# Mapping Raw Neural Activity to Learned Representations: A Deep Encoding Approach for BCIs

Anonymous ICME submission

***Abstract*—Brain-computer interface (BCI) is a technology that enables direct connection and interaction of brain activity with external devices or systems. The encoding and decoding of neural signals play a crucial role in BCIs. The quality of such encodings are the key to robust and accurate information exchange and control between the brain and external devices. Currently, the limited capabilities of conventional brain signal processing is restricting a wider application of BCIs. One of the more enduring issue is the problem of uniquely mapping neural information to a deep learning model, or specifically, to its model parameters. Such a mapping would be advantageous for the interpretable analysis of neural signals within the framework of a deep model. In this work we first propose a deep learning framework capable of reproducible encoding of EEG signals. We confirm that the learned representations correspond to the specifics of the input data. We subsequently verify the stable connection between the deep learned representation and motor imagery labels. In this way we show that there is a robust and unique correspondence between neural activity and learned representations. This correspondence is important to establish interpretability between model parameters and actual input data. All analysis and visualization codes in the article are available at https://anonymous.4open.science/r/Mapping-correlation-between-model-parameters-and-EEG-ED5C.**

***Index Terms*—Electroencephalography, deep learning, feature extraction, representation learning, interpretability, representation dissimilarity analysis**

## I. Introduction

The study of *representations* is a fundamental task in machine learning. Representations, in this context, refers to the product of *compressing* or *summarizing* information, where information in its original form, which might be high-dimensional or -complexity, is transformed to a form which is much less complex or low-dimensional, but, ideally, still retains the essential content of the original form. For conciseness, we refer to the search for an optimal transformation which fulfills this task *well* (see below) as *representation learning*. For modern applications which optimize for computational effort and storage capacities, it is easy to see how such an transformation would be seen as highly-desirable, assuming that several important criteria (among others) could be fulfilled:

- Fidelity: this refers to how closely the representation actually stands for or embodies the actual information. In the process of representation learning, this is *the* fundamental requirement. There have been many works [1] [2] [3]dedicated to this specific aspect, but it remains an open problem, primarily due to the multitude of possible data modalities and frameworks;
- Interpretability: for a learned representation to be effective, it should be mappable to the corresponding features of the original data. Such correlability ensure that the representation is useful for transparency in downstream tasks, where understanding how different characteristics of the data impacts on the result might be crucial;
- Separability: this can be simply understood as a characteristic of the learning process which partitions the resulting representation in the same way as is intrinsic to the original data. In this sense, separability is a necessary but insufficient condition for interpretability, but still stands on its own as an essential feature of representation learning.

In recent years, there has been active cross-fertilization of ideas between the fields of machine learning and neuroscience, significantly enriching research in both domains. The ideas relating to representation learning as elucidated above have one-to-one counterparts in the field of neuroscience, from which machine learning has borrowed a number of important concepts. In fact, all three points mentioned above are strict requirements for many neuroscientific tasks, e.g., brain-computer interfaces [4] [5], or theoretical neuroscience [6], especially in the quest to understand the origins of innately neural processes, such as intention and decision-making. In particular, for many applications of BCIs, it is important to ensure that the learned representations is able to robustly produce expected results and actions.

A key element in this workflow is the encoding of EEG signals, which involves transforming the raw neural data into a format that can be accurately interpreted by machines. This encoding ensures precise correspondence between a user's neural intent and the external action performed by the system. For example, accurate encoding models can distinguish between various brain activities, enabling the system to respond to specific intentions or commands. In this work, we take a further step in this direction, to ensure that machine-learned representations of neural activity, as manifest in the form of model parameters, can be robustly and uniquely mapped to a movement associated to a specific imagery. Given information on this mapping, we can hope to leverage model parameters as a reliable representation of particular neural states, and subsequently harness this representation as

a surrogate to characterize these states. In particular, a deep model-based surrogate is important in the context of modern data-driven analytic methods, since we could hope to transfer this representation across different learning models, e.g., as prior distributions to aid in transfer learning. Finally, the construction of this mapping is a necessary step in setting up an interpretable neural-machine integration workflow, which is a current inadequacy in the larger field of BCIs.

