Published: 25 August 2024

KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 25 - 29, 2024
Barcelona, Spain

Conference Sponsors:
SIGMOD
SIGKDD

RESEARCH-ARTICLE

EEG2Rep: Enhancing Self-supervised EEG Representation Through Informative Masked Inputs

NAVID MOHAMMADI FOUMANI, Monash University, Melbourne, VIC, Australia

GEOFFREY MACKELLAR

SOHEILA GHANE

SAAD IRTZA

NAM NGUYEN

MAHSA SALEHI, Monash University, Melbourne, VIC, Australia

Open Access Support provided by:

Monash University

KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (August 2024)
https://doi.org/10.1145/3637528.3671600
ISBN: 9798400704901

EEG2Rep: Enhancing Self-supervised EEG Representation Through Informative Masked Inputs

Navid Mohammadi Foumani*
Monash University
Melbourne, Australia
navid.foumani@monash.edu

Geoffrey Mackellar
Emotiv Research
Sydney, Australia
geoff@emotiv.com

Soheila Ghane
Emotiv Research
Melbourne, Australia
soheila@emotiv.com

Saad Irtza
Emotiv Research
Sydney, Australia
saad@emotiv.com

Nam Nguyen
Emotiv Research
Sydney, Australia
namnguyen@emotiv.com

Mahsa Salehi
Monash University
Melbourne, Australia
mahsa.salehi@monash.edu

ABSTRACT

Self-supervised approaches for electroencephalography (EEG) representation learning face three specific challenges inherent to EEG data: (1) The low signal-to-noise ratio which challenges the quality of the representation learned, (2) The wide range of amplitudes from very small to relatively large due to factors such as the inter-subject variability, risks the models to be dominated by higher amplitude ranges, and (3) The absence of explicit segmentation in the continuous-valued sequences which can result in less informative representations. To address these challenges, we introduce *EEG2Rep*, a self-prediction approach for self-supervised representation learning from EEG. Two core novel components of EEG2Rep are as follows: 1) Instead of learning to predict the masked input from raw EEG, EEG2Rep learns to predict masked input in latent representation space, and 2) Instead of conventional masking methods, EEG2Rep uses a new semantic subsequence preserving (SSP) method which provides informative masked inputs to guide EEG2Rep to generate rich semantic representations. In experiments on 6 diverse EEG tasks with subject variability, EEG2Rep significantly outperforms state-of-the-art methods. We show that our semantic subsequence preserving improves the existing masking methods in self-prediction literature and find that preserving 50% of EEG recordings will result in the most accurate results on all 6 tasks on average. Finally, we show that EEG2Rep is robust to noise addressing a significant challenge that exists in EEG data. Models and code are available at:https://github.com/Navidfoumani/EEG2Rep

CCS CONCEPTS

• **Applied computing → Health care information systems**; • **Computing methodologies → Machine learning**.

KEYWORDS

EEG Representation Learning, EEG self-supervised Learning, EEG Masking, EEG Classification

KDD '24, August 25–29, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3671600

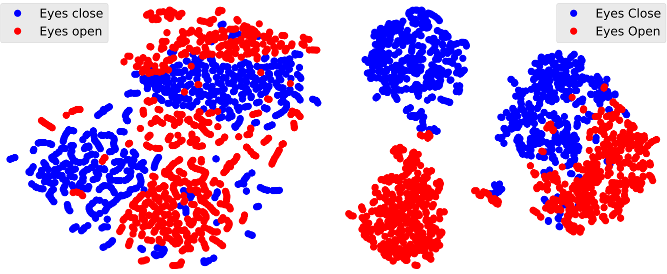
ACM Reference Format:
Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. 2024. EEG2Rep: Enhancing Self-supervised EEG Representation Through Informative Masked Inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671600

1 INTRODUCTION

An electroencephalogram (EEG) is a noninvasive method that captures brain data by placing electrodes on the patient's scalp surface, enabling the recording of electrical activity within the brain [1]. This specialized and complex biological electrical signal serves as a reflection of the brain's functional state, providing insights into the individual's mental condition [2]. From this data, valuable information can be extracted, including vital signs that facilitate continuous monitoring of the patient's health [3]. Additionally, EEG plays a crucial role in diagnosing and identifying various brain conditions, finding applications in diverse healthcare domains such as sleep medicine, neurological disorders, cardiovascular disease detection, and activity monitoring [4–6].

In the past decade, the integration of deep learning into biomedical research has experienced significant growth, demonstrating its ability to frequently outperform conventional machine learning methods across various tasks [4, 6–9]. However, the training of deep learning models for biomedical applications requires substantial amounts of data, annotated by experts, whose collection is often time and cost-prohibitive. Furthermore, deeper neural networks are susceptible to overfitting the EEG data, particularly in the presence of inter-subject variability [10, 11]. Self-supervised learning (SSL) has emerged as a prominent solution for such problems, as it allows learning powerful representations from vast unlabeled data by producing supervisory signals directly from the data [9, 10, 12, 13].

In EEG analysis, two commonly used self-supervised learning approaches are invariance-based methods and self-prediction methods [9, 14]. In invariance-based methods, the objective is to optimize an encoder to generate similar embeddings for two or more views of the same EEG time series [14]. These different views of the time series are usually crafted using a set of manual data augmentations, including techniques like jittering, permutation, and scaling [15, 16]. The main idea behind self-prediction methods is to remove or corrupt parts of the input and train the model to predict or reconstruct



(a) MAEEG (Acc:75.21%) **(b) EEG2Rep** (Acc: 81.66%)

Figure 1: Comparison of 2D t-SNE plots for representation learned by (a) MAEEG and (b) EEG2Rep on the Crowdsourced EEG dataset.

the altered content [14]. For example, approaches like Masked AutoEncoders (MAE) [17] learn representations by reconstructing randomly masked patches from an input [10, 17, 18].

Invariance-based pretraining methods can construct representations of high-level semantics by capturing essential features consistent across various data views. This strategy enables the model to identify and prioritize features important for understanding the underlying semantics of the data, as these features remain unchanged despite variations in the input. However, we believe that invariance-based methods perform well in computer vision and natural language processing due to the strong constraints present in image and text data. For example, the success with images arises from tasks related to object interpretation, where transformations such as scaling, blurring, and rotation assume that the resulting images will be similar to those generated in the original scenario with changes in camera zoom, stability, focus, or angle. However, there do not appear to be equivalent transformations that can be applied to EEG data. Augmentation methods applied to EEG data can inadvertently modify the semantic meaning, underlying distribution, and class representation of the signals [19, 20]. For example, augmentation may introduce synthetic patterns or artifacts that do not align with genuine brain activity.

Moreover, invariance-based methods may introduce significant biases, potentially hindering specific downstream tasks or even pretraining tasks with diverse data distributions. The generalization of these biases for tasks requiring varying levels of abstraction often remains unclear. Distinct data augmentation strategies may result in misinterpretation or misclassification by the model. For example, in scenarios such as sleep stage classification (involving low-frequency bands) and emotion recognition (involving high-frequency bands), identical augmentations may not be suitable for both cases.

In contrast to invariance-based methods, self-prediction pretraining tasks demand less prior knowledge and demonstrate ease of generalization across diverse downstream tasks [21, 22]. However, self-prediction pretraining faces unique challenges when applied to EEG data [10, 12] which makes it ineffective. Fig. 1a depicts the visualization of EEG representations by a state-of-the-art self-prediction EEG pretraining model, namely MAEEG [12], on Crowdsourced dataset [23] with two classes. The two classes are not easily separable in the learned representation space, resulting in low classification accuracy. Here, we outline the three main challenges that exist in EEG data:

- **Challenge 1:** The recorded EEG is invariably contaminated with noise, impacting the reconstruction loss function and potentially introducing significant errors. Even accurate predictions may yield high errors due to the pronounced impact of noise on the loss function.
- **Challenge 2:** EEG data has a wide range of amplitude values, which can be due to the variability between different subjects or the variability of electrode placement. The substantial ranges make it particularly challenging to reconstruct accurate values during the reconstruction process.
- **Challenge 3:** EEG signals differ from text and images in that there is no explicit segmentation for EEG data, as they are continuous-valued sequences. Hence, the implementation of a masking strategy becomes essential for a model to have sufficient enough information for the reconstruction of the masked EEG.

