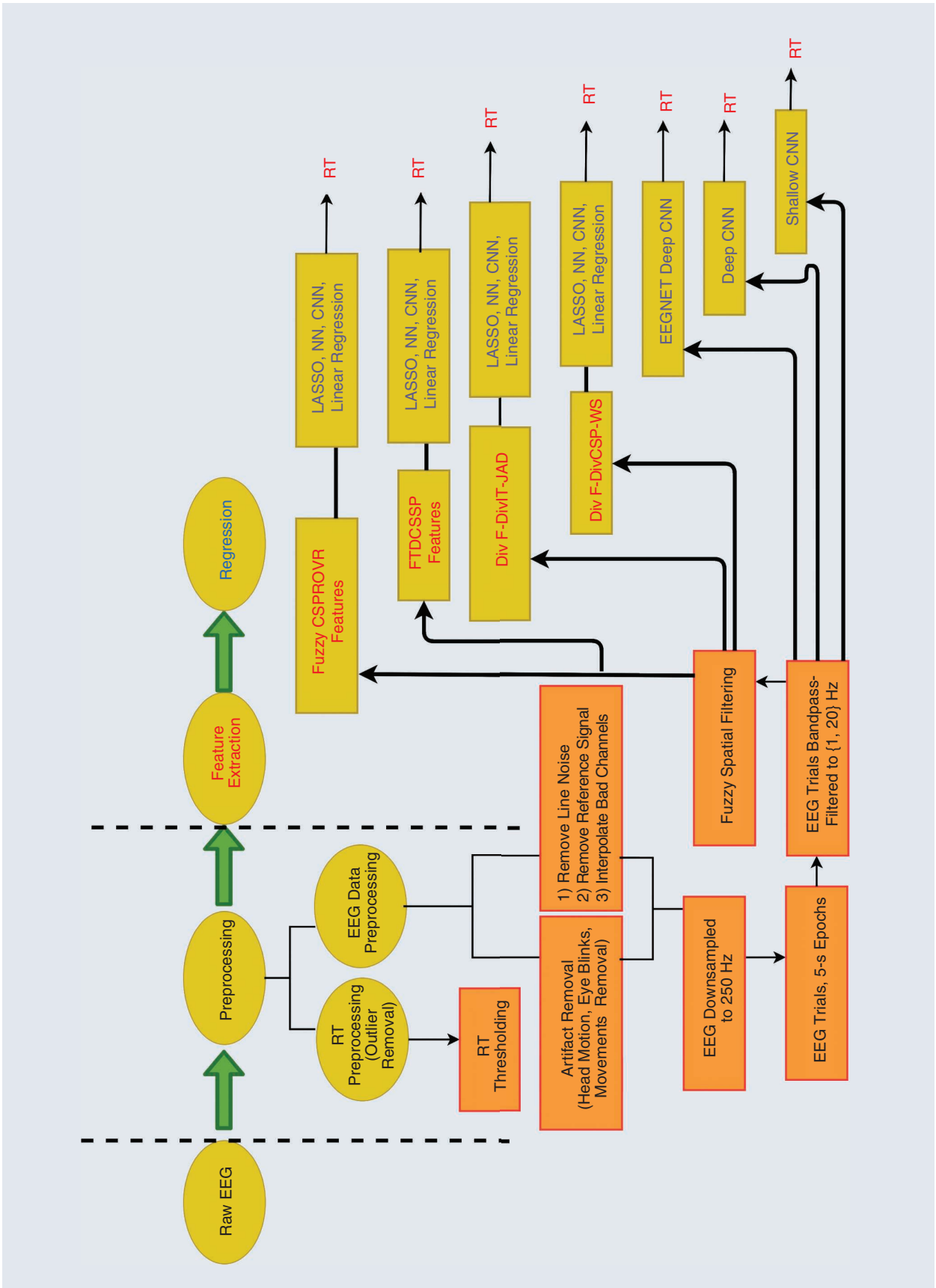

Figure 1. The state-of-the-art research.

CNNs

Deep learning has widely nullified the necessity for expertise in feature extraction, achieving advanced performance in computer vision and speech recognition. In particular, the use of deep CNNs has ramped up, in part, because of their success in many challenging computer vision classification problems [19], [20], overpowering approaches that depend on handcrafted, domain-specific features (see [21] for recent reviews). Despite these developments, a majority of BCI systems depend on bespoke features. In addition, several recent works [22] have focused on applying deep learning to EEG signals. For example, a shallow CNN [23] was implemented to classify cognitive and affective states. In [24], CNNs were developed to monitor epileptic subjects. These models are already employed for music information retrieval [25], detecting event-related potentials [26], motor imagery class prediction [27], sleep staging [28], anomaly discovery [29], and cognitive load prediction [30].

Estimating reaction times and drowsiness states from brain signals is a notable step in creating passive brain–computer interfaces.

In this article, we derive compact CNN models for regression: a shallow CNN–regression (CNN–R), a deep CNN–R, and an EEGNet–R. Contemporary research includes CNNs suited to EEG motor task classification: shallow and deep CNNs [27] and EEGNets [31]. A shallow CNN–R consists of a three-layer CNN with parameter settings optimized under robust empirical validation [27]. The primary layer is convolutional along the temporal dimension, while the subsequent layer is a convolutional across the EEG channels, i.e., along the spatial coordinates. The temporal convolution focuses on obtaining optimal bandpass kernels, while the spatial convolution aims at optimized spatial kernels. Further, the output signals are squared, followed by mean pooling to compute signals' band power, and the CNN ends with a fully connected linear regression layer. In summary, this network processes EEG data in a manner similar to the filter bank common spatial pattern (FBCSP) [32] (log-normalized power), followed by linear regression. Partitioning convolution into spatial and temporal operations [33] may improve accuracy for a large number of EEG input channels. This was the motivation for shallow and deep CNNs. In contrast to the FBCSP, these filters are simultaneously optimized, enabling them to outperform the FBCSP on motor EEG signals. All the preceding CNNs use minimally preprocessed EEG signals as input and filter them at 4–40 Hz.

Depth-wise convolution is a nice way of extracting optimal frequency-specific spatial filters. Further, we introduce depth-wise separable convolutions (depth-wise convolutions followed by pointwise convolutions). When applied to particular EEG signal processing, this procedure distributes learning into jointly pruned distinctive feature maps in time (the depth-wise convolution) and efficiently fuses the feature maps (the pointwise convolution). These advances have been realized in the form of an EEGNet–R CNN [31]. Several BCI baseline methods and learning models have been designed and fine-tuned for classification problems, but it is challenging to adapt them to regression scenarios. For the ease of notation, in this article, the deep CNN–R, EEGNet–R, and shallow CNN–R are referred to as the *deep CNN*, *EEGNet*, and *shallow CNN*, respectively.

Methods

Multichannel Persistent Attention Data Set

We examine driver demeanor and brain dynamics obtained from a 90-min persistent attention task (PAT) performed in a driving simulator. The information was gathered during 62 sessions in which 32-channel EEG electrode data were collected for 27 subjects who were driving on a four-lane road and directed to navigate in the center of a lane. Incidents were randomly triggered to make the car veer toward the left or right. A plenary trial consisted of three events: a deviation onset, response onset, and response offset. A subsequent trial, in which subjects were instructed to maneuver back to the primary cruising lane, started 5–10 s after the previous one. The temporal difference between the deviation onset and the response onset was a driver's RT, which is an objective indicator of drowsiness. In this article, a 5-s EEG trial prior to the deviation onset is used to assess a subject's drowsiness level via his or her RT.