In this paper, we used our unified EEG encoding framework (TFSICNet), which integrates the connectomic, hierarchical, and network properties of the brain. Our framework is designed to be versatile, allowing it to encode EEG signals for a range of tasks, such as motor imagery, visual signal decoding, and classification. We apply established methods in the field of machine explainability to construct the link between neural input data and learned machine representations. In particular, we compute the following metrics:

- Evoked [7] data: the Evoked data type is a container for EEG data that has been averaged over multiple epochs (equal length segments of data extracted around stimulus events or responses, e.g., in our case, the actual motor imagery events). Visualizations of the Evoked datatype is then expected to show consistent features, which is a clear marker of stable neural activity;
- Principal Component Analysis (PCA): PCA projects datapoints onto a set of *principal components* that order the data along the axes with the most variation. In our work, PCA allows us to compare the principal component axes of both the motor imagery activity and the model parameters, therefore giving strong correlational information between the two;
- Representational Similarity Analysis (RSA): RSA is a computational method which was specifically designed for the analysis of correlation matrices across different input modalities. In our work, we compute the RDA matrix which shows the correlation between neural activity and attention model parameters;
- Time-Frequency Analysis(TFA): This method examines the relationship between neural activity and attention model parameters by analyzing the spectral properties of signals and the corresponding attention weights. It highlights how different frequency components contribute to the model's attention mechanism.

The above-mentioned methods enable the computation of attentional parameter importances which we correlate with neural activity, hence establishing a correspondence between model-learned representations and actual neural structure.

Our key contributions in this work are:

- We propose a deep encoding model that captures the latent information within EEG signals while incorporating key structural and functional characteristics of the brain, enabling its application to diverse visual, motor, and BCI tasks.
- We show how to connect learned representations from our model with established interpretability methods in the BCI field. Our model shows strong correlative results with raw neural activity data;
- We will open source our code to reproduce the results.

This approach not only enhances the quality of signal decoding but also improves the overall reliability and performance of BCI systems. By unifying different embedding models under a single, general framework, we ensure that the encoding process is both flexible and precise, making it suitable for a wide array of downstream applications.

## II. Related Work

BCIs have evolved since Hans Berger's first EEG recordings [8], leading to applications like the P300 speller [9]. Modern BCIs use EEG signals for device control, often through motor imagery [10]. Early BCIs focused on classification, but recent work emphasizes encoding EEG signals for flexible downstream tasks [11], [12]. Deep learning's ability to handle multimodal, high-dimensional data has made it a popular choice for EEG analysis [13]. This flexibility enables advanced encoding strategies across time, frequency, and space domains [14]. Decoding—reconstructing stimuli from brain signals—has also advanced through deep learning, particularly with generative models like diffusion models [15], [16]. These approaches enhance BCI performance, driving applications in "mind-reading" technologies [17].

Analysis of EEG signals, in particular in the context of its correlation with external stimuli, is an important field of research in the BCI community. RSA is the most widely employed tool in this context [18], [19] due to its flexibility in handling cross-modal data, as well as its general robustness [20]. In our work we compute the RSA matrices between raw EEG data and attention coefficients, as learned via the TFSICNet.

## III. Methods

### A. Dataset And Data Preprocessing

We evaluated our proposed encoder on four datasets, including three motor imagery datasets and one object recognition dataset. Each dataset was preprocessed to extract relevant EEG signal segments, which were then fed into our encoder for further analysis:

- BNCI 2014-001 Motor Imagery Dataset: This dataset includes EEG data from 9 subjects performing four motor imagery tasks: imagining movement of the left hand, right hand, both feet, and tongue.
- BNCI 2015-001 Motor Imagery Dataset: In this dataset, subjects perform motor imagery tasks involving sustained left hand versus both feet movement.
- BNCI 2015-004 Motor Imagery Dataset: Users performed, follow- ing a cue-guided experimental paradigm, five distinct mental tasks (MT). MTs include mental word association , mental subtraction, spatial navigation, right hand motor imagery and feet motor imagery.

