Brant-2: Foundation Model for Brain Signals

Zhizhang Yuan
Zhejiang University
zhizhangyuan@zju.edu.cn

Daoze Zhang
Zhejiang University
zhangdz@zju.edu.cn

Junru Chen
Zhejiang University
jrchen_cali@zju.edu.cn

Gefei Gu
Zhejiang University
frankgu@zju.edu.cn

Yang Yang†
Zhejiang University
yangya@zju.edu.cn

arXiv:2402.10251v4 [q-bio.NC] 28 Mar 2024

ABSTRACT

Foundational models benefit from pre-training on large amounts of unlabeled data and enable strong performance in a wide variety of applications with a small amount of labeled data. Such models can be particularly effective in analyzing brain signals, as this field encompasses numerous application scenarios, and it is costly to perform large-scale annotation. In this work, we present the largest foundation model in brain signals, *Brant-2*. Compared to Brant, a foundation model designed for intracranial neural signals, *Brant-2* not only exhibits robustness towards data variations and modeling scales but also can be applied to a broader range of brain neural data. By experimenting on an extensive range of tasks, we demonstrate that *Brant-2* is adaptive to various application scenarios in brain signals. Further analyses reveal the scalability of the *Brant-2*, validate each component's effectiveness, and showcase our model's ability to maintain performance in scenarios with scarce labels.

KEYWORDS

Foundation model, Brain signal, Pre-training

1 INTRODUCTION

Brain signals refer to the biometric information collected from the brain [57]. Their patterns provide valuable insights towards understanding the physiological function of the brain and the mechanism of related diseases, leading to various applications like neurological disorders [1, 9], sleep health research [34, 36, 43], emotion recognition [41, 42] and so on. Brain signals are usually measured by invasive methods like stereoelectroencephalography (SEEG) or non-invasive methods like scalp electroencephalography (EEG). SEEG requires extra surgeries to implant the recording devices, resulting in a high cost [24]. However, it manifests advantages by providing stereotactic and detailed information about deep brain structures. As a non-invasive method, EEG fails to capture deep brain information accurately and contains more noise due to the placement of electrodes on the scalp. However, compared to SEEG, EEG is more accessible to implement without surgery, leading to more application scenarios. Despite the differences, both SEEG and EEG data use the same principle of electrical activity recording [8] and share the same physiological basis.

The field of brain signals encompasses a wide array of downstream tasks. It represents a cutting-edge domain that will continue to unveil new research directions and application scenarios in the future [57]. In addition, after collecting the brain signals, the annotation work highly relies on experts in the corresponding field, making large-scale data labeling infeasible [13]. However, existing works in this field are mainly designed to solve specific tasks [9, 54, 59]. Moreover, many of them [20, 27] require training models from scratch, which tends to have a high dependency on labels. Very limited research provides an off-the-shelf model, known as the foundation model, that can be applied to multiple scenarios in this field and serve as a tool for further investigation of the brain. Foundation models have shown great potential in language [44, 45, 50] and vision [48, 53], which not only allow for customization for diverse applications but also reduce costs for data annotation [6]. Furthermore, by leveraging foundation models, researchers and developers no longer need to build models from scratch, which saves time and costs, benefiting the advancement of the field. Therefore, we aim to build a foundation model for brain signals that can effectively solve numerous downstream tasks for both SEEG and EEG data. However, when building a foundation model for brain signals, it is inevitable to encounter challenging issues along the way.

Firstly, different data exhibit differences in terms of sampling rates as well as the positions and quantities of electrodes. SEEG data is often recorded at a high sampling rate of at least 1000Hz [16] and exhibits significant inter-individual variability in electrode numbers and locations. EEG data is usually sampled at a lower frequency than SEEG [10] and can vary significantly in terms of montage (the number and the places of electrodes placed on the scalp) [51]. Secondly, brain signals collected from different scenarios contain distinct physiological characteristics, leading to varying modeling scales. For example, a sleep stage in sleep studies is often defined as lasting up to 30 seconds [20, 34, 36, 43], seizure detection may utilize a time scale of less than 10 seconds [9, 54], and existing works for emotion recognition adopt modeling scales of 5 seconds or shorter [42, 51]. Thirdly, there is substantial diversity among different tasks in the field of brain signals. For example, in seizure detection (identifying whether a segment includes seizure waves), the model is required to extract information from the target signal, such as capturing spikes and sharp waves within the signal. On the other hand, in seizure prediction (predicting whether there will be future epileptic seizures), the model needs to anticipate future changes of the target signal.