Machine Learning Methods

The algorithms we evaluate have been applied to EEG-based motor imagery classification, cognitive load prediction, and so on, which are BCI problems, and the methods have produced remarkable results. It is quite important to note, in line with the opinions of the authors of [23], that motor imagery and drowsiness detection both involve changes in oscillatory activity (variations in the EEG signal power across various frequency bands); therefore, the approaches applicable to the former can be adapted to the latter. Hence, we devised the evaluation procedure discussed in this article. We also explored novel alternatives to these approaches. Altogether, we studied and proposed algorithms for regression: the fuzzy CSPROVR and FTDCSSP and stationarity-based approaches, such as the F-DivCSP-WS, F-DivIT-JAD, and deep CNNs.

CSP

The CSP is a supervised learning method for increasing the binary classification results in oscillatory EEG BCIs. This algorithm finds optimal spatial filters (which are nothing but a superposition of actual EEG channels) while ensuring that the variance of a filtered signal is amplified for one binary class and decayed for the other. We describe the OVR-CSP proposed to adapt a conventional CSP from a binary classification to M ($M > 2$) classes. For each class m, the OVR-CSP finds a matrix $\mathbf{W}_m^* \in \mathbb{R}^{C\times L}$, where L is the number of spatial filters to maximize the variance of class m against the rest:

$$\mathbf{W}_m^* = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{\operatorname{Tr}(\mathbf{W}^\top \tilde{\Sigma}_m \mathbf{W})}{\operatorname{Tr}(\mathbf{W}^\top \sum_{j\neq m} \tilde{\Sigma}_j \mathbf{W})}. \quad (1)$$

Here, $\tilde{\Sigma}_m$ is the mean covariance matrix of trials in class m, and $\mathbf{W}_m^*$ is the concatenation of the L eigenvectors associated with the L largest eigenvalues of the matrix $(\sum_{i\neq m} \tilde{\Sigma}_i)^{-1}\tilde{\Sigma}_m$. We concatenate the obtained L filters for each of the M classes to obtain $\mathbf{W}^* = [\mathbf{W}_1^*, ...\mathbf{W}_\mathbf{M}^*] \in \mathbb{R}^{C\times ML}$. Then, one can extract a spatially transformed trial by $\mathbf{X}_n' = \mathbf{W}^{*\top}\mathbf{X}_n, \quad n = 1,..,N$. Further, the feature vector is given by $\mathbf{F} = \begin{bmatrix} F_1 \\ .. \\ F_{\mathrm{ML}} \end{bmatrix}$, where F_i is given by $\log_{10} \frac{\| \mathbf{X}_i' \|^2}{\sum_{j=1}^{ML} \| \mathbf{X}_j' \|^2}$.

The obtained features are passed through LASSO and NN modules for the RT regression. Fuzzy CSP also requires the EEG signals to be bandpass-filtered in the range of {1, 20} Hz prior to spatial filtering.

Fuzzy CSPROVR

Let $\mathbf{X}^r \in \mathbb{R}^{C\times T}$, $r \in \{0, 1, .., N\}$ indicate the rth EEG trial pattern, and C and T describe the number of channels and temporal points, respectively. Primarily, the set $[0, 100]$ is broken into $K + 1$ adjoint blocks, with the separating points denoted by p_r, where

$$p_r = \frac{100r}{K+1}, r \in \{1, 2, ...K\}. \quad (2)$$

For every p_r, we associate its p_r th percentile point Y_{p_r} in the training set y_n (see Figure 2). Out of all such points, K sections are qualified as fuzzy sets. Now, it is possible to allocate training set y_n into K fuzzy classes. Every y_n pertains to a fuzzy class through a respective membership value $\in [0, 1]$. In Figure 2, $K = M$.

In the literature, OVR, one-versus-one (OVO), and JAD approaches calculate common spatial filters for multiclass problems. OVO computes CSPs for every two class combinations, and OVR calculates CSPs for every class against the rest of the classes considered jointly. In addition, one can derive a mean covariance matrix for each fuzzy class m' as

$$\widetilde{\Sigma}_m = \frac{\sum_{n=1}^{N} \mu_m(y_n)\mathbf{X}_n\mathbf{X}_n^\top}{\sum_{n=1}^{N} \mu_m(y_n)}, \quad m = 1, ..., K, \quad (3)$$

where $\mu_m(y_n)$ is the membership degree of y_n in fuzzy class m.

We propose the OVR-CSP to generalize the CSP from the binary classification to K classes. In particular, on a class m, the OVR-CSP calculates a transform $\mathbf{W}_m^* \in \mathbb{R}^{C\times F}$; here, F denotes the number of spatial filters:

$$\mathbf{W}_m^* = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{\operatorname{Tr}(\mathbf{W}^\top \widetilde{\Sigma}_m \mathbf{W})}{\operatorname{Tr}(\mathbf{W}^\top \sum_{j\neq m} \widetilde{\Sigma}_j \mathbf{W})}, \quad (4)$$

where $\mathbf{W}_m^*$ is the column-wise collection of the F eigenvectors corresponding to the F biggest eigenvalues of the matrix $(\sum_{i\neq m} \widetilde{\Sigma}_i)^{-1}\widetilde{\Sigma}_m$. We organize the obtained F filters column-wise for all K classes to obtain $\mathbf{W}^* = [\mathbf{W}_1^*, ...\mathbf{W}_\mathbf{K}^*] \in \mathbb{R}^{C\times KF}$. Later, one can calculate a spatially transformed trial by $\mathbf{X}_n' = \mathbf{W}^{*\top}\mathbf{X}_n, \; n = 1, .., N$.

JAD

The JAD algorithm is a popular alternative to the OVR-CSP for the classification of multiclass motor imagery. Given EEG data of K different classes, JAD finds a linear transformation $\mathbf{W} \in \mathbb{R}^{K\times K}$ that diagonalizes the class covariance matrices $\Sigma_i \in \mathbb{R}^{K\times K}$:

$$\mathbf{W}^\top \Sigma_i \mathbf{W} = \mathbf{D}_i, \quad i \in \{1, \cdots, \cdots, N\}, \quad (5)$$

where $\mathbf{D}_i \in \mathbb{R}^{K\times K}$ denotes diagonal matrices. The JAD formulation is motivated from the binary class, where one jointly diagonalizes two covariance matrices. Equation (5)

Prior to the deep learning era, estimating RTs and drowsiness from electroencephalogram signals was feasible only with moderate accuracy.

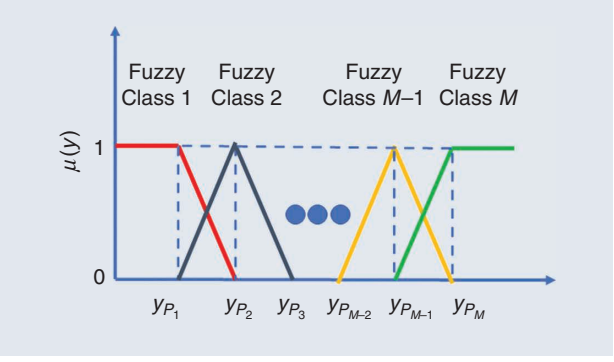

Figure 2. The M fuzzy classes for training RT values through triangular fuzzy membership.

can be solved using [34] and [35]. In the literature, the information theoretic filter extraction algorithm is the most commonly used technique for selecting filters obtained in matrix $\mathbf{W}$ [36].