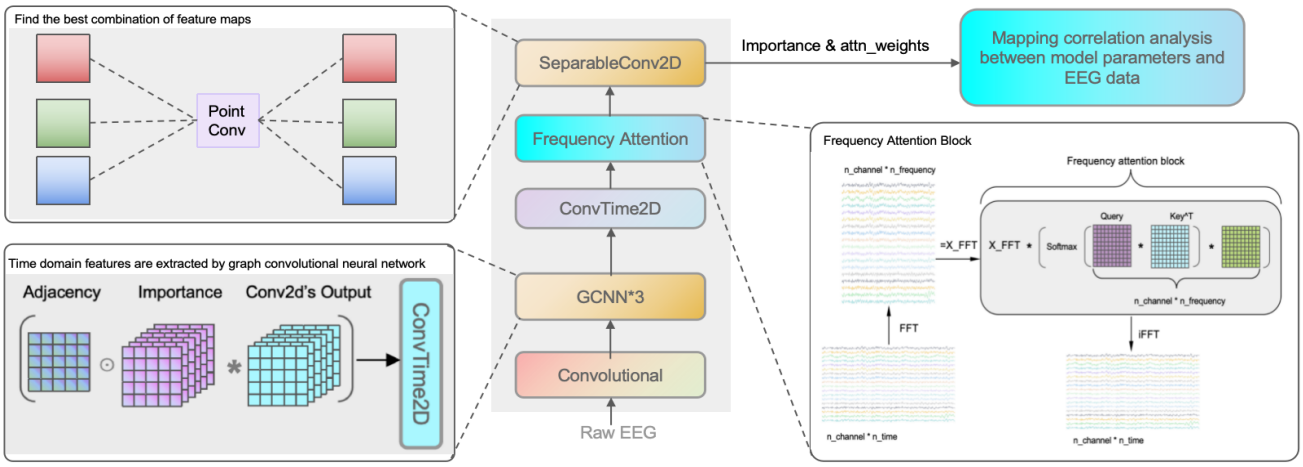



Fig. 1: Our proposed model TFSICNet, consisting of the following modules: frequency attention, graph convolution and convolution

### B. Encoder

The proposed deep encoder, TFSICNet, is composed of four key blocks designed to extract critical features from EEG data by leveraging brain structure and functional connectivity. Each block contributes to encoding different aspects of the data, ensuring that the full range of information is captured for downstream tasks. The workflow of TFSICNet show in Figure 1.

- Frequency Attention (FA) Block: The FA block enhances feature representation in the frequency domain by applying attention mechanisms, i.e., the attention mechanism then computes weights to emphasize the most informative frequency components for the frequency-transformed EEG data.
- Graph Convolution (GC) Block: The GC block captures connectomic information by constructing a graph-based representation of the EEG channels. It creates an adjacency matrix based on the Pearson correlation between EEG channels, treating the channels as nodes and their correlations as edges.
- Convolution (CV) Block: The convolution block encodes spatiotemporal features through sequential convolutions, reflecting the hierarchical organization of brain signals. It applies a convolution kernel across EEG channels to capture inter-channel relationships, followed by a temporal convolution that extracts sequential features from the EEG signals.

The combined output of these blocks forms an encoded vector for downstream tasks. The FA block processes the input in the frequency domain, applying attention, and the GC block builds a graph-based representation of brain connectivity. The convolution then encode spatiotemporal information before producing the final encoding vector.

### C. Multi-dimensional correlation and interpretability analysis

We investigate the high correlation between TFSICNet and raw EEG data using multiple analytical methods, including Representational Similarity Analysis (RSA), Time-Frequency Analysis (TFA), Principal Component Analysis (PCA), and Evoked Potential visualization. These analyses comprehensively validate the high interpretability of TFSICNet by demonstrating its ability to align closely with the underlying neural activity patterns captured in the EEG data.

## IV. Results

### A. Visualization of the Evoked data

From the different variants of Evoked visualizations, we present the scalp topographies, which highlights the actual EEG channels that contribute the most to a specific motor imagery. We then compare these topography plots with the weights distribution from our GCNN feature extractor network. Figure. 2 shows the overlapped channel and weight distribution plots. It is clear that there are significant correlations between the two plots. In both figures, the closer the brain electrode is to the bottom of the topology, the higher the energy or the greater the weight, showing that channel-wise neural activity corresponding to specific motor imageries can be reproducibly correlated with distribution of model parameters. From this result, we argue that the parameters of our learned model can be reproducibly associated to actual neural information, in the form of intensity of neural activity and electrode placement topology.