As Fig. 1 shows, to build such a foundation model, the first step is to gather a large amount of unlabeled brain neural data, which is then used for large-scale pre-training, overcoming the challenges above. For applications, as an off-the-shelf model, the pre-trained model can be applied to various downstream scenarios through fine-tuning. In the field of brain signals, Zhang et al. [56] propose a

† Corresponding author.

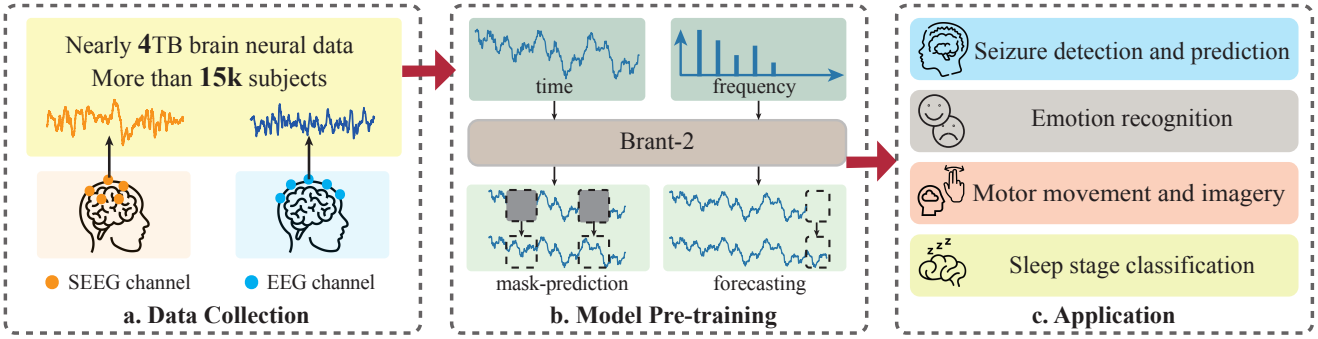

Figure 1: Overview of our work. We initially utilized approximately 4 TB of brain neural data from over 15k subjects to construct our pre-training corpus. Subsequently, we employ the corpus to train *Brant-2* using two pre-training tasks. Then the pre-trained model can be fine-tuned and applied to various application scenarios of brain signals.

foundation model for SEEG, Brant, which can capture long-term dependency, spatial correlation, and time-frequency information from SEEG signals. However, Brant has some limitations that prevent it from addressing the above challenges. Therefore, we propose *Brant-2*, a foundation model for brain signals, which excels in three main aspects. Firstly, the pre-training corpus of *Brant-2* is large and diverse. As shown in Fig. 1(a), *Brant-2* utilizes nearly 4 TB of mixed SEEG and EEG data with more than 15k subjects. Due to the large volume and diversity of data, *Brant-2* contains over 1 billion parameters. Secondly, *Brant-2* is robust to data variations and different modeling scales. Brant is pre-trained on multi-channel data with a fixed sampling rate and window length, by which it struggles to handle data variations and adapt to changes in modeling scales. During the pre-training process of *Brant-2*, we design a data augmentation module to further expand the diversity of the pre-training corpus, which enhances the robustness of *Brant-2* towards data variations and modeling scales. Thirdly, *Brant-2* can be applied to a broad range of tasks and scenarios. As shown in Fig. 1(b), compared to Brant, which is only pre-trained with mask-prediction, *Brant-2* learns more comprehensive semantic knowledge through two pre-training tasks, leading to better generalization abilities to a wider set of downstream tasks (shown in Fig. 1(c)). In summary, our key contributions comprise:

- We propose a foundation model *Brant-2*, the first off-the-shelf model that can be applied to scenarios of both SEEG and EEG. *Brant-2* is the largest model in brain signals pre-trained with nearly 4 TB brain signal data from more than 15k subjects.
- The pre-training framework we designed not only enhances the robustness of the model to significant data variations and different modeling scales but also empowers the ability to adapt to diverse downstream tasks in brain signals.
- We evaluate *Brant-2* on a wide range of downstream tasks to illustrate the generalization ability of our model. By conducting additional analysis experiments, we demonstrate our model's scalability, confirming each component's efficacy and highlighting its ability to sustain performance in scenarios with limited labels.

