Dear PhD Austin J. Brockmeier

We truly appreciate all the valuable comments and recommendations provided. We thank you for the extensive revision of the manuscript. Your insights have been crucial in improving the quality and clarity of the work. Changes made on the document can be found in red color. Our point-to-point responses to the reviewers' comments are given below:

Point 1. *Overall the introduction and background is comprehensive and thorough with references. The proposed approach seems well founded and results are extensive.*

Response. We appreciate your recognition and We are pleased to hear that you found the manuscript comprehensive and thorough with appropriate references.

Point 2. *here are some concerns with the baselines not being as competitive. Ablation study of spectral filtering is needed or allowing spectral filtering for other FC baselines would be useful.*

Response. Thank you for your comment. We acknowledge that the document was not clear about both PLV and CCF being compared using the same frequency bands and time windows as specified in equation (5-6). To address this, we have revised the first two sentences of the experimental setup for clarity. The updated text is as follows:

(i) Preprocessing and trial-based extraction of *s-t-f* representations. For extracting the subject EEG dynamics over time accurately, the sliding window length of feature extraction is fixed to the next values: $\tau = [0.5, 1.0, 1.5, 2.0]$ s, having an overlap of 75%. Additionally, similar to authors in [**?**] the frequency bands of interest are selected with the lower frequency limit set at 4 Hz, the upper limit set to 40 Hz, and an overlap of 50%. [**?**] (ii) The estimation of the single-trial functional connectivity from the extracted *s-t-f* features. For comparison, the proposed KCS-FC is contrasted with two commonly used single-trial FC measures. To facilitate this comparison, the KCS-FC in Equation (**??**) is interchanged with new single-trial FC measures. These measures can be estimated from a frequency n, a time window w_t, and a pair of channels cc' (with $c \neq c'$, for all cc' in C) as described in [**?**].

$$\rho(\boldsymbol{x}^{c}_{rnw_t}, \boldsymbol{x}^{c'}_{rnw_t}) = \left\langle \boldsymbol{x}^{c}_{rnw_t}, \boldsymbol{x}^{c'}_{rnw_t} \right\rangle \tag{1}$$

$$\Delta\phi(\boldsymbol{x}^{c}_{rnw_t}, \boldsymbol{x}^{c}_{rnw_t}) = |\exp(j(\boldsymbol{\phi}^{c}_{rnw_t} - \boldsymbol{\phi}^{c}_{rnw_t}))| \tag{2}$$

Point 3. *One major concern is the validating of the hyper-parameter selection. Many times it is suggested that the 'best' is chosen. Is the 10-fold cross-validation split (mentioned on page 57) for trails or subjects? If the cross-validation is used how is the hyper-parameter search done? On the first split, or internally in each split, or only after all splits (test) are seen? If it is the latter than the hyper-parameter selection runs the risk of overfitting a dataset. and On page 68 it is clear that the split is done for trials of a subject. However, it is not clear from the text whether the hyper-parameter, but "GridSearchCV" is mentioned.*

But this uses a gridsearch based on the CV performance meaning it gives the best hyper-parameters after all test splits are seen. Again, a more rigorous way is to do internal hyper-parameter selection. This may be a common issue with the other baselines too—I've reviewed papers at top tier venues where this is brought up and papers are rejected due to this invalid hyper-parameter selection approach. In any case, the thesis should make this caveat clear.

Response. Thank you for your insight, we understand the concern about hyper-parameters overfitting. However, dDue to the low number of samples, we couldn't afford a nested cross-validation strategy. Instead, we used a 5-fold cross-validation at the subject level. While our approach doesn't completely eliminate overfitting risks, the small number of parameters in our model reduces this risk. Future work should focus on assessing model stability to overfitting with more robust validation techniques. EEG MI data has high intra-subject variability, significantly affecting training and testing sets. Our validation strategy follows the schema in [**?**]. Moreover, to check hyper-parameter sensitivity, we include the next two figure containing the CV mean test accuracy across all subjects for models KCS-FCnet and RKCS-FCnet. The figure above shows the cross-validation mean test

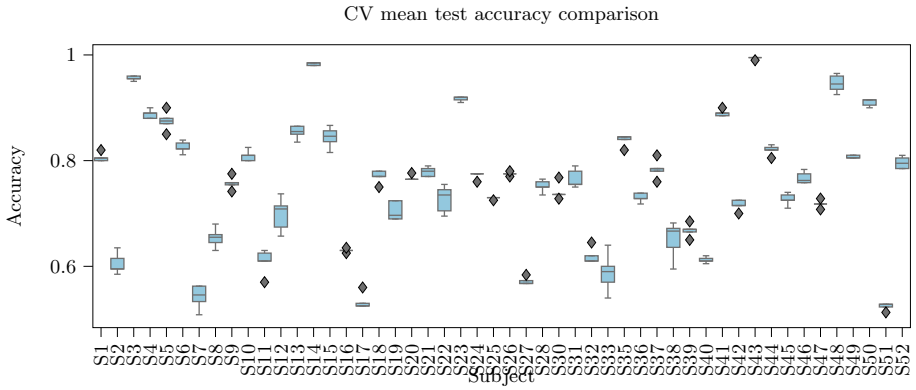

Fig. 1: Cross validation mean test accuracy comparison for KCS-FCnet model

accuracy for different subjects using the KCS-FCnet model. It indicates that the mean test validation accuracy remains relatively consistent regardless of the hyper-parameters used, suggesting robustness to hyper-parameter variations.

The image displays more variability in mean test accuracy across subjects. However, there is still a general trend of consistency, indicating that while some subjects exhibit more variability, the model maintains reasonable stability overall.

Point 4. *Page 4 Alpha "all ages and represents white matter" Not clear from context regarding white matter.*

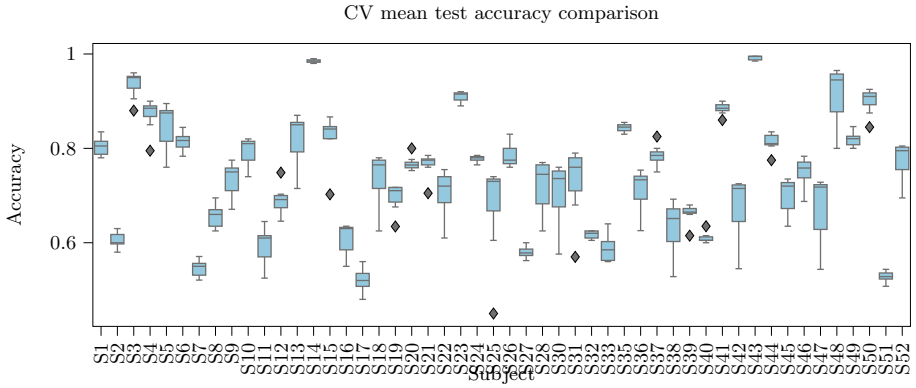

Fig. 2: Cross validation mean test accuracy comparison for RKCS-FCnet model

Response. Thanks for the recommendation. Indeed, it was a mistake. We have changed the text to make it clearer: Seen in all ages and can indicate white matter health.

Point 5. *"between a defined neighboring" not clear*

Response. We appreciate your observation. We modify the text to be to make it clearer Simpler strategies like Peak-valley representations, which extract local maximum and local minimum points within a defined timestamp, can also be used to predict MI tasks.

Point 6. *n example of this effectiveness was demonstrated by [Hassanpour et al., 2019], which found no statistical difference when testing DL methods with and without artifact removal strategies.*

Response. Thank you for the recommendation. The phrase was rewriten to give more clarity about the benefits of DL models. On ther other hand, Deep Learning (DL) strategies have shown no significant difference in testing results between models with and without clasical artifact removing strategies. This suggest that merly using DL models may be sufficient to handle artifcat removal [**?**,**?**].

Point 7. *Figure 1-10 "Potentially loss information" -¿ ""Potentially lose information"*

Response. Thank you for the comment, Same problem spoted in Figure 1-9. Images were modofied accordanly.

Point 8. *Last paragraph of Page 23. The last sentence "Finally, Renyi's entropy..." Seems out of place with respect to the interpretability paragraph.*

Response. Thank you for the observation, we rewrite the sentence as follows: To compare results we include a regularization technique based on Renyi's entropy to analyze how interpretability is affected when the cross-information potential of the internal FC of the KCS-FCnet is maximized.

Point 9. *Page 24-25. Not clear why "[Hz]" is in square brackets.*

Response. Thank you for pointing that out, We fix all unit notations.