### B. Visualization of the representational dissimilarity matrix

We also compute and present the Representational Dissimilarity Matrix (RDM) [18] between raw neural activity data and the attention parameters. Figure. 3 shows the corresponding matrices for both these modalities. The dimensionality of all three matrices corresponds to the number of different motor imagery actions. It can be seen from the same RDM that the

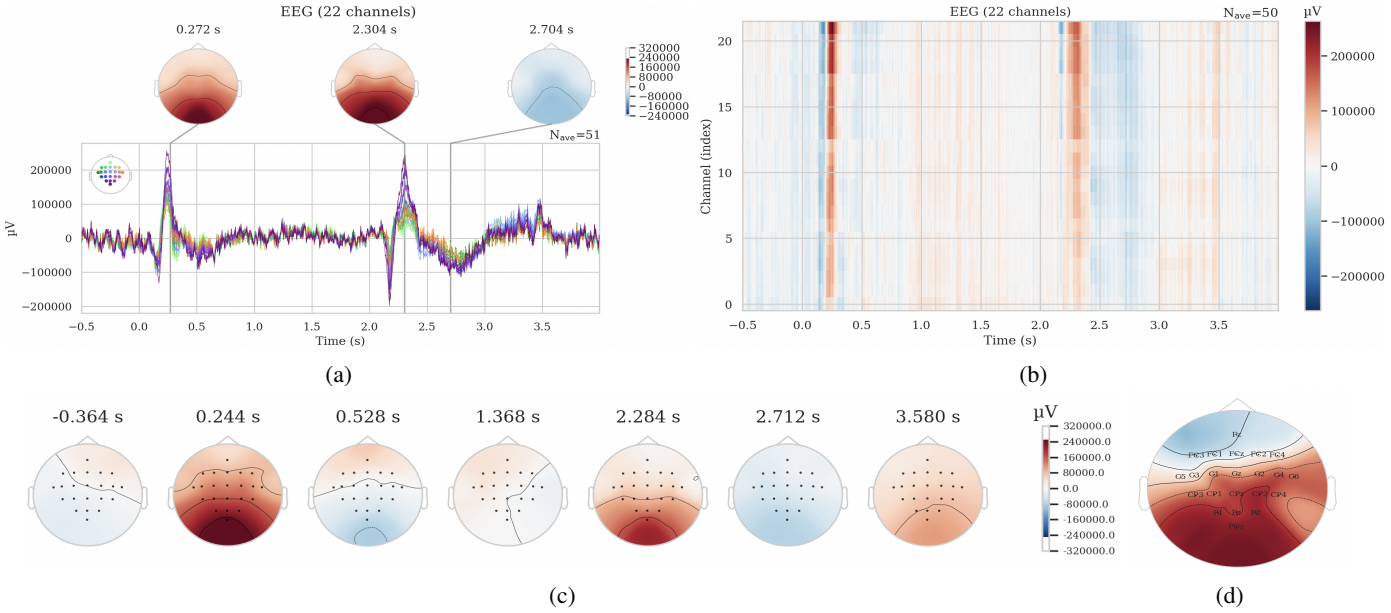



Fig. 2: (a) Visualization of evoked data calculated on the BNCI2014_001 dataset. (b) The picture evoked data in the form of two-dimensional pictures. (c) The picture shows scalp region topologies to evoked data at different time points. (d) The Importance matrix weights in GCNN trained in the BNCI2014_001 dataset are mapped to the scalp region topology.