FTDCSSP

Fuzzy time-delayed filters are used, generating the extended state space model

$$\mathbf{Z_k} \approx \sum_{\tau=0}^{2} \mu_{(\tau)} \mathbf{W}_{(\tau)} * (\delta_\tau \mathbf{X_k}), \tag{6}$$

where

$$\delta_{(\tau)}(\mathbf{X}_k) = \mathbf{X}_{(k-\tau)} \tag{7}$$

is the delay operator across the signal state space, μ_τ is the fuzzy membership value for the variable τ, and $\mathbf{W}_{(\tau)}$ is the optimized fuzzy CSSP weights matrix. Further, the terms in (6) can be simplified to obtain

$$\mathbf{Z}_k = \left[\mathbf{W}_{(0)}\mathbf{W}_{(1)}\mathbf{W}_{(2)}\right]\begin{bmatrix}\mu_0\mathbf{X}_{(k)}\\ \mu_1\mathbf{X}_{(k-1)}\\ \mu_2\mathbf{X}_{(k-2)}\end{bmatrix}. \tag{8}$$

The fuzzy CSSP filters, which are the rows in the matrices $\mathbf{W}_{(0)}, \mathbf{W}_{(1)}$ and $\mathbf{W}_{(2)}$, maximize the fuzzy mutual information criterion [5, eq. (22)], [37]. Each of the matrices $\mathbf{W}_{(0)}, \mathbf{W}_{(1)}$, and $\mathbf{W}_{(2)}$ apply to $\mu_0\mathbf{X}_{(k)}$, $\mu_1\mathbf{X}_{(k-1)}$, and $\mu_2\mathbf{X}_{(k-2)}$, respectively. In the CSP method [38], we select at least two filters (pertaining to the largest and smallest variances) for every class. In this manner, $F = 2K = 6$ is chosen in the experiments for $K = 3$. In (8), estimating three spatial transforms consists of calculating $3 \times 2K = 18$ row vectors.

Stationarity-Based Approaches

Nonstationarities are very frequent and can arise at different time instances. They are mainly caused by eye blinking, head/body movements, and drowsiness during the course of a trial. Between sessions, they can be triggered by different calibration settings and by constantly changing the positions of electrodes. In addition, subjects have physiological differences, leading to various signal probability distributions. These result in time-varying feature vectors. In fact, several traditional methods, including the fuzzy CSP algorithm, produce poor results from this feature space. We discuss approaches to deal with the problem of nonstationarity in EEG regression machine learning problems. One of them is the divergence-based technique. We generalize the notion of divergence-based CSP for regression through the concept of fuzzy sets.

F-DivCSP-WS

We propose a cost function for regression by deploying fuzzy covariance matrices. We begin by using two fuzzy classes and later generalize for multiple classes ($K > 2$). The conditional probability of every fuzzy class is normal, i.e., $\mathcal{N}(0, \overline{\Sigma}_1)$ and $\mathcal{N}(0, \overline{\Sigma}_2)$ for two fuzzy classes, respectively ($\overline{\Sigma}_1$ and $\overline{\Sigma}_2$ denote the fuzzy class covariances). The Kullback–Leibler (KL) divergence across two D variate gaussians $p_1 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $p_2 \sim \mathcal{N}(\mu_1, \Sigma_1)$ is calculated as

$$\begin{aligned} D_{kl}(p_1 \,\|\, p_2)) = \frac{1}{2}\Big(\log\Big(\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\Big) + trace(\Sigma_1^{-1}\Sigma_0), \\ + (\mu_1 - \mu_0)^\top \Sigma_1^{-1}(\mu_1 - \mu_0) - D). \end{aligned} \tag{9}$$

EEG trial $\mathbf{X}$ is to be spatially filtered, and one calculates the spatial filters by some approach; for instance, fuzzy CSP with regression gives $\mathbf{W}$, where $\mathbf{Y} = \mathbf{W}^\top\mathbf{X}$. Thus, the conditional probability of spatially transformed EEG trials is approximated by $p_1 = \mathcal{N}(0, W^\top \overline{\Sigma}_1 \mathbf{W})$ and $p_2 = \mathcal{N}(0, \mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W})$. One can compute the symmetric KL divergence between two distributions p_1 and p_2 as

$$\begin{aligned} F(\mathbf{W}) &= sD_{kl}(p_1 \,\|\, p_2), \\ &= D_{kl}(p_1 \,\|\, p_2) + D_{kl}(p_2 \,\|\, p_1), \\ &= \frac{1}{2}[\log\frac{\det(\mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W})}{\det(\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W})}, \\ &\quad + Tr((\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W})^{-1}(\mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W}))], \\ &\quad + \frac{1}{2}\Big[\log\frac{\det(\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W})}{\det(\mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W})}, \\ &\quad + Tr((\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W})^{-1}(\mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W}))\Big], \\ &\quad - d, \\ &= \frac{1}{2}[Tr((\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W})^{-1}(\mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W})), \\ &\quad + Tr((\mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W})^{-1}(\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W})) - 2d]. \end{aligned} \tag{10}$$

Here, $F(\mathbf{W})$ denotes the regression cost function and represents the regression approach's forecasting strength. The symmetric KL divergence [39] $sD_{kl}(p_1 \,\|\, p_2)$ linking the conditional probabilities of two fuzzy classes, after spatial filtering, can be written as

$$sD_{\text{kl}}(p_1 \,\|\, p_2) = \frac{1}{2}\Big[\frac{\mathbf{w}^\top \overline{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \overline{\Sigma}_2 \mathbf{w}} + \frac{\mathbf{w}^\top \overline{\Sigma}_2 \mathbf{w}}{\mathbf{w}^\top \overline{\Sigma}_1 \mathbf{w}} - 2\Big], \tag{11}$$

which is synonymous with the CSP objective function given by (13):

$$\mathbf{W}_{\text{skl}} = \underset{\mathbf{w}}{\text{argmin}}\ sD_{\text{kl}}(\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W} \,\|\, \mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W}), \tag{12}$$

$$\mathbf{W}^* = \underset{\mathbf{w}}{\text{argmax}} \frac{\mathbf{W}^\top \overline{\Sigma}_1 \mathbf{W}}{\mathbf{W}^\top \overline{\Sigma}_2 \mathbf{W}}. \tag{13}$$

In this article, we address an EEG-based driving scenario in which we examine the stationarity within sessions for every subject. The regularization function $G(\mathbf{W})$ is designed to optimize stationarity covering every fuzzy class:

$$G(\mathbf{W}) = \frac{1}{N_1 + N_2}\sum_{i=1}^{2}\sum_{j=1}^{N_i} \mu_{j,i} D_{\text{kl}}(\mathbf{W}^\top \Sigma_{\mathbf{j,i}} \mathbf{W} \,\|\, \mathbf{W}^\top \Sigma_{\mathbf{i}} \mathbf{W}), \tag{14}$$

where N_i captures the number of trials in the ith fuzzy class. In (14), $\Sigma_{\mathbf{j,i}}$ and $\Sigma_{\mathbf{i}}$ indicate the trial and class fuzzy covariances. We therefore put together a mixed objective function that conjointly optimizes coupled prediction and stationarity objectives:

$$\delta(\mathbf{W}) = \alpha F(\mathbf{W}) - (1-\alpha) G(\mathbf{W}), \tag{15}$$

where α is regularization multiplier. The optimization strategy to be maintained is the subspace approach utilizing gradient descent on an orthogonal manifold [39, p. 5, Algorithm 1]. In (15), we negate the regularization function as we maximize the stationarity [which minimizes divergence $G(\mathbf{W})$ and maximizes the forecasting ability of the model for every fuzzy class, boosting the divergence $F(\mathbf{W})$].