To address the challenges mentioned above, we introduce EEG2Rep to enhance the semantic quality of EEG representations without relying on prior knowledge about downstream tasks. In contrast to the existing self-prediction methods that learn EEG representations by reconstructing the raw EEG data space [10, 12], EEG2Rep is trained to reconstruct more abstract features of EEG data in the latent space. Such approaches have shown to be effective in image and text representation learning [22, 24, 25]. Our motivation is that the existing noise in EEG is less likely to remain in the abstract features of EEG, and by learning to reconstruct the abstract features we potentially eliminate the unnecessary noise that exists in raw EEG data (addressing challenge 1). Additionally, as the abstract features are normalized within the representation space, the reconstruction of these features becomes more straightforward compared to reconstructing the potentially high amplitude value range of raw EEG data (addressing challenge 2). Finally, EEG2Rep leverages our novel *Semantic Subsequence Preserving* method to ensure that the context has sufficiently meaningful and rich semantic information (addressing challenge 3).

Fig. 1b displays the visualization of EEG representations learned by EEG2Rep. The two classes in this figure are easily separable which highlights the effectiveness of EEG2Rep in enhancing EEG representations and, consequently, improving classification accuracy. Another core component of EEG2Rep is its efficient multi-masking design. Specifically, we reuse the same target representation for various masked versions of each sample. Additionally, we predict the representation of various target blocks for a single masked input to improve efficiency further.

This work follows from the project with Emotiv Research [1], a bioinformatics research company based in Australia, and Emotiv, a global technology company specializing in the development and manufacturing of wearable EEG products. In our prior work, we looked at detecting distraction episodes in drivers by analyzing their brain EEG as a case study. One significant challenge we encountered was the presence of noise in the recorded EEG datasets and the inability of current supervised detection models to learn patterns of distraction within the presence of noise. This paper addresses this issue, along with the two additional challenges mentioned above.

[1] www.emotiv.com/research

2 METHODOLOGY OF EEG2REP

2.1 Problem Definition

Our goal is to address the problem of learning a nonlinear embedding function that can effectively map each EEG sample $X_i = \{x_1, x_2, \ldots, x_L\}$ from a given dataset D into a concise and meaningful representation $R_i \in \mathbb{R}^{d_e}$, where d_e indicates the desired representation dimension. The EEG dataset D consists of n samples, denoted as $D = \{X_1, X_2, ..., X_n\}$, with each X_i representing a C-channel EEG sequence of length L. To evaluate the quality of our learned representation $\mathbf{R} = \{R_1, R_2, \ldots, R_n\}$, we examine two scenarios based on the availability of labeled data: i) *Linear Probing*: We first pre-train a model without labels through a self-supervised pretext task. Upon completing the pre-training phase, we freeze the encoder and add a linear classifier on top of the pre-trained model's output or intermediate representations. The linear classifier is then trained on a downstream task, typically a classification task, utilizing the pre-trained representations as inputs, and ii) *Fine-Tuning*: Initially, we pre-train a model without labels through a self-supervised pretext task. Next, we perform fine-tuning by training the entire model for a few epochs using a labeled dataset in a fully supervised manner.

2.2 Model architecture

Illustrated in Fig. 2, the EEG2Rep model is introduced with the main goal of predicting the representation of a given EEG sample based on a masked view of the same EEG input. We now explain each component of this architecture separately in the following subsections.

Input Embedding. Building upon established works [11, 15, 20, 26–28], we adopt a 3-layer convolutional neural network as input embedding to convert the raw EEG data into patches. Specifically, we feed EEG sample X_i to the first layer that incorporates a depth-wise convolutional layer, specifically designed to capture the spatial correlations between channels [11, 27]. This is succeeded by a linear spatial filter to amplify the signal-to-noise ratio. The spatial filters leverage the fact that neural signals exhibit specific spatial patterns across the scalp, while noise sources may manifest more random spatial patterns [26].

Following the spatial filter, we integrate max pooling and spatial padding to ensure translation equivalence and effectively address edge effects. The output of this network is a set of EEG patches $\hat{S}_{\mathbf{x}} = \{\hat{S}_{x_1}, ..., \hat{S}_{x_l}\}$ where $\hat{S}_{x_i} \in \mathbb{R}^{d_x}$, and d_x is the embedding dimension, and l is the number of patches. Lastly, for every patch $\hat{S}_{x_i}$, the positional embedding feature of the i^{th} position is added to it, resulting in the EEG patches $S_{\mathbf{x}} = \{S_{x_1}, ..., S_{x_l}\}$ (shown in blue in Fig. 2). Please note for each EEG sequence X_i, we will have a set of EEG patches $S^i_{\mathbf{x}}$ as the output of the input embedding network. However, for simplicity, we will drop the superscript i from $S^i_{\mathbf{x}}$ and use $S_{\mathbf{x}}$ as an input to the following components in the EEG2Rep architecture.

Context-driven target prediction. Previous self-prediction methods for EEG data have primarily followed the approach of Masked Autoencoders (MAE) [17], reconstructing local windows of the raw input EEG [12], or have adopted a BERT-like method [29], predicting discrete representations [10]. However, the resulting representations often demonstrate lower semantic quality than invariance-based methods during off-the-shelf evaluations, such as linear probing, due to the intrinsic characteristics of EEG data mentioned earlier. Low signal-to-noise ratio in EEG challenges the reconstruction task and a wide range of amplitudes in EEG further complicates the optimization process.

To enhance the semantic depth of self-supervised representation learning, we introduce the concept of *context-driven target prediction* for EEG data. In this approach, the model is trained to predict the representations of the original unmasked training data based on an encoding derived from the masked sample in abstract representation space. Compared to self-prediction methods that predict in raw space, EEG2Rep uses abstract prediction targets for which unnecessary raw-level details or noise are potentially eliminated, thereby leading the model to learn richer semantic features.

EEG2Rep comprises three main components: *Target network* uses the complete (unmasked) input embedding $S_{\mathbf{x}}$ to generate semantically rich representations, allowing each patch to encode knowledge of all others through its self-attention architecture. The resulting output serves as the target for our learning task. *Context network* shares the architecture with the target network, differing only in its use of the masked version of the input embedding to generate representations for visible patches. Our innovative semantic subsequence preserving method ensures that the context network's output contains sufficient semantic information for reconstruction purposes. We use a standard transformer [29, 30] for both target and context networks. *Predictor network* leverages the output of the context network and randomly chosen masked tokens to regress the targets, i.e., the output of the target network. We enhance contextualized information integration for the predictor network by incorporating cross-attention-based transformers [29]. Below, we elaborate on the process of creating the target, context, and predictor networks for the training task, providing clarification on how these networks are parameterized.

Target Network

Given an EEG embedding $S_{\mathbf{x}} = \{S_{x_1}, \ldots, S_{x_l}\}$, we feed it through the target-encoder $f_{\bar{\theta}}$ to obtain a corresponding patch-level representation $\mathbf{y} = \{y_1, \ldots, y_l\}$:

$$\mathbf{y} = f_{\bar{\theta}}(S_{\mathrm{x}}) \tag{1}$$

where $y_i \in \mathbf{R}^{d_e}$ is the representation associated with the S_{x_i} patch and d_e is the transformer's embedding dimension (shown in green in Fig. 2). We apply normalization to these representations, preventing the model from collapsing into a constant representation for all time steps and ensuring that values with high amplitude do not dominate the target features.