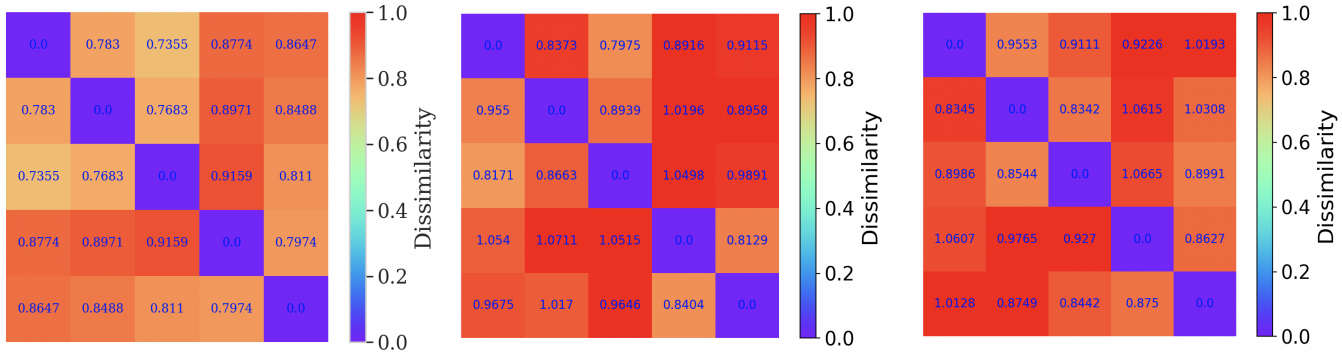
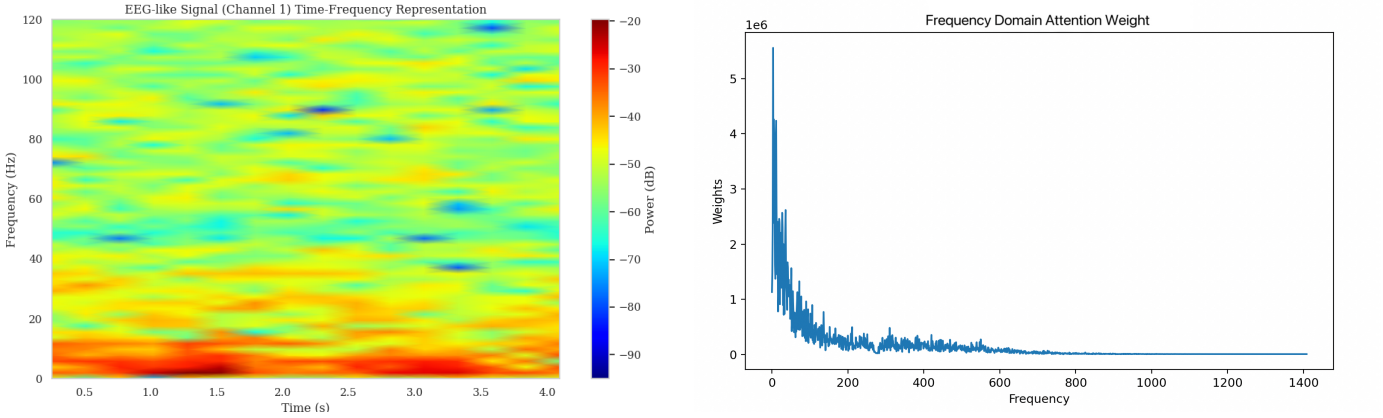



Fig. 3: (a) The RDM of the BNCI2015_004 data set is shown. (b) The picture shows the RDM of attn_weights matrix weights in the frequency domain attention module extracted in the case of BNCI2015_004 data set. (c) The figure shows the RDM for training the Importance matrix weights in the extracted GCNN in the case of the BNCI2015_004 dataset.

dissimilarity values of different action stimuli are significant. We conclude that the attn_weights in our model and the weights in the importance matrix have learned such dissimilar features. In addition, it can be seen that the three are very similar among different RDM. The averaged representational dissimilarity score between them is **0.0389** between Fig. (a) and Fig. (b), and **0.0447** between Fig. (a) and Fig. (c), calculated by using the $p$ value of Pearson correlation coefficient. The averaged score between Fig. (b) and Fig. (c) is **0.1268**. The the $p$-value computed for the Pearson correlation between Figs. (a) and (b) implies statistically significant correlation between the raw neural data and the attentional weights *within the same set of motor imagery class*. This result is a further validation of our claim of model interpretability for BCI, specifically motor imagery tasks.