The framework so formulated is also generalized for multiple fuzzy classes by employing an OVR approach. The objective function is

$$F^{\text{OVR}}{}_j(\mathbf{W}) = sD_{\text{kl}}(\mathbf{W}^\top \overline{\Sigma}_j \mathbf{W} \parallel \mathbf{W}^\top \overline{\Sigma}_{\text{OVR}_j} \mathbf{W}), \tag{16}$$

$$\Sigma_{\text{OVR}_j} = \frac{1}{K} \Sigma_{K=1}^{K} \Sigma_K,$$

where $K > 2$ is total number of fuzzy classes; and

$$G_{\text{multiclass}}(\mathbf{W}) = \frac{1}{\sum_{k=1}^{K} N_k} \sum_{i=1}^{K} \sum_{j=1}^{N_i} \mu_{j,i} D_{\text{kl}}(\mathbf{W}^\top \Sigma_{\mathbf{j,i}} \mathbf{W} \parallel \mathbf{W}^\top \overline{\Sigma}_i \mathbf{W}), \tag{17}$$

where N_k is the number of trials in the kth fuzzy class and $\overline{\Sigma}_i$ and $\Sigma_{\mathbf{j,i}}$ denote the fuzzy class covariance and trial covariances, respectively. Membership value $\mu_{j,i}$ is interpreted for the jth trial pertaining to the ith. After including a regularization objective in the OVR framework, we arrive at

$$\mathbf{W}^*{}_i = \underset{\mathbf{W}}{\operatorname{argmin}} (\alpha F^{\text{OVR}}{}_i(\mathbf{W})) - (1-\alpha) G_{\text{multiclass}}(\mathbf{W}). \tag{18}$$

We estimate the spatial filter column for every OVR model by optimizing (18); $\alpha \in (0,1)$ is the regularization parameter. Filter matrix $\mathbf{W}$ is optimized using a subspace technique, where a group of filters is collectively optimized. Further, the filter ($\mathbf{W}$) is broken as a multiplication of whitening matrix ($\mathbf{T}$) and an orthogonal transform ($\mathbf{R}$); d' indicates the stationary subspace rank that is to be selected through cross validation/leave-one-out validation: $\mathbf{W}^\top = \widetilde{\mathbf{R}}T$, $\tilde{\mathbf{R}} = \mathbf{I}_d \mathbf{R}$, $\mathbf{W} \in \mathbb{R}^{D \times d}$, $\mathbf{T} \in \mathbb{R}^{D \times D}$, and $\mathbf{T}^\top(\Sigma_1 + \Sigma_2)\mathbf{T} = \mathbf{I}$. The optimal filter is calculated on an orthogonal subspace/manifold; i.e., $\mathbf{R}\mathbf{R}^T = \mathbf{I}$. The cost functions now rely on orthogonal matrix $\mathbf{R}$

$$\Delta(\mathbf{R}) = \underbrace{(\alpha)\mathbf{F}^{OVR}{}_j(\mathbf{I}_d \mathbf{R})}_{\text{Fuzzy OVR CSP}} - \underbrace{(1-\alpha) G_{\text{multiclass}}(\mathbf{I}_d \mathbf{R})}_{\text{Stationary}},$$

where $d < D$: for a subspace approach with $\alpha \in (0,1)$.

F-DivIT-JAD

Equation (5) presents the basic JAD formulation. It finds minima of the KL divergence between covariances of the transformed trials and a diagonal version of the divergence. For instance, if $\mathbf{Y}$ is a matrix and Σ is a diagonal matrix, by a Pythagorean decomposition, one obtains

$$D_{\text{kl}}(\mathbf{Y} \parallel \Sigma) = D_{\text{kl}}(\mathbf{Y} \parallel diag(\mathbf{Y})) + D_{\text{kl}}(diag(\mathbf{Y}) \parallel \Sigma), \tag{19}$$

where $diag(\mathbf{X})$ is a matrix array whose elements in a diagonal are same as the diagonal elements of $\mathbf{Y}$. Minimizing $D_{\text{kl}}(\mathbf{Y} \parallel \Sigma)$, the expression of Σ equals $diag(Y)$. In other words, the formulation of the F-DivIT-JAD appears in (21):

$$F_J(\mathbf{W}) = \sum_{i=1}^{K} (p_k D_{\text{kl}}(\mathbf{W}^\top \Sigma_{\mathbf{i}} \mathbf{W} \parallel diag(\mathbf{W}^\top \Sigma_{\mathbf{i}} \mathbf{W})), \tag{20}$$

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} F_J(\mathbf{W}), \tag{21}$$

$$G_J(\mathbf{W}) = \frac{1}{\sum_{k=1}^{K} N_k} \sum_{i=1}^{K} \sum_{j=1}^{N_i} \mu_{j,i} D_{kl}(\mathbf{W}^\top \mathbf{\Sigma}_{\mathbf{j,i}} \mathbf{W} \parallel \mathbf{W}^\top \overline{\mathbf{\Sigma}}_i \mathbf{W}). \tag{22}$$

We propose another formulation of the stationarity-based method for regression deploying fuzzy covariances in (22). In it, we integrate within a session the stationarity developed in (14) with an information theoretic formulation of JAD as follows:

$$\Delta(\mathbf{W}) = \left[(1-\alpha) \left(\sum_{i=1}^{K} p_i D_{\text{kl}}(\mathbf{W}^\top \Sigma_{\mathbf{i}} \mathbf{W} \parallel diag(\mathbf{W}^\top \Sigma_{\mathbf{i}} \mathbf{W})) \right), + \alpha \left(\frac{1}{\sum_{k=1}^{K} N_k} \sum_{i=1}^{K} \sum_{j=1}^{N_i} \mu_{j,i} D_{\text{kl}}(\mathbf{W}^\top \Sigma_{\mathbf{j,i}} \mathbf{W} \parallel \mathbf{W}^\top \overline{\Sigma}_i \mathbf{W}) \right) \right].$$

In contrast to the DivCSP-WS method, as indicated by [39, eq. (10)], the regularizing expression is "added" because we are minimizing the diagonalization term as well as the group stationarity term. The spatial filters $\mathbf{W}^*$ are estimated such that

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \Delta(\mathbf{W}). \tag{23}$$

We optimize (23) on the subspace $\mathbf{W}$:

$$\mathbf{R}^* = \underset{\mathbf{R}}{\operatorname{argmin}} (1-\alpha) J(\mathbf{R}) + \alpha J_s(\mathbf{I}_d \mathbf{R}), \tag{24}$$

where J and J_s denote the diagonalization cost and the stationarity cost in terms of $\mathbf{R}$ (the orthogonal transform). In (24), we optimize the whole subspace for the JAD term. But for optimizing stationarity, we use $\mathbf{I}_d\mathbf{R}$ in place of $\mathbf{R}$. In other words, $\mathbf{I}_d\mathbf{R}$ points to the selection of the first d rows of the orthogonal transform R while choosing the first d columns of the filter transform $\mathbf{W}$ to incorporate stationarity. Using the proposed approach, we accomplish a pair of objectives: the JAD of the matrices and the imposition of stationarity on the primary d elements of the transform.

Spatial filters $\mathbf{W}$ are optimized using the following approaches:

- *Subspace approach*: a collection of jointly optimized filters
- *Deflation technique*: the sequential optimization of filters.

Filters ($\mathbf{W}$) are broken into a product of whitening matrix ($\mathbf{S}$) and orthogonal matrix ($\mathbf{R}$); d' represents the dimension of the stationary subspace tuned by cross validation:

$$\mathbf{W}^T = \tilde{\mathbf{R}}\mathbf{S} \quad \tilde{\mathbf{R}} = \mathbf{I}_d\mathbf{R} \quad \mathbf{W} \in \mathbb{R}^{D\times d},\ \mathbf{S} \in \mathbb{R}^{D\times D},$$
$$\mathbf{S}^T(\Sigma_1 + \Sigma_2)\mathbf{S} = \mathbf{I}.$$

Optimization is conducted on an orthogonal manifold; i.e., $\mathbf{R}\mathbf{R}^T = \mathbf{I}$. The objective functions now depend on orthogonal matrix $\mathbf{R}$:

$$\mathbf{R}^* = \underset{\mathbf{R}}{\operatorname{argmin}}\ (1-\alpha)J(\mathbf{R}) + \alpha J_s(\mathbf{I_d}\mathbf{R}),$$

where $d < D$: for the subspace method and $d = 1$: for the sequential optimization or deflation method.