To obtain the targets for our reconstruction loss, we randomly sample M blocks from the target representation $\mathbf{y}$. B_i is the i^{th} masked block in $\mathbf{y}$ where $i \in \{1, 2, \ldots, M\}$ and $\mathbf{y}(i) = \{y_j\}_{j \in B_i}$ (e.g., B_1 and B_2 is annotated in red in Fig. 2 and $\mathbf{y(1)}$ contains all $\{y_j\}_{j \in B_1}$). Note that the target blocks are chosen from the output of the target encoder, not the input of the target encoder. This distinction is crucial to ensure each target representation retains a high semantic level as it encodes the knowledge of all input patches through the self-attention mechanism in Transformers architecture [21, 22, 24].

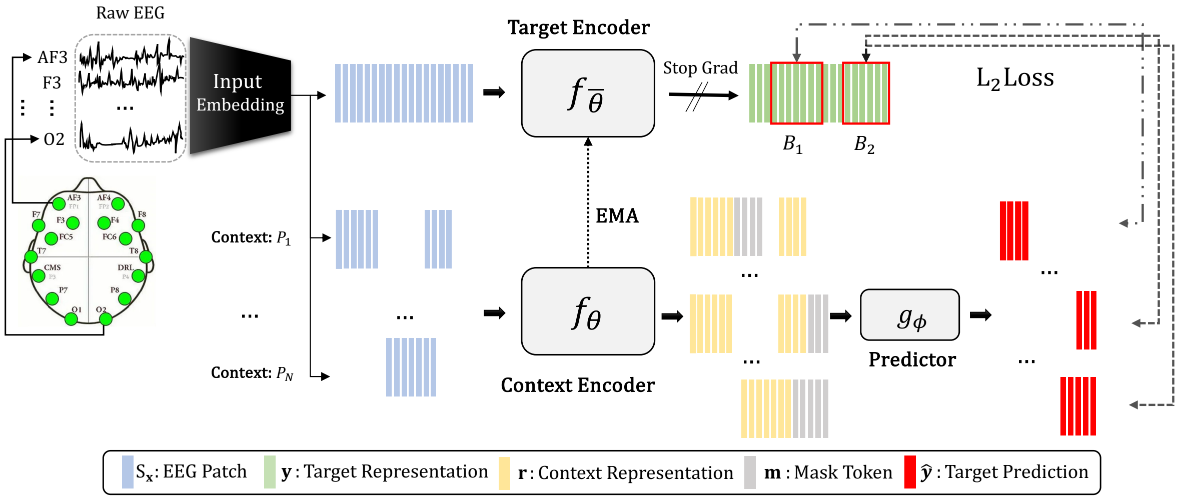


Figure 2: Architecture of EEG2Rep

Context Network

To obtain the context in EEG2Rep, we initially sample N contexts from the EEG patches $S_{\mathbf{x}}$ (refer to Sec.2.3 for details on our novel context selection). Only visible context patches are processed via context-encoder model f_θ to obtain context representations:

$$\mathbf{r}(q) = f_\theta(P_q) = f_\theta(\{S_{x_t}\}_{t\in P_q}) \tag{2}$$

Where P_q correspond to the q^{th} context which contains a subset of EEG patches from $S_{\mathbf{x}}$ and $\mathbf{r}(q) = \{r_t\}_{t\in P_q}$ is its patch level representation (shown in yellow in Fig. 2). Since the target blocks are sampled independently from the contexts, there may be significant overlap. We exclude any overlapping regions from the target block in the loss calculation to ensure a non-trivial prediction task. Examples of various contexts (P_1 and P_N) and target blocks (B_1 and B_2) are illustrated in Fig. 2. The remaining patches are called masked tokens (shown in grey in Fig. 2).

Predictor Network

For the Predictor network, we use a 4-layer cross-attention transformer to enhance the effective correlation among the context representation and masked tokens. The "cross-attention" aspect of the transformer is capable of blending two distinct embedding sequences, which may vary in length and originate from different sources [30]. In this setup, the masked token serves as a query, while the key and values are sourced from the context encoder.

Given the output of the context encoder $\mathbf{r}(q)$, our goal is to predict M target block representations $\{\mathbf{y}(1), \ldots, \mathbf{y}(M)\}$. For each target block $\mathbf{y}(i)$, the predictor g_ϕ takes as input the output of the context encoder $\mathbf{r}(q)$ and a mask token $m(i)$ for each patch we wish to predict and outputs a patch-level prediction:

$$\hat{\mathbf{y}}(i) = g_\phi(\mathbf{r}(q), m(i)) \tag{3}$$

where $m(i) = \{m_j\}_{j\in B_i}$. The mask tokens are parameterized by a shared learnable vector with an added positional embedding. Since we wish to make predictions for M target blocks, we apply our predictor M times, each time conditioning on the mask tokens corresponding to the target block locations we wish to predict.

loss: Given context-driven training targets $\mathbf{y}(i)$ and the predicted patch-level representation $\hat{\mathbf{y}}(i)$, we use the L2 loss:

$$Loss_{rec} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j\in B_i} ||y_j - \hat{y}_j||_2^2 \tag{4}$$

EEG2Rep Weights

The predictor parameters ϕ and the context network parameters θ are optimized through gradient-based methods. However, the target network's parameters $\bar{\theta}$ undergoes updates using an exponentially moving average (EMA) [21, 31] of the context network parameters:

$$\bar{\theta} = \tau\bar{\theta} + (1-\tau)\theta \tag{5}$$

We implement a schedule for the hyperparameter τ that linearly increases from an initial value τ_0 to the target value τ_e during the first τ_n updates. After this initial phase, the value remains constant for the rest of the training. This approach ensures that the target network is updated more frequently in the early stages of training when the context network is random, and less frequently in later stages when more robust parameters have been learned.

2.3 Semantic Subsequence Preserving (SSP)

Random masking has proven successful for Masked Autoencoders [17], where random patches are sampled without replacement, following a uniform distribution, resulting in significant efficiency improvements. Fig. 3 illustrates various masking strategies applied to EEG samples, with the top subplot showcasing the original EEG sample from Crowdsourced datasets [23] recorded for 0.5 seconds at 128 Hz. The subsequent subplots demonstrate different masking techniques, each with a 50% masking ratio. The second subplot exhibits random masking, following the Masked Autoencoders style. While the MAE approach has demonstrated success in computer vision tasks, it may pose challenges in building semantic representations for EEG data due to the lack of structure in the created masks (e.g., random binary masking). Notably, most of the masked regions can be reconstructed trivially, for example, through interpolation. On the other hand, block masking [21, 24], involves masking entire

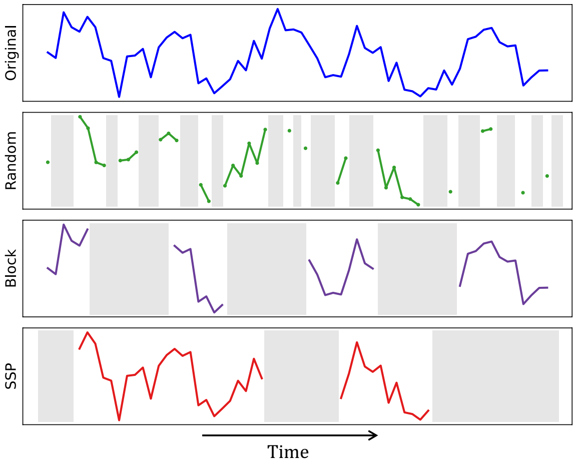


Figure 3: Different masking strategies applied to EEG samples from Crowdsourced datasets [23] with a 50% masking ratio. From top to bottom: (1) *Original* EEG sample, (2) *Random* masking, (3) *Block* masking, and (4) Semantic Subsequence Preserving (*SSP*).

blocks of time-steps or patches. However, this approach does not guarantee that large continuous portions of the training sample remain unmasked. As shown in Fig. 3, block masking primarily focuses on the structure of the mask rather than context. Our objective is to enable the context encoder to construct semantically rich representations over local regions of the sample.