Point 10. *Page 24 and Figure 4-2. The verbal description "subjects where asked to continue the MI task until the cross disappeared six seconds later" doesn't match the illustration which has only 3 seconds.* **Response**. Thank you for your observation. Indeed, there is a typo, we change six for three. three

Point 11. *Page 29 Around equation 4.1 the sentence lack punctuation.*

Response. We appreciate your feedback. We restructure the paragraph as: MI involves the neural simulation of a movement that, although not physically executed, activates the same areas of the brain as actual movements. EEGs capture these activations, and machine learning models can be built to interpret these signals, allowing the prediction of human actions purely from the EEGs.

From the mathematical perspective, let $\{\boldsymbol{X}_r\}_{r=1}^{R}$ be a multi-channel EEG observation from trial r, where each element $\boldsymbol{X}_r = \{\boldsymbol{x}_r^c \in \mathbb{R}^{N_t}\}_{c=1}^{N_c}$ contains $N_t \in \mathbb{N}$ time instants and $N_c \in \mathbb{N}$ number of channels. Moreover, there exists a function $\mathcal{F}$ (**??**) that exactly maps each trial $\boldsymbol{X}_r$ into the label space $y_r \in \{0, 1, \cdots, N_y\}$ representing the type of motor imagery with $N_y \in \mathbb{N}$ denoting the number of classes. The goal of EEG-MI classification is to find the best possible estimation function $\hat{\mathcal{F}}(\cdot; \boldsymbol{w})$ that approximates the true function $\mathcal{F}$ in (**??**), depending on a set of trainable parameters denoted as $\boldsymbol{w}$

$$\mathcal{F} : \boldsymbol{X}_r \mapsto y_r \quad \forall r \in \{1, \ldots, R\} \tag{3}$$

Point 12. *Pages 29–30. Superscripts R are misplaced in denominators of (4-4) and (4-5). Commas after equations that are part of sentences and followed by "where ... ".*

Response. Thank you for the feedback, we fix denominators position and add commas to equations.

$$TPR = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\sum_{r=1}^{R} \delta_K(\hat{y}r, y_r)\delta_K(y_r, 1)}{\sum_{r=1}^{R} \delta_K(y_r, 1)}, \tag{4}$$

$$FPR = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\sum_{r=1}^{R} \delta_K(\hat{y}_r, 1)\delta_K(y_r, 0)}{\sum_{r=1}^{R} \delta_K(y_r, 0)}, \tag{5}$$

Point 13. *Page 30. "Serves as a robust mathematical construct" "robust" is good word choice here. "The first moment relates to the mean" -¿ "The first moment is the mean".*

Response. We appreciate the comment, We change the phrase as suggested The first moment is the mean

Point 14. *Notation before equation 4.6 and equations 4-6, 4-7, and 4-8 is not clear. The use of $x^c(t)$ for a random variable but then X^c and dummy variable of integration should be x not x^c. $P_d(x^c)$ is odd construction. If the goal is index by channel then why not $p_{x_c}(x)$? The notation should make distinct the vector from scalar random variable case, to distinguish 4-8 and 4-9.*

Response. Thank you for the comment, We rewrite the notation to make it clearer as: he first moment, also known as the expectation, of a random variable $x_r^c(\tau)$, represents the expected value or mean of the variable. It serves as a measure of the center of the data distribution. The expectation can be mathematically defined as follows.

$$\mathbb{E}_\tau\{x_r^c(\tau)\} = \mu_r^c = \int_{-\infty}^{+\infty} \tau P_r^c(\tau) d\tau \tag{6}$$

where $P_r^c(\tau)$ is the probability density function of the channel c at trial r.

$$\begin{aligned}\tilde{\mathcal{R}}_r^c(\tau) = \mathbb{E}_\tau\{(x_r^c(\tau) - \mu_r^c)^2\} &= \int_{-\infty}^{+\infty} (\tau - \mu_r^c)^2 P_r^c(\tau) d\tau \\ &= \mathbb{E}_\tau\{(x_r^c(\tau))^2\} - (\mathbb{E}\{x_r^c(\tau)\})^2\end{aligned} \tag{7}$$

In the context of EEG-based MI-BCI systems, there is a common practice of centering each channel by eliminating the mean value. Consequently, when $\mu_r^c = 0$, the variance equation can be reformulated as:

$$\begin{aligned}\tilde{\mathcal{R}}_r^c(\tau) = \mathbb{E}_\tau\{(x_r^c(\tau))^2\} &= \int_{-\infty}^{+\infty} (\tau)^2 P_r^c(\tau) d\tau \\ &= \mathbb{E}_\tau\{(x_r^c(\tau))^2\}\end{aligned} \tag{8}$$

Point 15. *The text states Cov but the notation uses $\mathcal{R}$ ... Then on page 32 it is R*

Response. We value the feedback provided, we chnage notacion as: If two channels ar trial r x_r^c and $x_r^{c'}$ are centered, $\mu_r^c = 0$ and $\mu_r^{c'} = 0$, the covariance $\tilde{\mathcal{R}}_r^{c,c'}$

$$\tilde{\mathcal{R}}_r^{c,c'}(\tau) = \mathbb{E}_\tau\left[x_r^c(\tau) x_r^{c'}(\tau)\right] = \mathbb{E}_\tau\left[x_r^{c'}(\tau) x_r^c(\tau)\right] \tag{9}$$

Point 16. *On page 33 the covariance matrices are denoted as Σ but R_{CK} now is the number of trials in the Kth class. Here the assumption of zero-mean is used but not stated...*

Response. Thank you for the valuable comment. To avoid confussion between trials and autocorrelation functional we change the autocorrelation functional notation as: If two channels ar trial r x_r^c and $x_r^{c'}$ are centered, $\mu_r^c = 0$ and $\mu_r^{c'} = 0$, the covariance $\tilde{\mathcal{R}}_r^{c,c'}$

$$\tilde{\mathcal{R}}_r^{c,c'}(\tau) = \mathbb{E}_\tau \left[x_r^c(\tau) x_r^{c'}(\tau) \right] = \mathbb{E}_\tau \left[x_r^{c'}(\tau) x_r^c(\tau) \right] \tag{10}$$

Point 17. *Equation 4-19 not clear why diag is needed with the var operation (which would yield a vector). If the Cov (or properly typeset* Cov*) operator is used the* diag *operation is needed. It should be clarified that log is applied element wise (also* log *and* Tr *as these aren't variables but operations).*

Response. We value the feedback provided. We modify the notation as follows to include the suggestions.

Common Spatial Pattern (CSP) is a widely used technique for extracting features from EEG signals, especially for MI tasks. CSP helps in identifying spatial filters that maximize the difference of variance between classes. the core concept of CSP is the simultaneous diagonalization of two covariance matrices. CSP relies on **??** to estimate the sample covariance matrix for each trial as follows:

$$Cov(\boldsymbol{X}_r) = \boldsymbol{\Sigma}_r = \frac{1}{N_t - 1} \boldsymbol{X}_r \boldsymbol{X}_r^T \tag{11}$$

Where $Cov(\cdot)$ is the covariance fucntion, $\boldsymbol{\Sigma}_r \in \mathbb{R}^{N_c N_c}$ is the covariance matrix for any trial r, N_t is the number of time instances and $\boldsymbol{X}_r^T$ is the transpose of the EEG data matrix for trial r. Next, the average class covariance matrices Σ_{C1} and Σ_{C2} are calculated taking the sum of all covariance matrices for a class divided by the total number of trials in that class. For instance, if we have a total of R_{C1} trials in the set $C1$, the average clas matrix Σ_{C1} can be computed as follows:

$$\boldsymbol{\Sigma}_{C1} = \frac{1}{R_{C1}} \sum_{r=1}^{R_{C1}} \boldsymbol{\Sigma}_r; \quad \forall r \in C1 \tag{12}$$

$$\boldsymbol{z}_r = log \left(\frac{diag\left(Cov(\boldsymbol{S}_r)\right)}{Tr\left(Cov(\boldsymbol{S}_r)\right)} \right) \tag{13}$$

Where $Tr\left(\cdot\right)$ stands for the trace operator, $diag\left(\cdot\right)$ is the diagonal operator that extracts the elements in the principal diagonal, and $log(\cdot)$ represent the element wise logarithmic operator.