CNNs for Regression

In a nutshell, a CNN is a multilayer feedforward NN designed to learn spatial dependencies by employing fundamental subsystems as convolution layers, pooling layers, and fully connected layers. Convolution and pooling layers extract features, while fully connected layers transform the computed features to an output, which is useful for classification and regression. Recently, several studies developed novel models for CNNs adapted to motor imagery classification and P300 classification in EEGs, including the shallow CNN, deep CNN [27], and EEGNet [31]. A deep CNN is composed of four convolution maximum pooling blocks, with a specified primary block outlined to work with an EEG input, accompanied by three typical convolution maximum pooling clusters and a dense softmax classification output. The idea of split convolutions was used in the first convolution cluster. It consisted of a temporal convolution followed by spatial filtering. This was followed by maximum pooling and subsequent convolutional layers. Linear units were used in the temporal convolution, and exponential linear units were employed in spatial convolution. A shallow CNN has three layers and tunable parameters. It was recently tested and validated for classification problems [27]. The preliminary layer performs convolution across the time direction, while the subsequent layer accomplishes convolution across the spatial dimension, i.e., across EEG channels.

Convolution across the time dimension focuses on optimizing bandpass filters, and spatial convolution seeks to optimize spatial filters. The obtained signal amplitude is squared and averaged in a pooling fashion to derive the band power. Then, the last one is a fully connected linear classification layer. Although the network processes the signal in a manner analogous to the FBCSP, there is a difference in terms of the convolutional network conjointly optimizing spatial and temporal filters. In summary, this CNN processes EEG data in a manner similar to the FBCSP and linear discriminant analysis. In contrast to the FBCSP, all these filters are simultaneously optimized, producing better performance using motor EEG signals. A shallow CNN uses minimally preprocessed EEG signals as input, so we filtered the signals at 4–40 Hz. In this article, the preceding CNN models are adapted and fine-tuned for regression. We implement three models: the shallow CNN, deep CNN, and EEGNet.

Implementation and Results

PAT and EEG Preprocessing

The PAT experiment is shown in Figure 3. The goal is to study the correlation between fatigue and driving performance, based on the proposition that poor vigilance leads to significant delays before drivers notice events. The experiment details, EEG trial preprocessing, and RT processing remain the same as reported in [5, Secs. 3(a), 3(b)(1), and 3(b)(2)]. The raw data are available for download from https://doi.org/10.6084/m9.figshare.6427334.v5 [3]. They are targeted to assess RTs by utilizing a 5-s EEG window shortly ahead of drowsiness/alert states.

Methods for Performance Comparison

In practice, three approaches—an EEGNet [31], a shallow CNN [23], and a deep CNN [27]—applied to EEG trials are generally used for predicting RTs. The fuzzy CSP [18] and FTDCSSP [5] feature extraction methods with LASSO-based RT prediction are additional methods used as baselines. We compare the performance of the proposed methods (F-DivCSP-WS- and F-DivIT-JAD-based feature extraction with LASSO-based RT prediction), with the preceding approaches used as baselines.

Hyperparameter Tuning and Performance Comparison

Figure 3 and 4 along with Figures 5 and 6 in the supplementary material depict the average performance of the methods on the RT data set. The analysis is done with a leave-one-session-out validation approach for each of the subjects. Subjects $S04, S06, S11, S23, S48, S52, S54$, and $S55$ are left out of the analysis, as their data were limited to a single session. The EEGNet, deep CNN, and shallow CNN architectures are trained via the Keras deep learning platform with a TensorFlow backend, and the server is an Intel Xeon CPU with a 2.2-GHz processor, a Tesla T4 15079MiB GPU, and 13 GB of random-access memory. Five hundred epochs were used to train the networks.

Hyperparameter Tuning in Each Method

Shallow CNN

The shallow CNN was the slowest, with an epoch taking 5–6 s. This is attributed to the custom activation functions that were implemented from scratch and not optimized for the model. To reduce the time complexity, we changed the filters in the first convolution layer to 20, from the original 40. The learning rate obtained from the line search was 0.0001, and the average model showed better results. The network had approximately 26,500 trainable parameters.

Deep CNN

This method had the largest number of trainable parameters—around 300,000—which made it the most complex model to train. Even with a dropout of 0.5, it was clearly overfitting. The trick to training it, again, was to tune the learning rate by using a line search ($\eta = 0.001$), and the model yielded better results in leave-one-session-out cross-validation. The computational time to train it was the longest of all the CNNs in this article.

EEGNet

The EEGNet demonstrated the best results of the three CNNs. We ran the standard EEGNet model, without any changes, for a filter size of 128 and a standard learning rate of 0.01. After changing the learning rate for the shallow and deep CNNS and seeing the results improve, we decreased the learning rate. The learning rate ($\eta = 0.001$) and filter size (64) were hyperparameters tuned for the EEGNet through a grid search using leave-one-session-out cross-validation. This was also the fastest method of the three described. It was the most optimized of the three networks, with slightly more than 2,000 trainable parameters.

Fuzzy CSP

We implemented the fuzzy CSP method with LASSO regression, treating the number of fuzzy classes and filters per fuzzy class as hyperparameters and performing tuning