To address this, we propose *Semantic Subsequence Preserving (SSP)*. Instead of determining which time steps to mask, this approach decides which time steps to preserve in a block-wise manner. The process begins by randomly selecting β starting points and symmetrically expanding them until the block reaches a width given by:

$$\text{Block Size} = \lceil \frac{(1-\rho)\times l}{\beta} \rceil \tag{6}$$

Here, l represents the total number of time steps (patches) in a training sample, β is the desired number of blocks, and ρ is the mask ratio. We allow visible blocks to overlap, leading to oversized subsequences preserving and some variability in the number of visible time steps for each EEG sample. As shown in Fig. 3, the first preserved subsequence arises from overlapping two blocks. As we encode only visible time steps, we adopt a simple strategy to maintain uniformity in the number of masked time steps across all samples in a batch. Following the semantic subsequence preserving step, we randomly preserve individual time steps until we reach the targeted number of preserved time steps, calculated as $(1-\rho)\times l$.

2.4 Efficiency: Multiple masking

As discussed in earlier sections, our model comprises three main components: the target, context, and predictor networks. The processing of each sample through these networks can be computationally intensive. Additionally, computing activations for the target network is less efficient compared to the others, as it requires processing the entire unmasked EEG input. Hence, to mitigate the computation cost, we implement a two-stage approach with multiple masking and multiple predictions.

Multiple Masking We explore N different masked versions of the training sample and compute the loss with respect to the same target representation. This is feasible because target representations are based on the full unmasked version of the sample. As N increases, the computational overhead of computing target representations becomes negligible.

Multiple Prediction Subsequently, having obtained representations of various masked versions of the same input, we predict the representation of various target blocks for a single masked input to further enhance efficiency. Now, the reconstruction loss in Equation 4 is updated to:

$$Loss_{rec^*} = \frac{1}{M\times N}\sum_{q=1}^{N}\sum_{i=1}^{M}\sum_{j\in B_i} ||y_j - \hat{y}_j||_2^2 \tag{7}$$

2.5 Loss Regularisation

A common challenge faced by algorithms generating and predicting their own targets is representation collapse, where the model produces similar representations for all masked segments, making the problem trivial to solve [32, 33]. Various strategies have been proposed to address this issue such as contrastive learning[34] for invariance-based methods and stop gradient[21, 31, 35] and additional clustering [33] for self-prediction methods. As we mentioned earlier, we use the exponentially moving average and stop gradient methods to prevent representation collapse similar to other models in CV and NLP like BYOL [35], Data2vec [21], IBOT [31]. However, in our experiment, we found that preventing collapse does not guarantee that the model learns high-quality representations. To further enhance the representation learning process, we add Variance-Invariance-Covariance (VICReg) [36] regularization to our current reconstruction loss. VICReg encourages more variety among the data in the batch by using a hinge loss that limits how much the standard deviation can change. It also involves a covariance loss, which penalizes the off-diagonal elements of the covariance matrix of the representation, encouraging less correlation between features. The final loss of our model is updated to:

$$Total\ Loss = \lambda Loss_{rec^*} + \mu v(R) + \gamma c(R) \tag{8}$$

Where *Total Loss* is minimized over batches of samples and $v(R)$ and $c(R)$ denotes the variance, and covariance losses on the representations, respectively. λ, μ, γ are hyperparameters controlling the balance between these loss components.

3 RELATED WORK

Traditional approaches to extract useful features from EEG are surveyed in [37]. In recent years, advanced self-supervised learning methods have been proposed aiming to derive representations from sparsely labeled EEG data. Here, we discuss two primary self-supervised approaches for EEG representation learning: invariance-based and self-prediction, and refer interested readers to [9, 14] for more details.

3.1 Invariance-based self-supervised learning

In the field of EEG, a series of studies have drawn inspiration from the SimCLR [38] framework in computer vision and ALBERT [39] in NLP, aiming to generate consistent embeddings for various yet compatible views of the same input [9, 14–16, 40]. This involves capturing consistent characteristics across different views, where EEG views are typically formed using a set of carefully designed data augmentations. For instance, SeqCLR [16] applies diverse strategies like amplitude scaling, temporal shifting, DC shifting, and band-pass filtering to create different views. Similarly, Time-series Temporal and Contextual Contrasting (TS-TCC) [15] use weak and strong augmentations to transform input series into two views, utilizing a temporal contrasting module to learn robust temporal representations. The contrasting contextual module is then built upon the contexts from the temporal contrasting module and aims to maximize similarity among contexts of the same sample while minimizing similarity among contexts of different samples.

Time-Frequency Consistency (TF-C) [40] leverages the frequency domain to achieve better representation. It proposes that the time-based and frequency-based representations, learned from the same time series sample, should be more similar to each other in the time-frequency space compared to representations of different time series samples. Kan et al. [41] introduced a new augmentation method called 'meiosis' that involves randomly exchanging data segments between samples to create positive correlations, using group-level contrastive learning to distinguish emotional states. Furthermore, Shen et al. [42] implemented cross-subject contrastive learning. They compared negative pairs from different individuals with positive pairs from the same subjects during similar events. This helps the model generalize EEG representations while reducing variations due to individual differences. While invariance-based methods can generate informative representations, they introduce strong biases that may lead to model misinterpretation or misclassification. Additionally, defining generalizable augmentations is more challenging in EEG data compared to other types of data.

3.2 Self-Prediction based self-supervised learning

The primary goal of self-prediction-based models is to reconstruct corrupted or masked input. Following the success of models like BERT [29], BErt-inspired Neural Data Representations (BENDR) [10] utilizes a mask-autoencoder approach to generate representations and minimize reconstruction loss between masked and reconstructed features. BENDR encodes multi-channel EEG signals into temporal embeddings using a 1D convolution block, inspired by the wave2vec approach [43]. It then employs a transformer encoder to process locally masked EEG signals for feature and representation extraction. Pre-trained on the extensive TUH EEG dataset [44, 45], the model shows improved performance across tasks like motor imagery, sleep stage classification, and event recognition. Similarly, MAEEG [12] refines this approach by adding two layers to convert the transformer outputs back to the original EEG dimensions, reconstructing the EEG signal from contextual features and minimizing the loss between original and reconstructed signals. Li et al. [46] introduce a multi-view mask autoencoder that reconstructs masked EEG content across spectral, spatial, and temporal dimensions to derive emotion-related EEG features.

In contrast to invariance-based methods, self-prediction pretraining tasks demand less prior knowledge and demonstrate ease of generalization across diverse downstream tasks [21]. However, the resulting representations are typically of a lower semantic level and may underperform invariance-based pretraining in off-the-shelf evaluations like linear probing and pretraining, mainly due to the intrinsic nature of EEG data. In this study, we explore ways to enhance the semantic level of self-supervised representations without relying on additional prior knowledge encoded through data augmentation. To achieve this, we introduce EEG2Rep to improve self-supervised EEG representations through informative masked inputs.