Point 18. *In equation 4-20 ,* $\mathbf{C}$ *has not been defined. I'm guessing its definition assumes full rank covariances, Equation 4-21 is not well formed.* $\mathbf{W}$ *nor* $\mathbf{w}_c$ *have been defined and isn't in the argument of the optimization (argmax), and 4-21 4-22 4-23 should all have trade-off parameters on the penalty/regularizer.*

Response. Thank you for the comment. We rewrite the formulas to match the sample covariances notation and continue with the CSP notation rewriting all regularization techniques as follows:

$$\hat{\boldsymbol{\Sigma}} = \{\boldsymbol{\Sigma} + \alpha \mathbf{I}\} \tag{14}$$

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C1} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C2} \boldsymbol{w}} - \alpha \left\| \boldsymbol{w} \right\|_1 \right) \tag{15}$$

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C1} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C2} \boldsymbol{w}} - \alpha \left\| \boldsymbol{w} \right\|_2 \right) \tag{16}$$

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C1} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C2} \boldsymbol{w}} - \alpha \left\| \boldsymbol{w} \right\|_1 \frac{1-\alpha}{2} \left\| \boldsymbol{w} \right\|_2 \right) \tag{17}$$

$$\tilde{\boldsymbol{\Sigma}} = \{\boldsymbol{\Sigma} + \alpha \operatorname{diag}(\boldsymbol{a})\} \tag{18}$$

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C1} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{\Sigma}_{C2} \boldsymbol{w}} - \alpha \left\| \boldsymbol{w} \right\|_{p,q} \right) \tag{19}$$

Point 19. *Equation 4-24. It is not clear what A is . If it's a vector and diag is needed then why not lowercase?*

Response. Thank you for the comment, $\boldsymbol{a}$ is a vector so we change the notation accordanly.

$$\tilde{\boldsymbol{\Sigma}} = \{\boldsymbol{\Sigma} + \alpha \operatorname{diag}(\boldsymbol{a})\} \tag{20}$$

Point 20. *Page 35, 2nd to last paragraph first sentence "during a predefined time window" is not clear.*

Response. Thank you for the feedback. We change the paragraph as follows: EEG signals often exhibit different patterns across different time intervals and frequency bands during a predefined mental taks as EEG patterns change with different mental tasks.

Point 21. *While continuous time notation (t) is used in 4.2.2 it seems that it is discrete time by choice of indexing $\{\}_{t=0}^{N_t}$ (also won't there be be N_t time points if discrete?). However 4-13 assumes continuous time but 4-31 uses n to discrete time with square brackets. (Later in 4-42 there is a mix of square brackets and t.)*

Response. We value you comment. To avoid confussion we set τ to be the continous variable and t is the n-th value of a discritization over τ. the next paragraphs were changed:

$$\mathbb{E}_\tau\{x_r^c(\tau)\} = \mu_r^c = \int_{-\infty}^{+\infty} \tau P_r^c(\tau) d\tau, \tag{21}$$

where $P_r^c(\tau)$ is the probability density function of the channel c at trial r.

Second Moment The second moment of a random variable, often referred to as the variance, denoted as $\tilde{\mathcal{R}}_r^c$, serves as the measure of how much the potential outcomes of the variable deviate from its mean value. This quantifies the spread or dispersion of the data. The formulation for variance can be computed as follows.

$$\begin{aligned}\tilde{\mathcal{R}}_r^c(\tau) = \mathbb{E}_\tau\{(x_r^c(\tau) - \mu_r^c)^2\} &= \int_{-\infty}^{+\infty} (\tau - \mu_r^c)^2 P_r^c(\tau) d\tau \\ &= \mathbb{E}_\tau\{(x_r^c(\tau))^2\} - (\mathbb{E}\{x_r^c(\tau)\})^2\end{aligned} \tag{22}$$

In the context of EEG-based MI-BCI systems, there is a common practice of centering each channel by eliminating the mean value. Consequently, when $\mu_r^c = 0$, the variance equation can be reformulated as:

$$\begin{aligned}\tilde{\mathcal{R}}_r^c(\tau) = \mathbb{E}_\tau\{(x_r^c(\tau))^2\} &= \int_{-\infty}^{+\infty} (\tau)^2 P_r^c(\tau) d\tau \\ &= \mathbb{E}_\tau\{(x_r^c(\tau))^2\}\end{aligned} \tag{23}$$

The Covariance, another important concept intertwined with variance, measures the joint variability or spread of two random variables. It determines how much the variables change together and quantifies their dependency. If two channels ar trial r x_r^c and $x_r^{c'}$ are centered, $\mu_r^c = 0$ and $\mu_r^{c'} = 0$, the covariance $\tilde{\mathcal{R}}_r^{c,c'}$ can be computed as follows.

$$\tilde{\mathcal{R}}_r^{c,c'}(\tau) = \mathbb{E}_\tau\left[x_r^c(\tau) x_r^{c'}(\tau)\right] = \mathbb{E}_\tau\left[x_r^{c'}(\tau) x_r^c(\tau)\right] \tag{24}$$

Covariance is essential in many strategies, such as Common Spatial Patterns (CSP), used for MI-BCI, as they capture the relationships between different EEG channels c, and c' at trial r. These relationships contain relevant information regarding neuronal oscillations and synchronization, both of which are crucial aspects of MI tasks.

Wide/Weak-Sense Stationarity Stochastic Processes A stochastic process is deemed wide-sense stationary (WSS) or weak-sense stationary if it satisfies the following two conditions.

1. The mean function or the first moment of the process is constant. This implies that the expected value or the average value of the process should be constant over time and not rely on the underlying time. Mathematically, it is represented by the equation:

$$\mathbb{E}_\tau\{x_r^c(\tau)\} = \mu_r^c = \text{constant, for all } \tau \tag{25}$$

   where $\mathbb{E}_\tau\{x_r^c(\tau)\}$ denotes the expected value of the random variable at any time τ and μ_r^c is the constant mean value.
2. The autocorrelation function or the second moment of the process depends only on the difference in time and not the actual time. This suggests that the correlation between two variables taken at different periods should only depend on the difference between those periods. It can be mathematically expressed as follows.

$$\mathbb{E}_\tau\{(x_r^c(\tau_1) - \mu_r^c)(x_r^c(\tau_2) - \mu_r^c)\} = \tilde{\mathcal{R}}_r^c(\tau_1 - \tau_2) = \tilde{\mathcal{R}}_r^c(\tau_\Delta) \tag{26}$$

   Where the autocorrelation function $\tilde{\mathcal{R}}_r^c(\tau_1 - \tau_2) = \tilde{\mathcal{R}}_r^c(\tau_\Delta)$ between two points at time instances τ_1 and τ_2 is only a function of their difference $\tau_\Delta = (\tau_2 - \tau_1)$.

Wiener Khinchin Theorem According to the Wiener Khinchin theorem, the autocorrelation function $\tilde{\mathcal{R}}_r^c(\tau)$ of a WSS random process is a Fourier transform pair with its power spectral density $\tilde{P}_r^c[\omega]$. This means that the autocorrelation function in the time domain corresponds to the power spectral density in the frequency domain and vice versa. This can be mathematically represented as follows.

$$\tilde{\mathcal{R}}_r^c(\tau) = \int_{\mathbb{R}} \tilde{P}_r^c[\omega] e^{j2\pi f\tau} df \tag{27}$$

Likewise, the power spectral density is given by the Fourier transform of the autocorrelation function.

$$\tilde{P}_r^c[\omega] = \int_{\mho} \tilde{\mathcal{R}}_r^c(\tau) e^{-j2\pi f\tau} d\tau \tag{28}$$

In the context of EEG-based MI-BCI systems, the Wiener-Khinchin theorem provides a useful tool for analyzing EEG signals. By transforming from the time domain to the frequency domain and vice versa, we gain insights into the spectral and temporal properties of the underlying stochastic processes, enabling efficient feature extraction and system identification.

Point 22. *In definition of cross-correlation, "cross" is misspelled as "Corss".*

Response. Thank you For pointing this out, We have corrected the typo.

$$\text{Cross-corr}_r^{c,c'}(\delta) = \frac{\mathbb{E}_t\left[\left(x_r^c[t] - \mu^c\right)\left(x_r^{c'}[t+\delta] - \mu^{c'}\right)\right]}{\sigma^c \sigma^{c'}} \tag{29}$$

Point 23. *Similar notational ambiguities exist between discrete Fourier time (like in FFT) from discrete-time Fourier transform or continuous time Fourier transform (4-12) and (4-13).*

Response. Thank you for the comment. We rewrite the formulas to avoid confussion, Wiener Khinchin uses the continuos time Fourier transform.