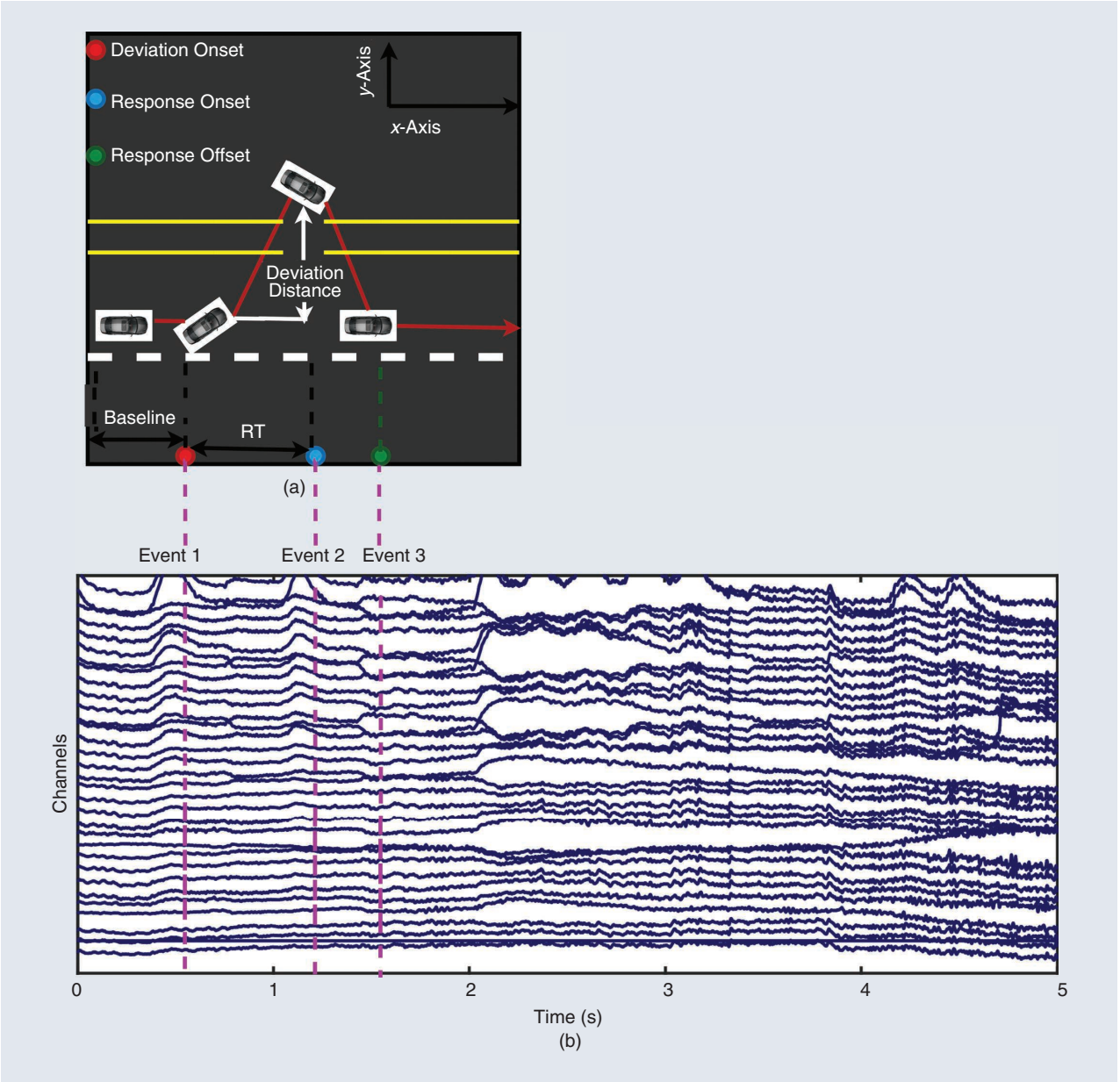

Figure 3. The EEG PAT driving scenario. (a) A pictorial description of the experimental protocol design and (b) the corresponding time information.

with a grid-search in a leave-one-session-out cross-validation paradigm. The average results for each subject were plotted: $K \in \{3,6,9,12\}$, and F was fixed at six.

FTDCSSP

The tunable hyperparameters consisted of the number of spatial filters F, which lay in the set $\{5,10,15,20\}$. We fixed the number of fuzzy classes at three (a number borrowed from the literature) and varied F.

F-DivCSP-WS

We used $K = 3; F = 3,$ where F was the number of filters per fuzzy class; and $d = 2$ (the number of OVR components), which was an estimate of the size of the subspace of filters. This was borrowed from the selection of the number of spatial filters in OVR-CSP [40, Sec. 4]. Also, $\alpha = 0.5$ was obtained through leave-one-out validation.

F-DivIT-JAD

We used $K = 3; F = 3,$ where F was the number of filters per fuzzy class; and $d = 2$ (the number of OVR components), which was an estimate of the size of subspace of filters. Also, $\alpha = 0.6$ was obtained through leave-one-session/subject-out cross-validation.

Evaluation Criterion

The RMSE constitutes the criteria used for measuring the regression results. Consider N training samples, with y_{di} denoting the actual RT for the ith example and y_i representing the predicted RT:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{di} - y_i)^2}{N}}. \tag{25}$$

Performance Comparison

Figure 3 and 4 along with Figures 5 and 6 in the supplementary material depict the average RMSE performance of all the methods applied to the RT data set when analyzed through a leave-one-session-out validation for all the sessions of a particular subject. The respective percentage performance improvements obtained by using the seven methods (proposed and baseline) are given in Figures 5 and 6 in the supplementary material. For instance, the terms *F-DivIT-JAD/F-DivCSP-WS* represents the improvement of the F-DivIT-JAD representations over the F-DivCSP-WS representations. *F-DivIT-JAD/EEGNET* denotes the improvement of the F-DivIT-JAD features over the EEGNet CNN-based RT prediction. On an average, using the F-DivIT-JAD, we recorded a 21.2% smaller RMSE than for the EEGNET, a 41.14% smaller RMSE than for the deep CNN, a 51.39% smaller RMSE than for the shallow CNN, a 21.93% smaller RMSE than for fuzzy CSPROVR, a 16.03% smaller RMSE than for FTDCSSP, and a 5.67% smaller RMSE than for F-DivCSP-WS. These reductions translate into a significantly smaller driving distance error at a constant speed of 100 km/h.

Figures 7–10, along with figures 11 and 12 in the supplementary material depict the average performance of all the methods applied to the RT data set when analyzed in a leave-one-subject-out validation for all subjects. On an average, using the F-DivIT-JAD, we recorded a 47.12% smaller RMSE than for the EEGNET, a 45.65% smaller RMSE than for the deep CNN, a 39.05% smaller RMSE than for the shallow CNN, a 27.24% smaller RMSE than for the fuzzy CSPROVR, a 17.61% smaller RMSE than for the FTDCSSP, and an 8.96% smaller RMSE than for the F-DivCSP-WS. These reductions translate to a significantly smaller driving distance error at a constant speed of 100 km/h.

Regression RMSE values are reported in Figure 4. We ran a two-way variance analysis to evaluate an algorithm's impact on the EEG RT data set, regarding the calibration setting (subject-specific/independent), while treating the subjects as random factors. The results are presented in Tables 1 and 2 along with Table 3 in the supplementary material (p *value* < 0.05); they indicate that there are statistically significant differences in the RMSE for different algorithms for a variety of calibration settings (subject-specific/independent). In other words, the selection of a regression method has a significant effect on the performance metric RMSE (p *value* < 0.05); see Tables 1 and 2 along with Table 3 in the supplementary material.

Furthermore, multiple comparisons in the form of paired t-tests are employed to find out if the difference between any couple of algorithms is statistically significant, with the p value corrected through the false discovery rate method [41]. The p values are in Table 1, where bold-face entries values are statistically significant, with large effects, and the other p values corresponded to medium effects. The size of the p value does not necessarily estimate the size of the effect; statistics, such as the partial eta squared, give effect sizes. Effect sizes for p value < 0.2 are reported as the

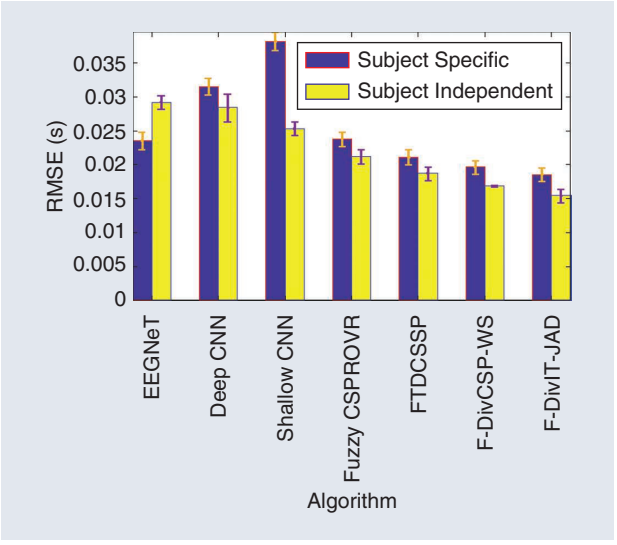

Figure 4. The RT prediction RMSE of each algorithm applied to the driving data set.

partial eta squared (η^2_{partial}), whose values can be benchmarked against Cohen's [43] criteria of small (0.01), medium (0.06), and large (0.14) effects, according to Richardson (2011) [42].

During this subject-independent (cross-subject) analysis, all the deep networks (the EEGNet, shallow CNN, and deep CNN) and methods (the fuzzy CSP and FTDCSSP) used hyperparameter settings that were similar to those mentioned in the preceding for the cross-session (subject-specific) analysis. In addition, 500 epochs were employed to train the deep networks. In summary, the F-DivIT-JAD performed better than the other algorithms for predicting the RT in subject-specific and subject-independent settings. Table 2 contains the average RMSE obtained from all the methods in subject-specific and subject-independent settings.