2 METHOD

As aforementioned, building a foundation model for brain signals primarily requires handling data variations, different modeling scales, and diverse tasks. To tackle the variations of the data and modeling scales, we employ data augmentation during pre-training to enhance the diversity of the training data, improving our model's robustness. To learn complex semantic representations and adapt to diverse downstream tasks, *Brant-2* integrates time and frequency information and simultaneously focuses on reconstructing the input and forecasting future signals based on partial observations.

Notations. We use $\boldsymbol{s}_i \in \mathbb{R}^{C\times(B+F)}$, $i \in \{1, 2..., N\}$ with C channel(s) and $B + F$ time steps to represent a segment of SEEG or EEG signals obtained from the pre-training corpus, where N denotes the total sample number. The sample $\boldsymbol{s}_i$ is divided into two consecutive parts: the look back window $\boldsymbol{x}_i \in \mathbb{R}^{C\times B}$ and the future values $\boldsymbol{x}_i^{\text{fut}} \in \mathbb{R}^{C\times F}$, where the look back window $\boldsymbol{x}_i$ serves as the input.

2.1 Overall Architecture

The overall architecture of *Brant-2* is shown in Fig. 2, which mainly involves four modules: 1) patching; 2) data augmentation and masking module; 3) input embedding module; 4) encoder.

Patching. Aggregating time steps into subseries-level patches can not only enhance the locality and capture comprehensive semantic information, but also reduce computation cost [33, 56]. Thus, we divide the input sample $\boldsymbol{x}_i$ into non-overlapped patches with length P and generate a set of patches $\boldsymbol{p}_i \in \mathbb{R}^{C\times L\times P}$, where $L = \lfloor B/P \rfloor$ is context length (i.e., the number of consecutive patches).

Data augmentation and masking. The quality of the data is of paramount importance for training a foundation model [26]. When assessing data quality, the diversity of the data is a crucial metric. High data diversity is beneficial for improving model performance, while low diversity can introduce biases and inaccuracies in the training process. In the field of language, LLMs (Large Language Models) are pre-trained using text corpus sourced from various domains [44, 45, 50]. Furthermore, Lee et al. [25] measure the diversity of publicly available LLM datasets and conclude that these datasets are highly diverse, which emphasizes the significance of data diversity for a foundation model.