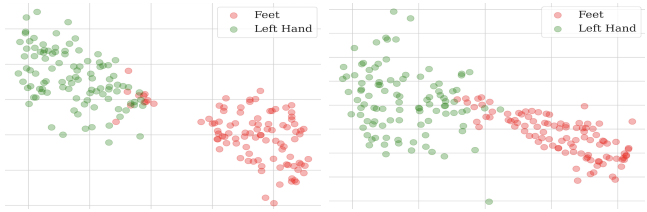
### C. *Time frequency diagram interpretability analysis*

Fig. 4(a) shows the time-frequency representation of EEG single-channel signals during the subject's left-hand motor imagery task. This figure, obtained through time-frequency analysis, reveals the distribution of signal power over time and frequency. It is observable that in the low-frequency range (particularly below 20 Hz), the signal power significantly increases, indicating that this frequency band may be associated with neural activity related to the motor imagery task. In terms of representation learning, the concentration of weights at this frequency band indicates that a well-trained learner should be able to clearly distinguish information contained




(a) (b)

Fig. 4: (a) BNCI2014_001 dataset EEG signal channel 1 time-frequency diagram. (b) BNCI2014_001 dataset frequency domain weighted attention matrix column-wise addition line chart.




(a) (b)

Fig. 5: (a) In the case of BNCI2015_001, the visualization results of extracting the trained GCNN model parameters and performing PCA dimensionality reduction for two categories are shown. (b) The visualization results of extracting the attn_weights matrix in the frequency domain attention module and performing PCA dimensionality reduction for two categories in the case of BNCI2015_001.

in this band from those showing more pronounced weighting at other frequency bands. In the case of our model, we can clearly see signs of this interpretability in Fig. 4(b). This figure illustrates frequency components' weights assigned by the frequency-domain attention mechanism in the deep learning model for the same task. It is evident that the model allocates higher weights in the low-frequency band, consistent with the enhanced power observed in Figure 4(a). This consistency in the frequency domain suggests that the model successfully captures the critical neural oscillatory features of the motor imagery task, further validating its biological interpretability and rationality.

### *D. PCA dimension reduction and identification of principal axes*

Figure. 5 shows a comparison between PCA-projected parameters for the GCNN and attention model parameteres, respectively. We see that the distribution of the parameters relative to the principal axes are oriented similarly between the plots, indicating that their relative distribution in the PC space is homogeneous. The homegeneity between the two distributions imply the same sources of variance within the respective data- and parameter sets, indicating a high degree of correlative information which can be extracted. Their similarity indicates that both the GCNN and attention modules have learned a common representation for the data, and both the parameters of GCNN and the parameters of the frequency domain attention module can effectively separate different imaginary actions.

## V. Discussion

In this work, by incorporating the structural and functional characteristics of the brain, we have shown that TFSICNet is capable of learning unique and robust representations of motor imagery and other brain states. Our multi-faceted interpretability analysis, including Representational Similarity Analysis (RSA), Principal Component Analysis (PCA), and Time-Frequency Analysis (TFA), provides substantial evidence for the model's ability to align closely with raw neural data, confirming its interpretability.

One of the primary contributions of this study is demonstrating the high correlation between TFSICNet's learned representations and the actual neural activity as captured by EEG signals. The use of Evoked Potential visualizations, PCA, and RSA matrices has provided strong evidence that the model can effectively learn and map neural activity to specific motor imagery tasks. The consistency observed between the model's attention weights and neural oscillatory patterns in the time-frequency domain further supports the idea that TFSICNet has successfully captured biologically meaningful features. These results suggest that TFSICNet can be used as a reliable framework for decoding neural signals, with potential applications in real-time BCI systems.

However, there are certain limitations to our approach that must be acknowledged. While the model demonstrates strong interpretability with respect to motor imagery tasks, the generalizability of the approach to other types of neural signals,

such as those associated with different cognitive states or complex motor tasks, needs further investigation. Additionally, although the results are promising, the model's performance could potentially be influenced by the quality and diversity of the EEG data used for training. Future studies could explore the incorporation of multimodal data sources, such as fNIRS or MEG, to further enhance the model's ability to capture diverse neural activity patterns.

Moreover, the computational efficiency of TFSICNet, especially with larger datasets, remains an area for optimization. While the current framework performs well within the scope of the data evaluated in this work, real-world BCI applications often require real-time processing of complex signals. Therefore, developing more efficient architectures, potentially through model pruning or hardware acceleration, could make TFSICNet more suitable for practical use.