Wiener Khinchin Theorem According to the Wiener Khinchin theorem, the autocorrelation function $\tilde{\mathcal{R}}_r^c(\tau)$ of a WSS random process is a Fourier transform pair with its power spectral density $\tilde{P}_r^c[\omega]$. This means that the autocorrelation function in the time domain corresponds to the power spectral density in the frequency domain and vice versa. This can be mathematically represented as follows.

$$\tilde{\mathcal{R}}_r^c(\tau) = \int_{\mathbb{R}} \tilde{P}_r^c[\omega] e^{j2\pi f \tau} df \tag{30}$$

Likewise, the power spectral density is given by the Fourier transform of the autocorrelation function.

$$\tilde{P}_r^c[\omega] = \int_{\mho} \tilde{\mathcal{R}}_r^c(\tau) e^{-j2\pi f \tau} d\tau \tag{31}$$

Point 24. *Equations 4-33 and 4-34 are not incorporated into sentences (with punctuation and capitalization).*

Response. Thanks fot the valuable comment. We incorporate the equations into the sentences as follows:

According to authors in [**?**], the phase difference between channel c and c' is calculated as follows:

$$\Delta\phi_{c,c'}[t] = \phi_{x^c}[t] - \phi_{x^{c'}}[t] \tag{32}$$

hence, the PLV between the two channels is then computed by averaging the phase difference over the time dimension and taking the absolute value:

$$PLV_r^{c,c'} = \left| \frac{1}{N_t} \sum_{t=1}^{N_t} e^{j\Delta\phi_{c,c'}[t]} \right| \tag{33}$$

where j is the imaginary unit, values range from 0 to 1, with 1 indicating perfect phase locking (i.e., constant phase difference over time) and 0 indicating a random phase relationship.

Point 25. *Cross-spectral density on page 37 has not been defined in notation.*

Response. We appreciate your observation. we where refering to the cross-spectrum, we change it accordanly to maintain the same notation.

Point 26. *Notation for sign differs between 4-35 and 4-36.*

Response. Thank you for the comment, we use the same notation and include the definition of the sign function.

$$WPLI_r^{c,c'} = \frac{\left|\mathbb{E}_f\left[\left|\Im[S_r^{c,c'}[\omega]]\right| \operatorname{sign}\left(\Im[S_r^{c,c'}[\omega]]\right)\right]\right|}{\mathbb{E}\left[\left|\Im[S_{X^c,X^{c'}}]\right|\right]}, \tag{34}$$

where the $S_r^{c,c'}[\omega]$ is the cross-spectrum of signals in channels c and c' at trial r, $\Im[\cdot]$ denotes the imaginary part of a complex number, andsign$(\cdot)$ denotes the sign function:

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases} \tag{35}$$

Point 27. *After 4-44 "epredictions" .* ***Point 28***. *Thanks for pointing this typo out. We fix the spelling. prediction*

Point 29. *"two among R signals" but N_c is number of signals?, Equation 4-45 has some mistake, Equation 4-46 should have C instead of X?, Equation 4-49 now uses X but has only N instead of N_c. But 4-50 is back to C... and 4-51 is back to X.*

Response. Thank you to the recommendation. DFT and PDC were rewritten as follows:

Directed Transfer Function The Directed Transfer Function (DTF) is based on the concept of Granger causality and estimated using multivariate autoregressive (MVAR) models, which uses all signals simultaneously [**?**]. First, the MVAR is calculated as follows:

$$\boldsymbol{X}_r[t] = \sum_{t'=1}^{\tilde{p}} \left(\boldsymbol{A}_r[t']\boldsymbol{X}_r[t-t']\right) + \boldsymbol{\epsilon}[t] \tag{36}$$

where $\boldsymbol{X}_r[t]$ is the matrix containing all channels at trial r and time sample t, $\boldsymbol{A}_r[t']$ is the coefficient matrix at lag t', $\tilde{p}$ is the model order, and $\boldsymbol{\epsilon}_r[t]$ is the residual matrix at time t and trial r. Second, the MVAR model can be transformed into its frequency domain version using the Fourier transform to obtain the transfer fucntion matrix $\boldsymbol{H}_r[\omega]$ as follows:

$$\boldsymbol{H}_r[\omega] = \left[I - \sum_{t'=1}^{\tilde{p}} \boldsymbol{A}_r[t']\mathscr{F}\left(\boldsymbol{X}_r[t-t']\right)\right]^{-1} \tag{37}$$

where $\boldsymbol{H}_r[\omega]$ represents the transfer matrix, ω is the frequency and $\mathscr{F}$ is the Fourier transform. Finally, the influence from channel c to channel c' can be calculated as follows:

$$\mathrm{DTF}_r^{c,c'}[\omega] = \frac{|h_r^{c,c'}[\omega]|^2}{\sum_{k=0}^{N_c} |h_r^{c,k}[\omega]|^2} \tag{38}$$

where $h_r^{c,c'}\omega$ is the element of the transfer matrix $H_r[\omega]$ from channel c to channel c', and N_c is the total number of cahnnels. The square of the magnitude of $h_r^{c,c'}[\omega]$ is normalized by the sum of squares of the magnitudes of all elements in channel c of $H_r[\omega]$.

Partial Directed Coherence Partial Directed Coherence (PDC) is a frequency-domain measure based on MVAR models, designed to examine directed interactions between any pair of channels c and c'. Similar to the DTF, the PDC is derived using Granger causality principles, but the normalization method differs, allowing the differenciation between direct and indirect interactions. As in DFT the first step is to fit the MVAR model to the data as follows:

$$\boldsymbol{X}_r[t] = \sum_{t'=1}^{\tilde{p}} (\boldsymbol{A}_r[t']\boldsymbol{X}_r[t-t']) + \boldsymbol{\epsilon}[t] \tag{39}$$

where $\boldsymbol{X}_r[t]$ is the matrix containing all channels at trial r and time sample t, $\boldsymbol{A}_r[t']$ is the coefficient matrix at lag t', $\tilde{p}$ is the model order, and $\boldsymbol{\epsilon}_r[t]$ is the residual matrix at time t and trial r. Second, the MVAR model can be transformed into its frequency domain version using the Fourier transform to obtain the matrix of coefficient in the frequency domain $\tilde{\boldsymbol{A}}_r[\omega]$ as follows:

$$\begin{aligned} \tilde{a}_r^{c,c'}[\omega] &= \sum_{t'=0}^{\tilde{p}} \mathscr{F}\left(a_r^{c,c'}[t']\right) \\ &= \sum_{t'=0}^{\tilde{p}} a_r^{c,c'}[t'] \exp^{j2\pi\omega\tau t'} \end{aligned} \tag{40}$$

where $\tilde{a}_r^{c,c'}[0] = 1$ where $\boldsymbol{H}_r[\omega]$ represents the transfer matrix, ω is the frequency and $\mathscr{F}$ is the Fourier transform. Finally, the PDC from channel c to channel c' at frequency ω and trial r can be written as follows:

$$\mathrm{PDC}_r^{c,c'}[\omega] = \frac{\left|\tilde{a}_r^{c,c'}[\omega]\right|}{\sqrt{\left|\sum_{k=0}^{N_c} \tilde{a}_r^{c,k}[\omega]\right|^2}} \tag{41}$$

PDC ranges from 0 to 1, providing the degree of a direct influence of one channel on another in the frequency domain [**?**].

Point 30. *Is the density in 4-55 known or approximated to be Gaussian? Otherwise how is estimation of the density done in practice.* **Response**. Thank you for the comment. From the text is not clear that the probability density fucntion is unknow, so we clarify it as follows:

Transfer Entropy Transfer Entropy (TE) is an alternative measure of direct functional connectivity based on information theory. TE does not require a model of the interaction and is inherently nonlinear. TE for two observed channels $\boldsymbol{x}_r^c$ and $\boldsymbol{x}_r^{c'}$ can be written as:

$$\mathrm{TE}(\boldsymbol{x}_r^c \to \boldsymbol{x}_r^{c'}) = \sum_{x_r^{c'}[t+u], \boldsymbol{x}^{c'}[t]^{d_{\boldsymbol{x}_r^{c'}}}, \boldsymbol{x}_r^c[t]^{d_{\boldsymbol{x}_r^c}}} P\left(x_r^{c'}[t+u], x_r^{c'}[t]^{d_{\boldsymbol{x}_r^{c'}}}, \boldsymbol{x}_r^c[t]^{d_{\boldsymbol{x}_r^c}}\right) \log\left(\frac{P\left(x_r^{c'}[t+u] | x_r^{c'}[t]^{d_{\boldsymbol{x}_r^{c'}}}, \boldsymbol{x}_r^c[t]^{d_{\boldsymbol{x}_r^c}}\right)}{P\left(x_r^{c'}[t+u] | x_r^{c'}[t]^{d_{\boldsymbol{x}_r^{c'}}}\right)}\right) \quad (42)$$

where t is a time-index and u denotes the prediction time. $x_r^{c'}[t]^{d_{\boldsymbol{x}_r^{c'}}}$ and $x_r^c[t]^{d_{\boldsymbol{x}_r^c}}$ are $d_{\boldsymbol{x}_r^{c'}}-$ and $d_{\boldsymbol{x}_r^c}-$ dimensional delay vectors [**?**]. Note that the probability density functions is unknown. Hence, the first step is to estimate it using methods like histograms [**?**], kernel estimates [**?**], or K-nearest neighbors [**?**].