Table 1: Overview of EEG Datasets

	Datasets	Rate	Dim	Len	#Samples	Task
Emotiv	DREAMER [47]	128Hz	14	2s	77,910	Emotion Detection
	STEW [48]	128Hz	14	2s	26,136	Mental workload Classification
	Crowdsourced [23]	128Hz	14	2s	12,296	Eyes open/close Detection
	Driver Distraction	128Hz	14	2s	66,197	Driver Distraction Detection
Temple	TUAB [44]	256Hz	16	10s	409,455	Abnormal EEG Classification
	TUEV [45]	256Hz	16	5s	112,464	Event Detection

4 EXPERIMENTAL RESULTS

4.1 Datasets

We employed two distinct dataset types, totaling six datasets, to evaluate the effectiveness of our EEG2Rep model: i) Four datasets were obtained using different types of Emotiv headsets(14-channel wireless headsets capable of data collection in real-world scenarios). Three of these datasets are publicly available. The fourth dataset, named "Driver Distraction" is a private dataset provided by Emotiv, and collected using an older version of the headset [2]. Reporting results on this private dataset along with other public ones allows us to assess our model's effectiveness across different Emotiv headset types, each producing specific types of noise. ii) We expanded our evaluation to another dataset type, the Temple University Hospital (TUH) Corpus [44, 45], collected in a different setting (in a laboratory setting using 24-36 channel braincaps). TUH is one of the largest open EEG data repositories, featuring diverse devices with varying channel numbers, all collected in clinical settings. An overview of the datasets is available in Table 1, and additional descriptions, including details on data preprocessing, are provided in Appendix A.

To assess our model's performance, we partitioned the Emotiv datasets into subject-wise train/validation/test sets. This setup poses a challenge for models to learn generalized patterns, given the inter-subject variability. As for TUAB and TUEV, the training and test separation is inherent in the dataset. Additionally, we further split TUEV and TUAB training sets into 20% validation and 80% training subject-wise.

4.2 Competitors and Implementation Setup

We conducted extensive comparisons against five state-of-the-art methods for EEG representation learning. These methods include invariance-based approaches TS-TCC [15], TF-C [40], and BIOT [13] as well as self-prediction methods BENDR [10] and MAEEG [12]. To ensure a fair evaluation, we used publicly available code for the baseline methods. All experiments utilized the PyTorch framework on a system featuring a single Nvidia A5000 GPU (24GB). Model and hyperparameter selection relied on the validation set. Table 2 and Table 3 present five sets of results with varied random seeds,

[2] https://www.emotiv.com/epoc/

Table 2: Performance of EEG2Rep in comparison to the competitors in a *Linear Probing* setting.

Models	DREAMER		Crowdsourced		STEW		DriverDistraction		TUAB		TUEV	
	Acc	AUROC	Acc	AUROC	Acc	AUROC	Acc	AUROC	Acc	AUROC	B-Acc	W-F1
BENDR [10]	51.48±1.87	51.68±1.98	70.46±4.14	70.56±4.32	63.03±1.07	63.03±1.07	68.40±3.08	55.21±3.17	72.78±4.17	79.87±4.11	37.38±3.11	61.31±3.18
MAEEG [12]	$\underline{54.24\pm2.04}$	$\underline{53.98\pm2.68}$	75.21±2.11	75.01±2.01	$\underline{67.99\pm1.86}$	$\underline{68.58\pm1.88}$	68.37±2.60	55.29±2.20	72.62±3.99	79.75±4.07	37.23±2.99	61.38±3.08
TS-TCC [15]	53.60±2.68	53.87±2.28	$\underline{77.75\pm2.94}$	$\underline{77.83\pm2.11}$	64.54±1.62	64.64±1.73	$\underline{76.36\pm3.48}$	56.27±2.88	74.39±3.06	81.02±2.97	35.98±2.85	60.67±2.98
TF-C [40]	52.52±1.88	52.26±2.08	64.82±7.23	65.32±6.12	58.84±2.36	58.69±2.43	64.85±3.89	53.87±4.02	69.33±5.78	75.75±3.75	30.12±4.06	56.23±4.05
BIOT [13]	53.35±2.41	53.42±1.91	76.23±4.56	76.33±4.12	67.54±2.08	67.69±3.60	63.93±1.28	$\underline{63.33\pm1.25}$	$\underline{75.11\pm2.79}$	$\underline{82.92\pm2.01}$	$\underline{40.02\pm1.87}$	$\underline{65.98\pm2.01}$
EEG2Rep	**58.45**±1.82	**55.19**±1.95	**81.66**±2.93	**81.67**±2.65	**69.04**±1.04	**69.10**±1.23	**76.88**±2.55	**65.59**±2.43	**76.55**±3.33	**83.24**±3.25	**43.25**±3.12	**69.95**±3.21

reporting mean and standard deviation values. We adhere to the original VICReg [36] for setting the hyperparameters λ, μ, and γ in the total loss. For further details on the evaluation metrics, refer to the Appendix B.

4.3 Linear Probing

Table 2 presents the average performance of EEG2Rep along with other state-of-the-art methods over five runs. For each dataset, the number in **bold** indicates the highest accuracy achieved, while the number underlined represents the second best (This formatting is consistent across all tables presented in this paper). The results presented in Table 2 indicate that our model, EEG2Rep, achieves the highest average performance on all EEG tasks.

The DREAMER and STEW datasets exhibit high inter-subject variance, stemming partly from the limited number of patients in the recordings and partly due to the complexity of the EEG tasks. As observed in the results, self-prediction-based methods like MAEEG and EEG2Rep tend to capture more general patterns that can be applied across different subjects. In contrast, tasks such as eyes open/eyes close in Crowdsourced dataset are easier to generalize among subjects due to the nature of the EEG task. TUAB also experiences low inter-subject variance as it encompasses a substantial number of subjects (more than 1000 patients). The results highlight that invariance-based methods like TS-TCC and BIOT outperform BENDER and MAEEG in Crowdsourced and TUAB. However, EEG2Rep while being a self-prediction technique, manages to improve the semantic level of representations resulting in the best average accuracy among all competitors.

4.4 Fine Tuning

Table 3 presents the average classification accuracy results across different datasets, comparing the performance of EEG2Rep with other pre-trained models that were initialized using their respective pretext tasks. These results are consistent with the ones presented in Table 2, and our EEG2Rep achieves the highest average performance on all EEG tasks.

The comparison also includes another version of the EEG2Rep model, initialized randomly and denoted as EEG2Rep (Random). This is equivalent to supervised training. The table shows that using pre-trained EEG2Rep leads to an average accuracy improvement of 6% on average compared to EEG2Rep (Random). Significant enhancements are evident in specific datasets, particularly DREAMER, DriverDistraction, and TUEV. In DREAMER and DriverDistraction, EEG2Rep demonstrates AUROC improvement of 5.81% and 2.26%, respectively, when compared to random initialization. Similarly, for TUEV, there is a 6.85% improvement in weighted F1, validating the effectiveness of incorporating informative context input in self-supervised methods for enhanced learning and improved EEG classification. It's worth noting that in datasets with pronounced subject invariance issues, such as DREAMER, STEW, Driverdistraction, and TUAB, the performance of the supervised models is notably subpar. In some cases, these models even exhibit lower performance compared to models pre-trained self-supervised and utilized for classification through a linear layer. For instance, in DREAMER, linear probed EEG2Rep achieves a 3.84% higher accuracy than its supervised counterpart.

4.5 Cross-Domain

We evaluated the performance of our EEG2Rep model in a cross-domain setting, where the model is trained on one dataset and tested on another. This approach assesses the model's ability to generalize across different types of EEG domains and tasks. We utilized two Emotiv datasets for the target tasks: STEW and Crowdsourced, both characterized by limited training samples, making cross-domain evaluation particularly relevant.

As depicted in Table 4, pre-training the model with datasets like DREAMER led to performance improvements for the STEW dataset, where the task is mental workload classification. This improvement can be attributed to the contextual alignment between activities recorded in the DREAMER and STEW datasets. Interestingly, even though the DriverDistraction dataset is not directly related to mental workload classification, the experimental environment and certain distraction classes are similar to those in the STEW dataset. This similarity allowed the EEG2Rep model to benefit from pre-training on the DriverDistraction dataset, as reflected in the performance.

For the Crowdsourced dataset, where the task is eye open/closed classification, the model achieved robust performance when pre-trained on all available Emotiv datasets. However, this performance did not surpass that achieved through in-domain pre-training. This outcome suggests that the alignment of brain activities across datasets is critical for optimal performance and indicates the potential need to utilize a larger number of pre-training datasets to cover various subjects and tasks comprehensively.