Deep learning has widely nullified the necessity for expertise in feature extraction, achieving advanced performance in computer vision and speech recognition.

Discussion and Conclusion

In this article, several intelligent machine learning systems were provided for driver drowsiness detection through EEG signals. The algorithms were tested under subject-specific and subject-independent calibration settings. In summary, we studied the EEGNet, shallow CNN, deep CNN, fuzzy CSPROVR with a LASSO, FTDCSSP with a LASSO, and two new methods (the F-DivCSP-WS with a LASSO and the F-DivIT-JAD with a LASSO). We observe that the EEGNet CNN performed decently in subject-specific and subject-independent settings. The deep and shallow CNNs demonstrated overfitting in the subject-independent setting. This is attributed to the depth-wise and separable convolutions used in the EEGNet architecture. The proposed methods (the F-DivCSP-WS with a LASSO and the F-DivIT-JAD with a LASSO) had a lower average RMSE than the baseline methods (the FTDCSSP, shallow CNN, deep CNN, and fuzzy CSPROVR). The fuzzy CSPROVR and FTDCSSP performed close to and even better than the EEGNet in subject-specific and subject-independent settings. This is attributed to the diverse nature of the filters learned in both methods by using optimal variance criterion (the fuzzy CSPROVR and FTDCSSP) and the mean-square-error loss (the EEGNet, deep CNN, and shallow CNN).

In general, the regression performances of the deep CNN and EEGNet were analogous across all cross-subject analyses, whereas the deep CNN performed worse for subject-specific analyses. One explanation for this is the multitude of data employed to train the model; in subject-independent analyses, the training set sizes were 15–20 times larger than those for subject-specific analyses. This helps us infer that the deep CNN is more data-hungry in comparison to the EEGNet, an expected result, provided that the architecture of the deep CNN is two times larger than the EEGNet. We presume that the argument is consistent with the findings originally published by the developers of the deep CNN [27], who mentioned that a training data augmentation technique was mandatory to record good classification performance on sensorimotor rhythm data. In contrast to that work, the EEGNet and the other proposed and baseline models performed well across all test settings, beyond the need for

Table 1. Paired t-test results (*p* values) for RMSE comparison across methods.

	Cross-Session (Leave-One-Session-Out Validation)					
	F-DivCSP-WS	**FTDCSSP**	**Fuzzy CSPROVR**	**Shallow CNN**	**Deep CNN**	**EEGNet**
F-DivIT-JAD	**0.002**	0.01	**0.001**	0.002	0.01	**0.001**
	Cross-Subject (Leave-One-Subject-Out Validation)					
	F-DivCSP-WS	**FTDCSSP**	**Fuzzy CSPROVR**	**Shallow CNN**	**Deep CNN**	**EEGNet**
F-DivIT-JAD	**0.001**	0.02	**0.002**	0.001	0.01	**0.0001**

Bold-face *p* values indicate large effects ($\eta^2_{\text{partial}} > 0.13$).

Table 2. A comparison of the average RMSE across methods for subject-independent and subject-specific (cross-session) validation.

Subject Specific (Leave-One-Session-Out Validation)						
EEGNet	**Deep CNN**	**Shallow CNN**	**Fuzzy CSPROVR**	**FTDCSSP**	**F-DivCSP-WS**	**F-DivIT-JAD**
0.024	0.032	0.038	0.024	0.022	0.02	0.019
Subject Independent (Leave-One-Session-Out Validation)						
EEGNet	**Deep CNN**	**Shallow CNN**	**Fuzzy CSPROVR**	**FTDCSSP**	**F-DivCSP-WS**	**F-DivIT-JAD**
0.029	0.028	0.025	0.021	0.019	0.017	0.015

data augmentation, rendering the models simpler, adaptable, and generic to use in practical settings.

For the proposed models (the F-Div-IT-JAD and F-DivCSP-WS), the hyperparameter space was sufficiently explored to choose optimal settings for reporting results. In the future, we will focus on the efficient selection of optimization parameters through cross-validation and the integration of subject-independent stationarity to estimate robust spatial filters. Such results enable us to suggest guidelines for which algorithm to use for RT predictions from EEGs. The F-DivIT-JAD and F-Div-CSP-WS are recommended for subject-specific and subject-independent calibration, whatever the amount of training data. The EEGNet CNN is recommended for subject-specific RT calibration (several hundreds of training trials), but its performance is slightly compromised, with an enormous number of trials (in which case the deep CNN can be used).

CNNs and the proposed stationarity-enforcing methods are machine learning assets for scientists and engineers whose aim is to decode drowsy states from EEG signals.

In summary, we extended the divergence framework of the CSP algorithm for multiple fuzzy class settings by using the OVR strategy. We also proposed a composite framework employing information theoretic JAD to optimize within session stationarity. The regression performance of the F-DivCSP-WS (the F-Div-CSPROVR reduces to fuzzy CSP in a binary fuzzy class setting) and F-DivIT-JAD follows a similar trend when the regularization parameter controlling stationarity is regulated, which further reinforces our formulated divergence-based JAD framework. Also, in multiple fuzzy class settings, the F-DivIT-JAD and F-DivCSP-WS techniques have similar performance for different values of α. One of the main advantages of the F-DivIT-JAD in comparison to the F-DivCSP-WS method is computational time. This is because, in the OVR framework, the gradient descent optimization is repeatedly performed for n iterations (where n is the number of classes in multiple fuzzy class settings). However, when the initialization of the orthogonal matrix for each OVR case is the corresponding fuzzy CSPROVR solution, the computation time in the F-DivCSP-WS drastically reduces and outperforms that of the F-Div-IT-JAD. In essence, our results demonstrate that CNNs and the proposed stationarity-enforcing methods are machine learning assets for scientists and engineers whose aim is to decode drowsy states from EEG signals using intelligent system models.

Limitations and Future Work

Real-world deployment of the proposed solutions in this article requires further validation through a resource utilization study consisting of a rigorous analysis of metrics such as the RMSE, memory trace, number parameters, operations enumeration, prediction time, and so on. This is a limitation and will be addressed in future work. In addition, for practical applications, high-density EEG systems can be replaced by ergonomic headsets, which employ few channels to perform several applications. Good examples include Neurosky MindWave, InteraXon Muse, Emotiv Epoc, Emotiv Insight, and OpenBCI. Also, including other intrusive drowsiness detection modalities with sensors, such as electrocardiogram, electromyogram, and electrodermal activity, can contribute to a significant improvement in the drowsiness detection system's performance. In addition, long short-term memory and recurrent neural network (RNN) models constitute a standard choice for the time series data, which need to be further explored for drowsiness detection. Approaches to the fusion of multiple sensor modalities and the incorporation of RNN models for regression can be explored in upcoming research.

Acknowledgments

We are grateful to Dr. Y.K. Wang for the valuable guidelines he provided for the driving data set used in this manuscript. This article has supplementary downloadable material available at https://doi.org/10.1109/MSMC.2021.3069145, provided by the authors.

About the Authors

Tharun Kumar Reddy (tharun@iitk.ac.in) is with the Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, 247667, India.

Laxmidhar Behera (lbehera@iitk.ac.in) is with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, 208016, India. He is a Senior Member of IEEE.

References

[1] J. J. Vidal, "Toward direct brain–computer communication," *Annu. Rev. Biophys. Bioeng.*, vol. 2, no. 1, pp. 157–180, 1973. doi: 10.1146/annurev.bb.02.060173.001105.