Point 31. *IN the RKHS section (pages 42-43) missing punctuation on equations and lemmas.*

Response. Thanks for the valuable comment. We rewrite the RKHS section as:

Reproducing Kernel Hilbert Spaces Let $\kappa(\cdot,\cdot)$ be a real-valued positive definite kernel, $\mathcal{H}$ a nonempty space, and $\phi(\cdot)$ be a function that maps any element in $\mathcal{X}$ to the $\mathcal{H}$ space. One can define a vector space by taking a linear combination of the form

$$\phi(\cdot) = \sum_{i=1}^{N} \alpha_i \kappa(\cdot, x_i) \quad (43)$$

where $N \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, and $\{x_i \in \mathcal{X}; \quad \forall i\}$. Thus, the dot product between ϕ and another function ϕ', defined as $\phi'(\cdot) = \sum_{j=1}^{N} \beta_j \kappa(\cdot, x_j)$, can be defined as follows:

$$\begin{aligned} \langle \phi, \phi' \rangle &= \sum_{i=1}^{N} \sum_{j=1}^{N'} \alpha_i \beta_j \kappa(x_i, x_j). \\ &= \sum_{j=1}^{N'} \beta_j \phi(x_j). \\ &= \sum_{i=1}^{N} \alpha_i \phi'(x_i). \end{aligned} \quad (44)$$

The dot product $\langle \cdot, \cdot \rangle$ is bilinear and symmetric, as $\langle \phi, \phi' \rangle = \langle \phi', \phi \rangle$. Furthermore, it is positive definite as for any function ϕ we have:

$$\langle \phi, \phi \rangle \sum_{i,j=1}^{N} \alpha_i \alpha_j \kappa(x_i, x_j) \geq 0. \quad (45)$$

Finally, the dot product is equivalent to the $\kappa(\cdot,\cdot)$ function as $\langle \kappa(\cdot,x),\phi\rangle = f(x)$ and $\langle \kappa(\cdot,x),\kappa(\cdot,x')\rangle = \kappa(x,x')$. By virtue of these properties, positive definite kernels κ are also called RKHS. In general, an RKHS can be defined as follows:

Definition 1 (Reproducing Knerl Hilbert Space). *Let $\mathcal{X}$ be a nonempty set and $\mathcal{H}$ a Hilbert space of functions $\phi : \mathcal{X} \to \mathcal{H}$. Then, $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with the dot product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, and the norm $|\phi|_{\mathcal{H}} = \sqrt{\langle\phi,\phi\rangle_{\mathcal{H}}}$, if there exists a bilinear function $\kappa : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ with the following properties.*

- *κ has the reproducing property $\langle \kappa(\cdot,x),\phi\rangle = f(x) \quad \forall f \in \mathcal{H}$, in particular $\langle \kappa(\cdot,x),\kappa(\cdot,x')\rangle = \kappa(x,x')$.*
- *κ spans $\mathcal{H}$, i.e. $\mathcal{H} = \overline{span\left\{\kappa\left(x,\cdot\right)|x\in\mathcal{X}\right\}}$, where $\overline{X}$ denotes the completion of the set X.*
- *RKHS uniquely determines κ, since the symmetry property is not fulfilled as $\langle \kappa(\cdot,x),\kappa'(\cdot,x')\rangle = \kappa(x,x')$ but $\langle \kappa'(\cdot,x),\kappa(\cdot,x')\rangle = \kappa'(x,x')$, then $\kappa(x,x') \neq \kappa'(x,x')$*

Point 32. *Equations 4-66 and 4-67 do not match exactly as 4-67 squares the entries of the kernel matrix.* **Response**. We value the comment. The two equations are not the same, the latter is a entropy-like measure emulating first equation. To avoid confussion we include the next paragraphs:

$$\hat{H}_2(X) = -\log\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\kappa_{\sqrt{2}\sigma}\left(x_i,x_j\right)\right), \tag{46}$$

the information potential, the quantity inside the log function, correspond to the squared norm $\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$, which by the law of large numbers converges to $\left|\mathbb{E}\left\{X\right\}\right|^2$. The empirical estimatior used in **??** links both Hilbert space representations and it can be rewritten using the terms of a Gram matrix $\boldsymbol{K}_{ij} = \kappa_{2\sigma}\left(x_i,x_j\right)$ as follows.

$$\hat{H}_2(X) = -\log\left(\frac{1}{n^2}\|\boldsymbol{K}\|_F^2\right) + C(\sigma), \tag{47}$$

where $C(\sigma)$ is the normalization factor of the Parzen window, the Frobenius norm of the Gram matrix $\boldsymbol{K}$, defined as $\|\boldsymbol{K}\|_F^2 = \mathrm{tr}(\boldsymbol{K}\boldsymbol{K})$, and $\mathrm{tr}(\cdot)$ is the matrix trace. The cross-inforamtion portential seeks to quantify the information contribution of each sample involved in a Parzen approximation and is directly related to Renyi's entropy by a strictly monotonic function [**?**]. Moreover, **??** is an entropy-like value possessing properties similar to Renyi's entropy without having to estimate probability distributions. Given a Gram matrix $\boldsymbol{K}$ with elements $k_{ij} = \kappa\left(x_i,x_j\right)$, a kernel-based formulation of Renyi's α-order entropy can be outlined as follows:

$$H_\alpha(\boldsymbol{K}) = \frac{1}{1-\alpha}\log\left(\mathrm{tr}\left(\boldsymbol{K}^\alpha\right)\right), \tag{48}$$

which holds that $\mathrm{tr}(\boldsymbol{K}) = 1$. The power α of $\boldsymbol{K}$ can be obtained using the spectral theorem [**?**].

Point 33. *Punctuation around 4-69...* **Response**. Thank you for the comment.

Punctuation was fixed around all equations in the document.

Point 34. *Section 4.2.7 is missing references especially for more modern approaches (GRU, LSTM, GCN, GAT, ELU)...*

Response. Thanks for your comment, the next are the references required. MLP: [?] CNN: [?] LSTM: [?] GRU: [?] GNN: [?] GCN: [?] GAT: [?] ELU: [?]

Point 35. *Is "TCFussionnet" spelled correctly?*

Response. Thanks for your comment, Indeed, "TCFussionnet" was spelled incorrectly, we rewrite as: TCNet-Fusion

Point 36. *Figure 5-1. The first equation in the blue box don't make sense as f is a dummy variable on right side. Same problem with 5-3... and What does infinitely long interval T mean? If infinite then what sense is T? Isn't T the dimension of the vectors? Then the multivariate frequencies need to integrate over T dimensions not just* $\mathbb{R}$.

Response. Thank you for the comment. The function 5-3 represents the integral of the cross-spectral density $S_r^{cc'}$ that makes sense in the case that the cross-spectral distribution is differentiable over $\varpi \in \Omega$, also the equation 5-3 integrates and 5-2 integrates over $\varpi \in \Omega$. Finally, we rewrite it as follows: The Wiener-Khinchin's theorem states that a real-valued auto-correlation function, $R_r^c(\tau)$, of a weak-sense stationary stochastic process $x_r^c(\tau)$ can be defined as [?]:

$$R_r^c(\tau) = \int_{\varpi \in \Omega} \exp(j2\pi\tau\varpi) dP_r^c(\varpi), \tag{49}$$

where $P_r^c(\varpi) \in \mathbb{R}[0,1]$ is a monotonic spectral distribution function absolutely continuous and differentiable over frequency $\varpi \in \Omega$. According to Bochner's theorem the univariable relationship in Equation (??) a stationary positive-definite kernel $\kappa_r^{cc'}(\Delta_x) = \kappa(\boldsymbol{x}_r^c, \boldsymbol{x}_r^{c'})$ can be related with its cross-spectral counterpart $S_r^{cc'}(\varpi)$ if and only if the following assumption holds between both spectral representations [?]:

$$\kappa_r^{cc'}(\Delta_x) = \int_{\varpi \in \Omega} \exp\left(j2\pi\Delta_x\varpi\right) S_r^{cc'}(\varpi) d\varpi, \tag{50}$$

where $\Delta_x = \boldsymbol{x}_r^c - \boldsymbol{x}_r^{c'}$ is the vector delay, $\varpi \subseteq \Omega$ is the frequency domain that contains the bandwidth set of analysis Ω, and $S_r^{cc'}(\varpi)$ is the cross-spectral density that preserves the following equality: $S_r^{cc'}(\varpi) = dP_r^{cc'}(\varpi)/d\varpi$, with $P_r^{cc'}(\varpi) \in \mathbb{R}[0,1]$ being the cross-spectral distribution that is related to the mapping kernel, $\kappa : \mathbb{R}^{N_t} \times \mathbb{R}^{N_t} \to \mathbb{R}$. As regards $\kappa(\boldsymbol{x}_r^c, \boldsymbol{x}_r^{c'}) = \langle \phi(\boldsymbol{x}_r^c), \phi(\boldsymbol{x}_r^{c'}) \rangle_{\mathcal{H}}$, it is a positive-definite stationary kernel inducing the nonlinear feature mapping $\phi(\cdot)$ to a Reproducing Kernel Hilbert Space, $\mathcal{H}$. Notation $\langle \cdot, \cdot \rangle$ represents the dot product. Therefore, we compute the cross-spectral distribution $P_r^{cc'}(\varpi)$ within a

bandwidth Ω, as below:

$$\begin{aligned} P_r^{cc'}(\varpi) &= 2\int_{\varpi\in\Omega} S_r^{cc'}(\varpi)d\varpi, \\ &= 2\int_{\varpi\in\Omega} \mathscr{F}\left\{\kappa(\boldsymbol{x}_r^c, \boldsymbol{x}_r^{c'})\right\} d, \end{aligned} \tag{51}$$

where notation $\mathscr{F}\{\cdot\}$ stands for the Fourier transform.