Table 3: Performance of EEG2Rep in comparison to the competitors in a *Fine Tuning* setting.

Models	DREAMER		Crowdsourced		STEW		DriverDistraction		TUAB		TUEV	
	Acc	AUROC	Acc	AUROC	Acc	AUROC	Acc	AUROC	Acc	AUROC	B-Acc	W-F1
BENDR [10]	54.45±2.11	53.02±3.11	83.78±2.35	83.8±2.63	69.74±2.11	69.77±2.03	74.31±2.38	59.86±2.6	76.96±3.98	83.97±3.44	41.17±2.89	67.31±2.96
MAEEG [12]	53.63±2.61	52.08±2.36	86.75±3.50	86.21±3.41	72.46±3.67	72.5±3.22	74.58±2.16	60.79±2.72	77.56±3.56	86.56±3.33	41.23±3.65	67.38±3.69
TS-TCC [15]	58.16±2.11	55.05±1.79	89.22±1.22	89.22±1.22	71.00±2.98	71.03±3.02	74.21±2.68	60.33±2.66	79.66±2.99	87.02±2.68	40.98±2.55	68.67±2.89
TF-C [40]	52.82±1.66	52.86±1.85	82.93±4.02	82.90±4.32	68.65±1.12	68.70±1.75	65.39±4.12	58.75±3.98	72.33±5.64	78.48±3.85	40.12±3.66	66.23±3.85
BIOT [13]	53.45±2.01	53.53±2.13	87.95±3.52	87.78±3.09	69.88±2.15	70.11±2.57	74.34±3.57	61.21±4.36	79.21±2.15	87.42±2.01	46.02±1.68	69.98±1.99
EEG2Rep (Random)	54.61±2.22	53.61±2.09	91.19±1.18	91.22±1.23	70.26±1.59	70.49±1.86	72.95±2.95	59.5±3.17	77.85±3.14	84.91±3.07	44.25±3.01	68.95±2.89
EEG2Rep (Pre-Trained)	**60.37**±1.52	**59.42**±1.45	**94.13**±2.11	**94.13**±2.17	**73.60**±1.47	**74.40**±1.50	**80.07**±2.63	**66.14**±2.44	**80.52**±2.22	**88.43**±3.09	**52.95**±1.58	**75.08**±1.21

Table 4: EEG2Rep model performance in cross-domain settings on STEW and Crowdsourced datasets.

Models	STEW		Crowdsourced	
	ACC	AUROC	ACC	AUROC
Random initialization	70.26±1.59	70.49±1.86	91.19±1.18	91.22±1.23
In-domain pre-trained	73.60±1.47	74.40±1.50	**94.13**±2.11	**94.13**±2.17
Pre-trained on DREAMER	73.75±1.95	74.95±2.77	93.91±1.78	93.86±1.80
Pre-trained on DriverDistraction	73.68±2.17	74.67±3.06	94.09±1.95	94.11±2.02
Pre-trained on All Emotiv	**74.11**±2.34	**77.38**±2.77	94.05±1.68	94.07±1.75

Table 5: Ablation Study of EEG2Rep Components

EEG2Rep	**Average Accuracy: 67.64**
Masking Strategy	
Block Masking	66.45 (↓ 1.19)
Random Masking	60.19 (↓ 7.45)
Targets	
Input-Space Prediction	54.52 (↓ 13.12)
Loss Component	
W/O Variance-Covariance	65.63 (↓ 2)

4.6 Ablation Study

Masking Strategy. We compare our semantic subsequence preserving (SSP) strategy with other random and block masking strategies. For random masking, we adopt an approach similar to masking autoencoders [17], where patches are shuffled randomly, and the initial 50% of these patches are selected. In the case of block masking, we follow the MAEEG [12] approach, emphasizing the masking of a continuous chunk. The results in Table 5 highlight the effectiveness of semantic subsequence preservation in guiding our model towards learning meaningful representations. The subpar performance in the random masking strategy could be attributed to the model primarily attempting to interpolate the masked values rather than focusing on learning a semantic representation. With block masking, the visible subsequence may have a shorter length than the natural time scale of the brain, posing a significant challenge to the reconstruction process.

Masking Ratio. The masking ratio during self-supervised pretraining is crucial as it determines both the difficulty of the self-prediction task and the quality of the learned representations. Therefore, identifying an optimal masking ratio is essential for effective representation learning from EEG data. Following the recent study on the effects of masking on representation learning [49], we pretrain our model using five different mask rates [10%, 25%, 50%, 75%, 90%] with our SSP method and random masking to evaluate the quality of representation learning. As shown in Table 6, several interesting findings emerged from the results. We observed that conservative masking (10%) leads to low performance, regardless of the masking strategy. However, as the masking ratio increases, the performance of the model improves, indicating that higher masking can result in more high-level representation learning. Nevertheless, performance degradation occurs with overly aggressive masking (90%), suggesting that representation does not become monotonically more high-level with increasing masking aggressiveness. In other words, overly aggressive masking also leads to low-level representations, similar to conservative masking. Moderate and high masking ratios yield the best results for both random and SSP masking strategies.

As shown in Table 6, our SSP masking strategy consistently achieves higher performance than random masking across various masking ratios. A key distinction between our proposed masking strategy in EEG2Rep and random masking lies in the arrangement of visible patches. In random masking, there is a possibility that visible patches are not consecutive, whereas in EEG2Rep, visible patches (referred to as preserved patches) are ensured to remain contiguous. Additionally, during the reconstruction of masked patches near the boundary between masked and unmasked regions, the model uses nearby visible patches to interpolate, thereby capturing low-level information, which resonates with the empirical observations in [17, 49]. Our proposed masking strategy is less susceptible to such phenomena compared to random masking, as it ensures that masked tokens remain contiguous. The high number of boundaries in random masking may result in less contextualized learning. Our empirical results further confirm that our proposed method yields higher performance compared to random masking, even with aggressive and conservative masking.

Masking Blocks. In EEG2Rep, our focus is on preserving subsequences to ensure information is retained in the masked input. Figure 4 depicts the average accuracy of EEG2Rep relative to the

Table 6: Performance comparison between random masking and SSP across different datasets and masking ratios.

	DREAMER		Crowdsourced		STEW		DriverDistraction		TUAB		TUEV	
Mask Ratio	Random	SSP	Random	SSP	Random	SSP	Random	SSP	Random	SSP	Random	SSP
10%	52.11	53.02	72.13	76.69	63.13	68.14	64.02	70.41	64.50	72.36	33.01	40.37
25%	**53.23**	54.73	75.05	77.28	**65.12**	**70.47**	**65.40**	72.94	64.75	73.73	33.14	42.53
50%	52.63	**59.89**	**76.69**	**81.24**	64.97	69.52	64.12	**76.12**	**68.12**	**76.55**	**34.59**	**44.23**
75%	51.77	52.03	72.10	76.24	63.27	64.69	63.12	65.03	66.34	70.81	35.28	42.19
90%	51.51	51.95	70.77	73.44	62.05	63.70	62.06	64.91	66.34	70.12	34.14	38.72
Average	52.25	**54.32**	73.35	**76.98**	63.71	**67.30**	63.74	**69.88**	66.01	**72.71**	34.03	**41.61**

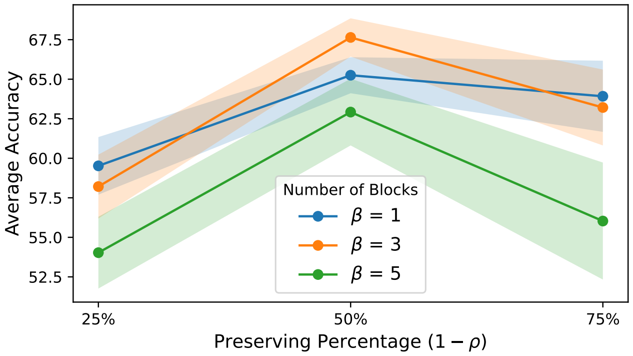


Figure 4: Effect of preserving percentage $(1 - \rho)$ on average accuracy across all EEG datasets: A comparison of accuracy variation across different numbers of blocks (β), with error bars indicating standard deviation.

number of masked blocks and the ratio of preserved subsequences (inverse of masking). Optimal results are observed when preserving 50% of the input in 3 blocks and masking the other half. When preserving only 75% of the input, it is more effective to mask a single chunk rather than multiple chunks, as the model tends to interpolate instead of learning semantic patterns in EEG data.