[2] J. Wolpaw and E. W. Wolpaw, *Brain–Computer Interfaces: Principles and Practice*. London: Oxford Univ. Press, 2012.

[3] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, "Multi-channel EEG recordings during a sustained-attention driving task," *Sci. Data*, vol. 6, no. 1, pp. 1–8, 2019. doi: 10.1038/s41597-019-0027-4.

[4] T. K. Reddy, V. Arora, S. Kumar, L. Behera, Y.-K. Wang, and C.-T. Lin, "Electroencephalogram based reaction time prediction with differential phase synchrony representations using co-operative multi-task deep neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 3, no. 5, pp. 369–379, 2019. doi: 10.1109/TETCI.2018.2881229.

[5] T. K. Reddy, V. Arora, L. Behera, Y-k. Wang, and C.-T. Lin, "Multiclass fuzzy time-delay common spatio-spectral patterns with fuzzy information theoretic optimization

for EEG-based regression problems in brain–computer interface (BCI)," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 10, pp. 1943–1951, 2019. doi: 10.1109/TFUZZ.2019.2892921.

[6] S. Nahavandi, "Robot-based motion simulators using washout filtering: Dynamic, immersive land, air, and sea vehicle training, vehicle virtual prototyping, and testing," *IEEE Syst., Man, Cybern. Mag.*, vol. 2, no. 3, pp. 6–10, 2016. doi: 10.1109/MSMC.2016.2566119.

[7] M. A. Boksem and M. Tops, "Mental fatigue: Costs and benefits," *Brain Res. Rev.*, vol. 59, no. 1, pp. 125–139, 2008. doi: 10.1016/j.brainresrev.2008.07.001.

[8] R. O. Phillips, "A review of definitions of fatigue–and a step towards a whole definition," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 29, pp. 48–56, Feb. 2015. doi: 10.1016/j.trf.2015.01.003.

[9] C.-T. Lin et al., "Exploring the brain responses to driving fatigue through simultaneous EEG and FNIRS measurements," *Int. J. Neural Syst.*, vol. 30, no. 1, p. 1,950,018, 2020. doi: 10.1142/S0129065719500187.

[10] K.-C. Huang et al., "An EEG-based fatigue detection and mitigation system," *Int. J. Neural Syst.*, vol. 26, no. 4, p. 1,650,018, 2016. doi: 10.1142/S0129065716500180.

[11] T. J. Bradberry, R. J. Gentili, and J. L. Contreras-Vidal, "Reconstructing three-dimensional hand movements from noninvasive electroencephalographic signals," *J. Neurosci.*, vol. 30, no. 9, pp. 3432–3437, 2010. doi: 10.1523/JNEUROSCI.6107-09.2010.

[12] T. J. Bradberry, R. J. Gentili, and J. L. Contreras-Vidal, "Fast attainment of computer cursor control with noninvasively acquired brain signals," *J. Neural Eng.*, vol. 8, no. 3, p. 036,010, 2011.

[13] T. K. Reddy, V. Arora, L. Behera, Y.-k. Wang, and C.-T. Lin, "Fuzzy divergence based analysis for EEG drowsiness detection brain computer interfaces," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, 2020, pp. 1–7. doi: 10.1109/FUZZ48607.2020.9177833.

[14] D. Wu, B. J. Lance, V. J. Lawhern, S. Gordon, T.-P. Jung, and C.-T. Lin, "EEG-based user reaction time estimation using Riemannian geometry features," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 2157–2168, 2017. doi: 10.1109/TNSRE.2017.2699784.

[15] T. K. Reddy, V. Arora, and L. Behera, "HJB-equation-based optimal learning scheme for neural networks with applications in brain–computer interface," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 2, pp. 159–170, 2018. doi: 10.1109/TETCI.2018.2858761.

[16] D. Wu, C.-H. Chuang, and C.-T. Lin, "Online driver's drowsiness estimation using domain adaptation with model fusion," in *Proc. Int. Conf. Affective Comput. Intell. Interaction (ACII)*, 2015, pp. 904–910. doi: 10.1109/ACII.2015.7344682.

[17] J. Sandhan, A. Mitra, and V. Subramanian, "Object counting in a single surveillance image," in *Proc. 23rd Nat. Conf. Commun. (NCC)*, 2017, pp. 1–4. doi: 10.1109/NCC.2017.8077076.

[18] D. Wu, J.-T. King, C.-H. Chuang, C.-T. Lin, and T.-P. Jung, "Spatial filtering for EEG-based regression problems in brain–computer interface (BCI)," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 771–781, 2017. doi: 10.1109/TFUZZ.2017.2688423.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2017, pp. 4700–4708.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539.

[22] G. Dai, J. Zhou, J. Huang, and N. Wang, "HS-CNN: A CNN with hybrid convolution scale for EEG motor imagery classification," *J. Neural Eng.*, vol. 17, no. 1, p. 016,025, 2020. doi: 10.1088/1741-2552/ab405f.

[23] A. Appriou, A. Cichocki, and F. Lotte, "Modern machine-learning algorithms: For classifying cognitive and affective states from electroencephalography signals," *IEEE Syst., Man, Cybern. Mag.*, vol. 6, no. 3, pp. 29–38, 2020. doi: 10.1109/MSMC.2020.2968638.

[24] A. Antoniades, L. Spyrou, C. C. Took, and S. Sanei, "Deep learning for epileptic intracranial EEG data," in *Proc. IEEE 26th Int. Workshop Mach. Learning Signal Process. (MLSP)*, 2016, pp. 1–6. doi: 10.1109/MLSP.2016.7738824.

[25] S. Stober, D. J. Cameron, and J. A. Grahn, "Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1449–1457.

[26] H. Cecotti and A. Graser, "Convolutional neural networks for p300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, 2010. doi: 10.1109/TPAMI.2010.125.

[27] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017. doi: 10.1002/hbm.23730.

[28] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2018. doi: 10.1109/TBME.2018.2872652.

[29] W. Li, G. Wu, and Q. Du, "Transferred deep learning for anomaly detection in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 597–601, 2017. doi: 10.1109/LGRS.2017.2657818.

[30] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, arXiv:1511.06448.

[31] 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, p. 056013, 2018. doi: 10.1088/1741-2552/aace8c.

[32] 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.)*, 2008, pp. 2390–2397.

[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016. [Online]. Available: http://arxiv. org/abs/1610.02357

[34] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 17, no. 1, pp. 161–164, 1996. doi: 10.1137/S0895479893259546.

[35] F. Bouchard, B. Afsari, J. Malick, and M. Congedo, "Approximate joint diagonalization with Riemannian optimization on the general linear group," *SIAM J. Matrix Anal. Appl.*, vol. 41, no. 1, pp. 152–170, 2020. doi: 10.1137/18M1232838.

[36] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, 2008. doi: 10.1109/TBME.2008.921154.

[37] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 121–131, 2010. doi: 10.1109/TBME.2010.2077291.

[38] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, 2010. doi: 10.1109/TBME.2010.2082539.

[39] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Rev. Biomed. Eng.*, vol. 7, pp. 50–72, Nov. 2014. doi: 10.1109/RBME.2013.2290621.

[40] S. Kumar, T. K. Reddy, V. Arora, and L. Behera, "Formulating divergence framework for multiclass motor imagery eeg brain computer interface," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2020)*, 2020, pp. 1344–1348.

[41] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. B, Methodol.*, vol. 57, no. 1, pp. 289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[42] J. T. Richardson, "Eta squared and partial eta squared as measures of effect size in educational research," *Educ. Res. Rev.*, vol. 6, no. 2, pp. 135–147, 2011. doi: 10.1016/j.edurev.2010.12.001.

[43] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic, 2013.

SMC