Point 37. *I think the first summation in 5-5 should be over n not r.* **Response**. We value the comment. Indeed, the summation shouldn't be over trials r but over filters and time windows.

$$\hat{P}_r^{cc'}(\boldsymbol{u}^{cc'}, \kappa_x(\cdot;\sigma)) = \sum_{n=1}^{N_f}\sum_{w_t=1}^{N_t} u_{nw_t}^{cc'}\kappa_x\left(\boldsymbol{x}_{rnw_t}^c, \boldsymbol{x}_{rnw_t}^{c'};\sigma\right), \tag{52}$$

where $\boldsymbol{u}^{cc'} \in \mathbb{R}^{N_f N_t}$ is the spatio-temporal-frequency (textits-t-f) relevance vector that codes the pairwise undirected dependency between channels and contains the values $u_{nw_t}^{cc'}$ holding the relevance value estimated at n-th frequency and w_t window.

Point 38. *The vector concatenation before 5-6 needs to be more explicit to include the time and frequency points* **Response**. Thank you for teh valuable comment. We rewrite the equations to be more explicit about the concatenation. we estimate the cross-spectral distribution $\hat{\boldsymbol{P}}_r^{cc'}$ between a pair of channels c and c' at trial r as follows:

$$\hat{P}_r^{cc'}(\boldsymbol{u}^{cc'}, \kappa_x(\cdot;\sigma)) = \sum_{n=1}^{N_f}\sum_{w_t=1}^{N_t} u_{nw_t}^{cc'}\kappa_x\left(\boldsymbol{x}_{rnw_t}^c, \boldsymbol{x}_{rnw_t}^{c'};\sigma\right), \tag{53}$$

where $\boldsymbol{u}^{cc'} \in \mathbb{R}^{N_f N_t}$ is the spatio-temporal-frequency (textits-t-f) relevance vector that codes the pairwise undirected dependency between channels and contains the values $u_{nw_t}^{cc'}$ holding the relevance value estimated at n-th frequency and w_t window.

Using the single-trial kernel-based spectral distribution representation in Equation (**??**), we propose extracting the sparse functional connectivity, using the Elastic Net regularization [**?**], aiming to finding discriminative and interpretable brain activity patterns.

$$\boldsymbol{u}^* = \arg\min_{\boldsymbol{u}} \mathbb{E}\left\{\sum_{r=1}^{R}\left\|\sum_{c,c'=1}^{N_c}\hat{P}_r^{cc'}(\boldsymbol{u}^{cc'}, \kappa_x(\cdot;\sigma)) - y_r\right\|_2^2\right\} + \alpha\sum_{c,c'=1}^{N_c}\|\boldsymbol{u}^{cc'}\|_1 + \frac{1-\alpha}{2}\sum_{c,c'=1}^{N_c}\|\boldsymbol{u}^{cc'}\|_2 \quad : \forall c < c', \tag{54}$$

where $\alpha \in \mathbb{R}^+$ is the regularization hyperparameters, $\|\cdot\|_q$ is the ℓ_q-norm.

Without loss of generallity, equation (**??**) can be rewritten as:

$$\tilde{\boldsymbol{u}}^* = \arg\min_{\tilde{\boldsymbol{u}}} \mathbb{E}\left\{\sum_{r=1}^{R} \|\tilde{\boldsymbol{u}}\tilde{\boldsymbol{p}}_{\boldsymbol{r}}^{\top} - y_r\|_2^2\right\} + \alpha\|\tilde{\boldsymbol{u}}\|_1 + \frac{1-\alpha}{2}\|\tilde{\boldsymbol{u}}\|_2, \tag{55}$$

where $\top$ is the vector transpose, $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{\kappa}}$ stand for the spatio-temporal-frequency relevance matrix and cross-spectral distribution vector concatenation obtained as:

$$\begin{aligned}\tilde{\boldsymbol{u}} &= \left[u_{11}^{12}, u_{11}^{13}, \cdots, u_{11}^{1N_c}, u_{11}^{23}, \cdots, u_{11}^{(N_c-1)N_c}, u_{12}^{12}, \cdots, u_{N_f N_t}^{(N_c-1)N_c}\right], \\ \tilde{\boldsymbol{p}}_r &= \left[\hat{P}_{r11}^{12}, \hat{P}_{r11}^{13}, \cdots, \hat{P}_{r11}^{1N_c}, \hat{P}_{r11}^{23}, \cdots, \hat{P}_{r11}^{(N_c-1)N_c}, \hat{P}_{r12}^{12}, \cdots, \hat{P}_{rN_f N_t}^{(N_c-1)N_c}\right],\end{aligned} \tag{56}$$

Note that for simplicity the term $\hat{P}_r^{cc'}(\boldsymbol{U}_r^{cc'}, \kappa_x(\cdot;\sigma))$ is rewritten as $\hat{P}_{rnw_t}^{cc'}$.

Point 39. *Page 58 "the sliding window's impact" -¿ "impact of the sliding window's length"* **Response**. Thanks for the comment, we rewrite the sentence as suggested. impact of the sliding window's length

Point 40. *"One the other hand" -¿ "Additionally" There is no contrast here and the "On the one hand" wasn't stated.*

Response. Thank you for the comment. we change as suggested: Additionally

Point 41. *It seems to be an unfair comparison to not use frequency bands for PLV and CCF. A better comparison and ablation study would allow a version of equation 5-6 for CCF and PLV when they are also frequency dependent. PLV (and the underlying Hilbert transform) in particular is more interpretable for a single frequency band.*

Response. Thank you for teh valuable comment. The document was not clear enough about both PLV and CCF being compared within the same frequency bands, time windows using equation 5-6. we change the frist two sentences of the experimental set-up as follows:

(i) Preprocessing and trial-based extraction of s-t-f representations. For extracting the subject EEG dynamics over time accurately, the sliding window length of feature extraction is fixed to the next values: $\tau = [0.5, 1.0, 1.5, 2.0]$ s, having an overlap of 75%. Additionally, similar to authors in [**?**] the frequency bands of interest are selected with the lower frequency limit set at 4 Hz, the upper limit set to 40 Hz, and an overlap of 50%.

(ii) The estimation of the single-trial functional connectivity from the extracted s-t-f features. For comparison, the proposed KCS-FC is contrasted with two commonly used single-trial FC measures. To facilitate this comparison, the KCS-FC in Equation (**??**) is interchanged with new single-trial FC measures. These measures can be estimated from a frequency n, a time window w_t, and a pair of channels cc' (with $c \neq c'$, for all cc' in C) as described in [**?**].

$$\rho(\boldsymbol{x}_{rnw_t}^{c}, \boldsymbol{x}_{rnw_t}^{c'}) = \left\langle \boldsymbol{x}_{rnw_t}^{c}, \boldsymbol{x}_{rnw_t}^{c'} \right\rangle \tag{57}$$

$$\Delta\phi(\boldsymbol{x}^c_{rnw_t}, \boldsymbol{x}^c_{rnw_t}) = |\exp(j(\boldsymbol{\phi}^c_{rnw_t} - \boldsymbol{\phi}^c_{rnw_t}))| \tag{58}$$

Point 42. *Relatedly on 5-6 couldn't a logistic regression model also be used? Where the logit is linear and passed through a sigmoid function.*

Response. Thank you for your comment. Indeed, logistic regression could be applied directly to Equation (5-6) by optimizing the model using cross-entropy loss. Since both LDA and logistic regression fundamentally operate as linear classifiers, the results between the two methods might not differ significantly. However, it's important to note that while logistic regression directly models the probability of class membership, LDA focuses on maximizing the separability between classes.

Point 43. *"Rather, the CFF" -¿ "Rather the CCF"*

Response. Thank for the comment. We fix the notation across the document.