Input-Space Prediction. Table 5 shows a comparison of linear probing performance when the loss is computed in input-space versus representation space. We hypothesize that a key component of EEG2Rep lies in computing the loss entirely in the representation space, enabling the target encoder to generate abstract prediction targets that eliminate irrelevant raw details. As evident from Table 5, predicting in input-space leads to a 13.12% degradation in linear probing performance, highlighting the impact of noise and wide range of amplitudes inherent to EEG data.

Loss Regularization. As discussed in Section 2.5, the self-supervised loss function comprises two components: reconstruction and regularization. The main role of the reconstruction term is to ensure similarity between the representations of masked and unmasked versions of the same sample. In contrast, the regularization term not only aids in preventing representation collapse but also enhances the representation learning task. As shown in Table 5, the regularization term contributes to a 2% accuracy improvement on average across all EEG tasks.

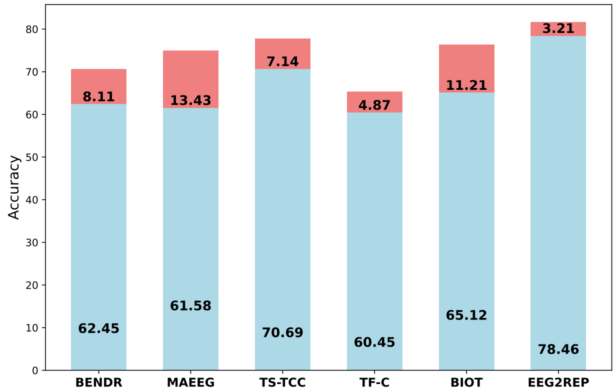


Figure 5: Model Robustness Comparison: Assessing model performance by introducing Gaussian noise, DC-shift, and amplitude changes to Crowdsourced data.

Robustness to Noise. In this experiment, our goal is to assess the robustness of our model to noise, recognizing the high signal-to-noise ratio as a significant challenge in learning representations from EEG data. To do so we introduce Gaussian noise, time shift, DC-shift, and amplitude scaling, as recommended by neurologists [16] to the Crowdsourced dataset. Such transformation do not change the interpretation of the EEG according to [16]. The details of noise types are described in appendix B.2. As depicted in Figure 5, our model demonstrates superior robustness to these noises in the data, while raw input-based models like BENDER and MAEEG exhibit the most degradation in accuracy. TS-TCC and TF-C show commendable robustness to these noises, possibly due to their utilization of similar augmentations for their invariance losses.

5 CONCLUSION

EEG2Rep emerges as a pioneering self-supervised approach tailored to address the inherent challenges of EEG data representation learning, including low signal-to-noise ratio. By innovatively predicting masked inputs in the latent representation space and employing a novel semantic subsequence preserving method, EEG2Rep facilitates the generation of rich semantic representations. Our extensive experiments across six diverse EEG tasks demonstrate that EEG2Rep not only significantly surpasses state-of-the-art methods but also highlights remarkable robustness to noise. We found that preserving 50% of EEG recordings optimizes accuracy across all tasks. In the future, we will explore the alignment of this finding with the natural time scale of the brain on different tasks.

REFERENCES

[1] M. Teplan *et al.*, "Fundamentals of eeg measurement," *Measurement science review*, vol. 2, no. 2, pp. 1–11, 2002.

[2] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields.* Lippincott Williams & Wilkins, 2005.

[3] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[4] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.

[5] J. Chen, Y. Yang, T. Yu, Y. Fan, X. Mo, and C. Yang, "Brainnet: Epileptic wave detection from seeg with hierarchical graph diffusion learning," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2741–2751.

[6] M. Raković, Y. Li, N. M. Foumani, M. Salehi, L. Kuhlmann, G. Mackellar, R. Martinez-Maldonado, G. Haffari, Z. Swiecki, X. Li *et al.*, "Measuring affective and motivational states as conditions for cognitive and metacognitive processing in self-regulated learning," in *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 2024, pp. 701–712.

[7] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[8] M.-P. Hosseini, A. Hosseini, and K. Ahi, "A review on machine learning for eeg signal processing in bioengineering," *IEEE reviews in biomedical engineering*, vol. 14, pp. 204–218, 2020.

[9] W. Weng, Y. Gu, S. Guo, Y. Ma, Z. Yang, Y. Liu, and Y. Chen, "Self-supervised learning for electroencephalogram: A systematic survey," *arXiv preprint arXiv:2401.05446*, 2024.

[10] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, vol. 15, 2021.

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

[12] H.-Y. S. Chien, H. Goh, C. M. Sandino, and J. Y. Cheng, "Maeeg: Masked auto-encoder for eeg representation learning," in *NeurIPS Workshop*, 2022.

[13] C. Yang, M. B. Westover, and J. Sun, "Biot: Biosignal transformer for cross-data learning in the wild," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[14] N. Mohammadi Foumani, L. Miller, C. W. Tan, G. I. Webb, G. Forestier, and M. Salehi, "Deep learning for time series classification and extrinsic regression: A current survey," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–45, 2024.

[15] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021, pp. 2352–2359.

[16] P. M. Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, "Contrastive representation learning for electroencephalogram classification," in *Machine Learning for Health Workshop, ML4H@NeurIPS 2020, Virtual Event, 11 December 2020*, ser. Proceedings of Machine Learning Research, vol. 136, 2020, pp. 238–253.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[18] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2114–2124.

[19] H. Ling, Y. Luyuan, L. Xinxin, and D. Bingliang, "Staging study of single-channel sleep eeg signals based on data augmentation," *Frontiers in Public Health*, vol. 10, p. 1038742, 2022.

[20] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, "Series2vec: Similarity-based self-supervised representation learning for time series classification," *arXiv preprint arXiv:2312.03998*, 2023.

[21] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning.* PMLR, 2022, pp. 1298–1312.

[22] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *International Conference on Machine Learning.* PMLR, 2023, pp. 1416–1429.

[23] N. S. Williams, W. King, G. Mackellar, R. Randeniya, A. McCormick, and N. A. Badcock, "Crowdsourced eeg experiments: A proof of concept for remote eeg acquisition using emotivpro builder and emotivlabs," *Heliyon*, vol. 9, no. 8, 2023.

[24] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[25] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "Image BERT pre-training with online tokenizer," in *International Conference on Learning Representations*, 2022.

[26] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[27] S. N. M. Foumani, C. W. Tan, and M. Salehi, "Disjoint-cnn for multivariate time series classification," in *2021 International Conference on Data Mining Workshops (ICDMW).* IEEE, 2021, pp. 760–769.

[28] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, "Improving position encoding of transformers for multivariate time series classification," *Data Mining and Knowledge Discovery*, vol. 38, no. 1, pp. 22–48, 2024.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "Image bert pre-training with online tokenizer," in *International Conference on Learning Representations*, 2021.

[32] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *International Conference on Learning Representations*, 2021.

[33] I. Ben-Shaul, R. Shwartz-Ziv, T. Galanti, S. Dekel, and Y. LeCun, "Reverse engineering self-supervised learning," *arXiv preprint arXiv:2305.15614*, 2023.

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning.* PMLR, 2020, pp. 1597–1607.