## VI. Conclusion

In conclusion, this study has demonstrated that TFSICNet provides a robust and interpretable framework for encoding and analyzing EEG signals. We have designed an interpretable deep learning model based on the Fourier attention mechanism, which effectively learns a unique and distinct representation of motor imagery data. By leveraging metrics such as Evoked potential, PCA, and Representational Dissimilarity Matrix (RDM), we have shown that our model's learned representations exhibit a high degree of correspondence with raw neural activity, confirming its high interpretability. The strong correlation between the model parameters and EEG data underscores the potential of TFSICNet in advancing the performance and reliability of brain-computer interface systems.

Looking ahead, an exciting direction for future research would be to explore the application of TFSICNet in closed-loop BCI systems. By combining the model's interpretability with its ability to accurately decode neural activity, we could pave the way for more intuitive and effective brain-machine interfaces that offer users greater control and interactivity. The integration of real-time feedback into such systems could further enhance user performance and experience, opening up new possibilities for BCI applications in rehabilitation, gaming, and beyond.

## References


[1] Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee, "High-fidelity synthesis with disentangled representation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 157–174.

[2] Kazi Nazmul Haque, Rajib Rana, and Björn W Schuller, "High-fidelity audio generation and representation learning with guided adversarial autoencoder," *IEEE Access*, vol. 8, pp. 223509–223528, 2020.

[3] Florian Bordes, Randall Balestriero, and Pascal Vincent, "High fidelity visualization of what your self-supervised representation knows about," *arXiv preprint arXiv:2112.09164*, 2021.

[4] Klaus-Robert Müller, Matthias Krauledat, Guido Dornhege, Gabriel Curio, and Benjamin Blankertz, "Machine learning techniques for brain-computer interfaces," *Biomed. Tech*, vol. 49, no. 1, pp. 11–22, 2004.

[5] Zhihan Lv, Liang Qiao, Qingjun Wang, and Francesco Piccialli, "Advanced machine-learning methods for brain-computer interfacing," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 5, pp. 1688–1698, 2020.

[6] Mai-Anh T Vu, Tülay Adalı, Demba Ba, György Buzsáki, David Carlson, Katherine Heller, Conor Liston, Cynthia Rudin, Vikaas S Sohal, Alik S Widge, et al., "A shared vision for machine learning in neuroscience," *Journal of Neuroscience*, vol. 38, no. 7, pp. 1601–1607, 2018.

[7] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, no. 267, pp. 1–13, 2013.

[8] Hans Berger, "Über das elektrenkephalogramm des menschen," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 87, no. 1, pp. 527–570, 1929.

[9] L. A. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988.

[10] Gert Pfurtscheller, Doris Flotzinger, and J. Kalcher, "Brain-computer interface: a new communication device for handicapped persons," *Journal of Microcomputer Applications*, vol. 16, pp. 293–299, 1993.

[11] Evan Hernandez and Jacob Andreas, "The low-dimensional linear geometry of contextualized word representations," *arXiv preprint arXiv:2105.07109*, 2021.

[12] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, no. 1, pp. 1–7, 2016.

[13] Dana Lahat, Tülay Adali, and Christian Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[14] Behnaz Ghoraani and Sridhar Krishnan, "Time–frequency matrix feature extraction and classification of environmental audio signals," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2197–2209, 2011.

[15] Qiongyi Zhou, Changde Du, Dan Li, Haibao Wang, Jian K Liu, and Huiguang He, "Neural encoding and decoding with a flow-based invertible generative model," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[16] Yu Takagi and Shinji Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14453–14463.

[17] Anu Realo, Jüri Allik, Aire Nõlvak, Raivo Valk, Tuuli Ruus, Monika Schmidt, and Tiina Eilola, "Mind-reading ability: Beliefs and performance," *Journal of Research in Personality*, vol. 37, no. 5, pp. 420–445, 2003.

[18] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini, "Representational similarity analysis - connecting the branches of systems neuroscience," *Frontiers in Systems Neuroscience*, vol. 2, 2008.

[19] Nikolaus Kriegeskorte and Rogier A. Kievit, "Representational geometry: integrating cognition, computation, and the brain," *Trends in Cognitive Sciences*, vol. 17, no. 8, pp. 401–412, 2013.

[20] H Nili, C Wingfield, A Walther, L Su, W Marslen-Wilson, and N Kriegeskorte, "A toolbox for representational similarity analysis," *PLoS Comput Biol*, vol. 10, no. 4, pp. e1003553, 2014.