Point 44. *The composition notation on 6-1 is non-standard in its usage as it is overloaded (a bivariate function being applied to each pair. The matrix/vector valued function should be defined explicitly by iterating over the channel pairs.*

Response. Thank you for the comment, we include an extra function to avoid overloaded equation.

$$\hat{\boldsymbol{P}}_r(\boldsymbol{w}_f) = \tilde{K}(\cdot;\sigma) \circ \varphi(\boldsymbol{X}_r; \boldsymbol{w}_f), \tag{59}$$

where $\hat{\boldsymbol{P}}_r\ (\boldsymbol{w}_f) \in [0,1]^{N_c \times N_c \times N_f}$, N_f is the number of convolutional filters, notation $\circ$ stands for function composition, $\varphi(\cdot; \boldsymbol{w}_f)$ is a 1-D convolutional layer that can be used to automatically extract frequency patterns ruled by the weight vector $\boldsymbol{w}_f \in \mathbb{R}^{\Delta_t}$, with $\Delta_t < N_t$. Of note, in Equation (**??**) function $\tilde{K}(\cdot;\sigma)$ is the convolutional filter concatenation of all pair-wise values $\kappa_x(\boldsymbol{x}^c_{rf}, \boldsymbol{x}^{\boldsymbol{c'}}_{\boldsymbol{rf}}; \sigma)$ and is obtained as:

$$\tilde{K}(\tilde{\boldsymbol{X}}_r; \sigma) = \left[\boldsymbol{K}_{r1}, \boldsymbol{K}_{r2}, \cdots, \boldsymbol{K}_{rf}, \cdots, \boldsymbol{K}_{rN_f}\right], \tag{60}$$

where $\boldsymbol{K}_{rf} \in \mathbb{R}^{N_c \times N_c}$ is the kernel matrix for a trial r at a convolutional filter f and it is calculated as follows:

$$\boldsymbol{K}_{rf} = \begin{bmatrix} \kappa_x(\boldsymbol{x}^1_{rf}, \boldsymbol{x}^1_{rf}; \sigma) & \kappa_x(\boldsymbol{x}^1_{rf}, \boldsymbol{x}^2_{rf}; \sigma) & \cdots & \kappa_x(\boldsymbol{x}^1_{rf}, \boldsymbol{x}^{N_c}_{rf}; \sigma) \\ \kappa_x(\boldsymbol{x}^2_{rf}, \boldsymbol{x}^1_{rf}; \sigma) & \kappa_x(\boldsymbol{x}^2_{rf}, \boldsymbol{x}^2_{rf}; \sigma) & \cdots & \kappa_x(\boldsymbol{x}^2_{rf}, \boldsymbol{x}^{N_c}_{rf}; \sigma) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_x(\boldsymbol{x}^{N_c}_{rf}, \boldsymbol{x}^1_{rf}; \sigma) & \kappa_x(\boldsymbol{x}^{N_c}_{rf}, \boldsymbol{x}^2_{rf}; \sigma) & \cdots & \kappa_x(\boldsymbol{x}^{N_c}_{rf}, \boldsymbol{x}^{N_c}_{rf}; \sigma). \end{bmatrix} \tag{61}$$

Point 45. *The notation in 6-2 is not clear (beyond the concerns with 6-1). The expectation is taken with respect of f as random variable following what distribution? Is it uniform? If so what does $\hat{\Omega}$ as an argument signify and is it related to the expectation?* **Response**. Thanks for the comment. We

rewrite the fucntion as: We compute the average functional connectivity measure $\tilde{\boldsymbol{P}}_r \in \mathbb{R}^{N_c \times N_c}$ over convolutional filters, as follows:

$$\tilde{\boldsymbol{P}}_r = \text{AvgPooling}_f \left(\hat{\boldsymbol{P}}_r(\boldsymbol{w}_f) \right), \tag{62}$$

where $\boldsymbol{w}_f$ is the f-th convolutional filter, N_f is the number of convolutional filters. This measure provides a way to analyze how different frequency bands of a single-trial EEG relate to each other across channels. After computing the average functional connectivity measure and taking advantage of the symmetric property of the Gaussian functional connectivity, the vectorized version of $\tilde{\boldsymbol{P}}_r$ is calculated as:

$$\overline{\boldsymbol{p}}_r = \left[\tilde{p}_r^{12}, \tilde{p}_r^{13}, \cdots, \tilde{p}_r^{cc'}, \cdots, \tilde{p}_r^{(N_c-1)N_c} \right]; \forall c < c', \tag{63}$$

where $\overline{\boldsymbol{p}}_r \in \mathbb{R}^{N_c(N_c-1)/2}$. Next, a the softmax-based output layer is applied over vector $\overline{\boldsymbol{p}}_r$ to obtain the MI class probability membership $\hat{\boldsymbol{y}}_r \in [0,1]^{N_y}$ as:

Point 46. *Notation in 6-3 has a couple mistakes. The output should be a vector so* $\mathbf{v}$ *is a matrix (and tensor product should be standard matrix product after transposing* $\mathbf{v}$*. (Really it should be* $\mathbf{V}$ *and the parentheses are not matching in size.)*

Response. We value the comment. You are right $\mathbf{v}$ is a matrix. we rewrite the equation as follows:

$$\hat{\boldsymbol{y}}_r = \text{softmax}\left(\boldsymbol{V}\overline{\boldsymbol{p}}_r + \boldsymbol{b} \right), \tag{64}$$

where $\boldsymbol{V} \in \mathbb{R}^{N_c(N_c-1)/2 \times N_y}$, $\boldsymbol{b} \in \mathbb{R}^{N_y}$. In addition, a gradient descent-based framework using back-propagation is employed to optimize the parameter set $\Theta = \{\boldsymbol{w}_f, \boldsymbol{V}, \boldsymbol{b}, \sigma; \forall f \in \{1, 2, \ldots, N_f\}\}$, as follows [**?**]:

Point 47. *Top of page 69 is missing reference in Figure.* **Response**. Thank you for the comment. We fix the figure refence.

Point 48. *In equation 7-4 (also in Figure 7-1) the composition notation is again non-standard. Composition is functions applied from right to left. So it should be there kernel on the left applied to a tuple of the filtered versions, which would be something like* $\kappa_G(\cdot,\cdot;\sigma) \circ \left(\phi(x^{c'}; w_f), \phi(x^c; w_f) \right)$.

Response. Thank you for the valuable comment. We rewrite the equation as suggested:

$$H(\chi; \mathbf{w}_f, \sigma) = -\log \left(\frac{1}{N_c^2} \sum_{c,c'=1}^{N_c} \kappa_x(\cdot,\cdot;\sigma) \circ \left(\varphi(x^{c'}; w_f), \varphi(x^c; w_f) \right) \right) \tag{65}$$

Point 49. *Description at bottom of page 80 is not precise. It is maximized when the each sample is only similar to its self, which will induce a diagonal* $\mathbf{K}$ *matrix. However, this case means no channels are similar which is not useful. There is a middle ground where there is structured sparsity in the* $\mathbf{K}$.

Response. Thank you for your comment. You are correct in pointing out that the description on page 80 regarding optimization could induce a diagonal $\mathbf{K}$ matrix. To clarify, the objective is similar to the information bottleneck principle, where we aim to maximize the cross-information potential while maintaining the accuracy obtained. We have now included an optimization problem which clarifies that the cross-information potential acts as a regularization technique on the usual cross-entropy optimization problem.

Since the main objective is to enhance the interpretability while maintaining the results obtained in **??**, we include the $H(\chi; \mathbf{w}_f, \sigma)$ into the optimization problem as follows:

$$\Theta^* = \arg\min_{\Theta} \quad \mathbb{E}_r\Big\{\mathcal{L}(\boldsymbol{y}_r, \hat{\boldsymbol{y}}_r|\Theta) - H(\tilde{\boldsymbol{P}}_r; \mathbf{w}_f, \sigma); \forall r \in \{1, 2, \ldots, R\}\Big\}, \tag{66}$$

where $\tilde{\boldsymbol{P}}_r$ is the average functional connectivity measure calculated in Equation (**??**) and $\hat{\boldsymbol{y}}$ is the class probability vector obtained in Equation (**??**). Here, the objective is akin to the information bottleneck principle, where we aim to maximize the cross-information potential while maintaining the accuracy obtained in **??**. In this context, the cross-information potential serves as a regularization technique on the typical cross-entropy optimization problem.