[35] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[36] A. Bardes, J. Ponce, and Y. Lecun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *ICLR 2022-International Conference on Learning Representations*, 2022.

[37] D. P. Subha, P. K. Joseph, R. Acharya U, and C. M. Lim, "Eeg signal analysis: a survey," *Journal of medical systems*, vol. 34, pp. 195–212, 2010.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning.* PMLR, 2020, pp. 1597–1607.

[39] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[40] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," in *Proceedings of Neural Information Processing Systems, NeurIPS*, 2022.

[41] H. Kan, J. Yu, J. Huang, Z. Liu, H. Wang, and H. Zhou, "Self-supervised group meiosis contrastive learning for eeg-based emotion recognition," *Applied Intelligence*, vol. 53, no. 22, p. 27207–27225, sep 2023.

[42] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2496–2511, 2023.

[43] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[44] S. Lopez, G. Suarez, D. Jungreis, I. Obeid, and J. Picone, "Automated identification of abnormal adult eegs," in *2015 IEEE signal processing in medicine and biology symposium (SPMB).* IEEE, 2015, pp. 1–5.

[45] A. Harati, M. Golmohammadi, S. Lopez, I. Obeid, and J. Picone, "Improved eeg event classification using differential energy," in *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB).* IEEE, 2015, pp. 1–4.

[46] R. Li, Y. Wang, W.-L. Zheng, and B.-L. Lu, "A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6–14.

[47] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[48] W. Lim, O. Sourina, and L. P. Wang, "Stew: Simultaneous task eeg workload data set," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2106–2114, 2018.

[49] L. Kong, M. Q. Ma, G. Chen, E. P. Xing, Y. Chi, L.-P. Morency, and K. Zhang, "Understanding masked autoencoders via hierarchical latent variable models," in

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7918–7928.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

A DATASET OVERVIEW AND PROCESSING

A.1 Emotiv

We applied a bandpass filter to all Emotiv datasets and windowed them into segments consisting of 256 time steps, each equivalent to 2 seconds of recording.

DREAMER

DREAMER is a multimodal database featuring electroencephalogram (EEG) and electrocardiogram (ECG) signals recorded during affect elicitation using audio-visual stimuli [47] using a 14-channel Emotiv EPOC headset. The dataset includes signals from 23 participants, accompanied by their self-assessments of affective states after each stimulus in terms of valence, arousal, and dominance. For our classification task, we specifically utilize the arousal labels. The DREAMER dataset can be obtained from here[3], and we employ the Torcheeg toolkit for preprocessing, which involves cropping and applying low-pass and high-pass filters[4]. It is important to note that, for our analysis, we solely focus on EEG data, and the ECG signals are excluded from consideration.

Crowdsourced

Crowdsourced EEG data was collected while participants were engaged in a resting state task, involving periods with eyes open and eyes closed, each lasting 2 minutes. Out of 60 participants, only 13 successfully completed both conditions using 14-channel EPOC+, EPOC X, and EPOC devices. The data was initially recorded at 2048 Hz and later downsampled to 128 Hz. Raw EEG data for these 13 participants, along with preprocessing, analysis, and visualization scripts, are openly available on the Open Science Framework (OSF)[5].

Simultaneous Task EEG Workload (STEW)

STEW dataset comprises raw EEG recordings from 48 participants using a 14-channel Emotiv EPOC headset involved in a multitasking workload experiment utilizing the SIMKAP multitasking test [48]. Additionally, the subjects' baseline brain activity at rest was recorded before the test. The data was captured using the Emotiv Epoc device with a sampling frequency of 128Hz and 14 channels, resulting in 2.5 minutes of EEG recording for each case. Participants were instructed to assess their perceived mental workload after each stage using a rating scale ranging from 1 to 9, and these ratings are available in a separate file. Moreover, this dataset includes binary class labels, considering a workload rating of more than 4 as high and otherwise as low. We utilize these labels for our specific problem. STEW can be accessed upon request through the IEEE DataPort[6].

DriverDistraction

The data were gathered by recording the EEG brain signals of 17 participants, each using a driving simulator for approximately 40 minutes. The participants performed various distraction activities while they were driving. These can be grouped into the following high-level activities: 1. Talking to a passenger 2. Using a phone (texting and calling) 3. Problem solving. The EEG data were sampled at a frequency of 128Hz, through 14 channels on the Emotiv Epoc EEG headset. The sampling result is a multivariate (14 input variables) time-series dataset containing approximately 5.5 million records. The data were then manually annotated with the activity being performed at each time point.

A.2 Temple University Hospital (TUH) EEG Corpus

TUH Abnormal EEG Corpus (TUAB)

The TUH Abnormal EEG Corpus (TUAB) is a subset of the Temple University Hospital (TUH) EEG Corpus, which is one of the largest publicly available collections of clinical EEG data. The TUAB specifically focuses on EEG recordings labeled as abnormal, making it a valuable resource for studies on neurological disorders, brain function anomalies, and the development of diagnostic tools [44].

TUH EEG Events (TUEV)

The TUH EEG Events Corpus (TUEV) contains annotations of EEG segments classified into six different categories: spike and sharp wave, generalized periodic epileptiform discharges, periodic lateralized epileptiform discharges, eye movement, artifact, and background [45].

Acquisition and Preprocessing

The TUH Abnormal EEG Corpus (TUAB) [44] and TUH EEG Events (TUEV) [45] can be accessed upon request through the Temple University Electroencephalography (EEG) Resources[7]. We processed both datasets to adhere to the 16 EEG montages [13], following the 10-20 international system, are as follows: "FP1-F7", "F7-T7", "T7-P7", "P7-O1", "FP2-F8", "F8-T8", "T8-P8", "P8-O2", "FP1-F3", "F3-C3", "C3-P3", "P3-O1", "FP2-F4", "F4-C4", "C4-P4", and "P4-O2".

B DETAILS OF EXPERIMENTAL SETTINGS

B.1 Parameter Setting

In our experiment, the EEG2Rep model employed one depth-wise and two spatial-wise convolution layers for input embedding, each with 16 filters. During training, a batch size of 256 was used, and we utilized the Adam optimization algorithm [50]. To prevent overfitting, we implemented an early stopping method based on the validation loss. The model was pre-trained for 500 epochs, after which logistic regression was applied to the representations for linear probing. Similar to the transformer-based model for multivariate time series classification (TST) [18] and the default transformer block [30], in our experiments, we employed eight attention heads in both the context and target encoder to capture diverse features from the EEG. The transformer encoding dimension was set to $d_e = 16$, and the feed-forward network (FFN) in the transformer block expanded the input size by a factor of 4 before projecting it

[3] https://zenodo.org/records/546113

[4] https://torcheeg.readthedocs.io/en/v1.1.0/torcheeg.datasets.html

[5] https://osf.io/9bvgh

[6] https://ieee-dataport.org/open-access/stew-simultaneous-task-eeg-workload-dataset

[7] https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml

back to its original size. For the learning rate, we start with 1×10^{-3} and use a cosine learning rate scheduler to adjust it over time [21].

B.2 Noise Types for Model Robustness

Table 7 provides the details of the noise types we added to the Crowdsourced dataset to test the robustness of EEG2Rep and benchmark models to noise.

Table 7: Noise Types Details

Transformation	Min	Max
Amplitude Scale	0.5	2
Time Shift	-50	50
DC Shift	-10	10
Additive Gaussian Noise	0	0.2

B.3 Evaluation Metrics

(Balanced) Accuracy is defined as the average recall obtained for each class. We use the term *ACC* for binary classification and *B-ACC* for multi-class classification. **AUROC** represents the area under the ROC curve, condensing the ROC curve into a single number that measures the model's performance across multiple thresholds. It is employed for binary classification. **Weighted F1** is utilized for multi-class classification in this paper. It is a weighted average of individual F1 scores for each class, with each score weighted by the number of samples in the corresponding class.