Point 50. *This chapter is missing a description of the optimization problem. The entropy is maximized with respect to what variables? The full optimization with the trade-off with respect to the cross-entropy should be detailed.*

Response. Thank you for the recommendation. We agree that a optimization problems is missing, so it is included as follows: Since the main objective is to enhance the interpretability while maintaining the results obtained in **??**, we include the $H(\chi; \mathbf{w}_f, \sigma)$ into the optimization problem as follows:

$$\Theta^* = \arg\min_{\Theta} \quad \mathbb{E}_r\Big\{\mathcal{L}(\boldsymbol{y}_r, \hat{\boldsymbol{y}}_r|\Theta) - H(\tilde{\boldsymbol{P}}_r; \mathbf{w}_f, \sigma); \forall r \in \{1, 2, \ldots, R\}\Big\}, \tag{67}$$

where $\tilde{\boldsymbol{P}}_r$ is the average functional connectivity measure calculated in Equation (**??**) and $\hat{\boldsymbol{y}}$ is the class probability vector obtained in Equation (**??**). Here, the objective is akin to the information bottleneck principle, where we aim to maximize the cross-information potential while maintaining the accuracy obtained in **??**. In this context, the cross-information potential serves as a regularization technique on the typical cross-entropy optimization problem.

Point 51. *Reference to ADAM optimizer should be given.* **Response**. Thank you for the comment, we add the reference as follows: The ADAM algorithm [**?**] is employed to tune the model parameters with a learning rate of 0.1

Point 52. *Legend is missing in 7-3 (although the colors are described in the text, it should be in the figure and/or caption).*

Response. Thank you for the recommendation, we include the legend to the figure and describe each color in the caption.

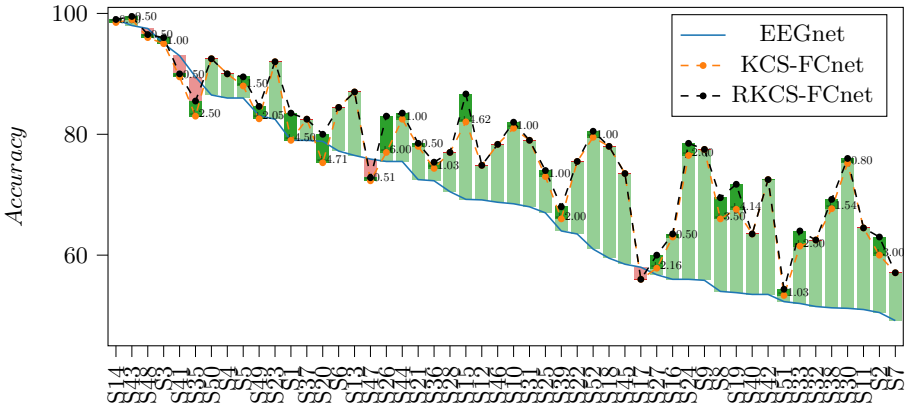

Fig. 3: Comparison of subject-specific average accuracy across EEGnet, KCS-FCnet, and RKCS-FCnet models. Subjects are arranged in descending order based on the accuracy of the EEGnet model. The color-coded bars illustrate performance shifts, with dark green(▮) indicating an increase in accuracy on the RKCS-FCnet over KCS-FCnet, light green(▮) indicating an increase in accuracy on KCS-FCnet over EEGnet, and while red ▮ indicates an accuracy reduction.

Point 53. *It seems some of the future work in the conclusion directly contradicts what was said earlier about avoiding complexity of graph neural networks. Perhaps those earlier comments could be tempered.*

Response. Thank you for the valuable comment, we agree and change the paragraph as: While Graph Neural Networks are capable of managing complex graph-structured data, they can be computationally demanding due to the necessity of processing all nodes and their neighbors within the graph. However, ongoing advancements in optimization techniques are continually improving their efficiency [**?**].

Point 54. *Units like mV, V, and Hz are not italics. Nor should reference names be italics. "Table X", "Figure X", "Section X", etc. should be capitalized and not abbreviated like in Chapter 6, but consistent throughout.*

Response. Thank you for the recommendation we removed all italics refering to Units and names. Also "Table X", "Figure X", "Section X" was changed as suggested.

References

1. Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S.U., Altuwaijri, G.A., Abdul, W., Bencherif, M.A., Faisal, M.: Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. Neural Computing and Applications **35**(20), 14681–14722 (2023)
2. Amin, S.U., Alsulaiman, M., Muhammad, G., Mekhtiche, M.A., Hossain, M.S.: Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion. Future Generation computer systems **101**, 542–554 (2019)
3. Ang, K.K., Chin, Z.Y., Zhang, H., Guan, C.: Filter bank common spatial pattern (fbcsp) in brain-computer interface. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). pp. 2390–2397. IEEE (2008)
4. Bochner, S.: Harmonic analysis and the theory of probability. University of California press (2020)
5. Cattai, T., Colonnese, S., Corsi, M.C., Bassett, D.S., Scarano, G., Fallani, F.D.V.: Phase/amplitude synchronization of brain signals during motor imagery bci tasks. IEEE Transactions on Neural Systems and Rehabilitation Engineering **29**, 1168–1177 (2021)
6. Choi, I., Kim, W.C.: Detecting and analyzing politically-themed stocks using text mining techniques and transfer entropy—focus on the republic of korea's case. Entropy **23**(6), 734 (2021)
7. Cohen, L.: The generalization of the wiener-khinchin theorem. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). vol. 3, pp. 1577–1580. IEEE (1998)
8. Demir, A., Koike-Akino, T., Wang, Y., Erdoğmuş, D.: Eeg-gat: graph attention networks for classification of electroencephalogram (eeg) signals. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 30–35. IEEE (2022)
9. Gao, H., Wang, X., Chen, Z., Wu, M., Cai, Z., Zhao, L., Li, J., Liu, C.: Graph convolutional network with connectivity uncertainty for eeg-based emotion recognition. IEEE Journal of Biomedical and Health Informatics (2024)
10. Gaxiola-Tirado, J.A., Salazar-Varas, R., Gutiérrez, D.: Using the partial directed coherence to assess functional connectivity in electroencephalography data for brain–computer interfaces. IEEE Transactions on Cognitive and Developmental Systems **10**(3), 776–783 (2017)
11. Giraldo, L.G.S., Rao, M., Principe, J.C.: Measures of entropy from data using infinitely divisible kernels. IEEE Transactions on Information Theory **61**(1), 535–548 (2014)
12. Hassanpour, A., Moradikia, M., Adeli, H., Khayami, S.R., Shamsinejadbabaki, P.: A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals. Expert Systems **36**(6), e12494 (2019)
13. Ju, C., Guan, C.: Graph neural networks on spd manifolds for motor imagery classification: A perspective from the time–frequency analysis. IEEE Transactions on Neural Networks and Learning Systems (2023)
14. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kuang, P.C.: Measuring information flow among international stock markets: An approach of entropy-based networks on multi time-scales. Physica A: Statistical Mechanics and its Applications **577**, 126068 (2021)

16. Kumar, S., Sharma, A., Tsunoda, T.: Brain wave classification using long short-term memory network based OPTICAL predictor. Scientific reports **9**(1), 1–13 (2019)
17. Luo, T.j., Zhou, C.l., Chao, F.: Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network. BMC bioinformatics **19**(1), 1–18 (2018)
18. Martínez-Cancino, R., Delorme, A., Wagner, J., Kreutz-Delgado, K., Sotero, R.C., Makeig, S.: What can local transfer entropy tell us about phase-amplitude coupling in electrophysiological signals? Entropy **22**(11), 1262 (2020)
19. Qiumei, Z., Dan, T., Fenghua, W.: Improved convolutional neural network based on fast exponentially linear unit activation function. Ieee Access **7**, 151359–151367 (2019)
20. Raeisi, K., Khazaei, M., Croce, P., Tamburro, G., Comani, S., Zappasodi, F.: A graph convolutional neural network for the automated detection of seizures in the neonatal eeg. Computer methods and programs in biomedicine **222**, 106950 (2022)
21. Rezaei, E., Shalbaf, A.: Classification of right/left hand motor imagery by effective connectivity based on transfer entropy in electroencephalogram signal. Basic and Clinical Neuroscience **14**(2), 213 (2023)
22. Rodrigues, P., Stefano, C., Attux, R., Castellano, G., Soriano, D.: Space-time recurrences for functional connectivity evaluation and feature extraction in motor imagery brain-computer interfaces. Medical & biological engineering & computing **57**(8), 1709–1725 (2019)
23. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization. Human brain mapping **38**(11), 5391–5420 (2017)
24. Tay, J.K., Narasimhan, B., Hastie, T.: Elastic net regularization paths for all generalized linear models. Journal of statistical software **106** (2023)
25. Zhang, A., Lipton, Z.C., Li, M., Smola, A.J.: Dive into deep learning. arXiv preprint arXiv:2106.11342 (2021)
26. Zhang, K., Robinson, N., Lee, S.W., Guan, C.: Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network. Neural Networks **136**, 1–10 (2021)