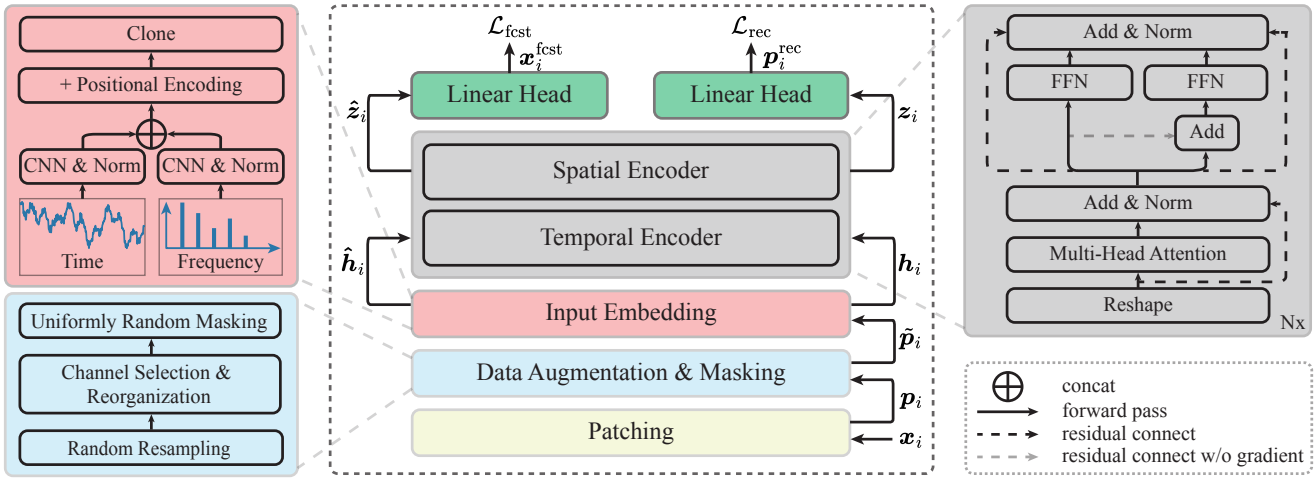

Figure 2: The architecture and pre-training framework of *Brant-2*. The input raw signal x_i is first processed to subseries-level patches p_i. Then we conduct data augmentation to increase the diversity of the training data and mask a subset of patches. We combine the information from both time and frequency domains to obtain the input embedding $h_i, \hat{h}_i$, which are then fed into the temporal and spatial encoder sequentially. The output representations $z_i, \hat{z}_i$ are linearly mapped to reconstruct the masked patches and forecast future signals.

In view of the significance of data diversity during pre-training, we introduce a data augmentation module to enhance the pre-training corpus in both the temporal and spatial dimensions, aiming to generate more diverse data. For the obtained patches $\boldsymbol{p}_i \in \mathbb{R}^{C\times L\times P}$, the temporal augmentation involves a random resampling to adjust the sampling rate of the input sample. The variation in sampling rates enriches the temporal scale of the samples, allowing the model to become more robust to handle changes in modeling scales. Formally, we choose an adjustment factor m_k from $\mathcal{M} = \{m_1, ..., m_K\}$ uniformly at random, then the input patches and future values are resampled by a factor of m_k, which are denoted as $\hat{\boldsymbol{p}}_i \in \mathbb{R}^{C\times L\times P_k}$ and $\hat{\boldsymbol{x}}_i^{\text{fut}} \in \mathbb{R}^{C\times F_k}$, where $P_k = P \times m_k$, $F_k = F \times m_k$. For the spatial augmentation, our goal is to enable the model to handle various numbers of channels, including single-channel data. Specifically, we first select a channel number $C_{k'}$ from $C = \{C_1, ..., C_{K'}\}$ ($C_{k'}$ is set equal to C when $C_{k'} > C$), where $C_1 = 1$. Then, we shuffle the channel dimension of the resampled patches $\hat{\boldsymbol{p}}_i$ and the future values $\hat{\boldsymbol{x}}_i^{\text{fut}}$ according to the same rule and the data is reorganized into n non-overlapping subset(s) along the channel dimension with $C_{k'}$ channel(s), where $n = \lfloor C/C_{k'} \rfloor$. In the masking procedure, we randomly mask a subset of $C_{k'} \times L$ patches with a fixed masking ratio, where the values of the masked patches are replaced by zeros. We denote the outputs of data augmentation and masking as $\tilde{\boldsymbol{p}}_i \in \mathbb{R}^{n\times C_{k'}\times L\times P_k}$ and $\tilde{\boldsymbol{x}}_i^{\text{fut}} \in \mathbb{R}^{n\times C_{k'}\times F_k}$.

Input embedding. For neural recordings, the time domain provides insights into the amplitude and duration of neural signals, while the frequency domain unveils oscillatory patterns and rhythmic activity [21]. By modeling neural signals in both domains, we can obtain a more comprehensive understanding of the underlying neurophysiological mechanisms [31]. Therefore, as shown in the top left corner of Fig. 2, we combine the features from both time and frequency domains to obtain the input embedding. We generate the frequency features $\tilde{\boldsymbol{p}}_i^{\text{F}}$ from the augmented data $\tilde{\boldsymbol{p}}_i$ by calculating the power spectral density [52] (PSD) that reveals the spectral power distribution in different frequency bands of the signals, which is associated with different brain functional states [56]. For example, during wakefulness, α (8-13 Hz) and β (13-30 Hz) waves are more active; during sleep, δ (less than 4 Hz) and θ (4-8 Hz) waves are more prominent.

We use non-linear encoders to map the time and frequency data $\tilde{\boldsymbol{p}}_i$, $\tilde{\boldsymbol{p}}_i^{\text{F}}$ to $\frac{D}{2}$-dimensional latent representations $\tilde{\boldsymbol{h}}_i$, $\tilde{\boldsymbol{h}}_i^{\text{F}} \in \mathbb{R}^{n\times C_{k'}\times L\times \frac{D}{2}}$. which are then concatenated and added with a learnable positional encoding $\mathbf{W}_{\text{pos}} \in \mathbb{R}^{L\times D}$ which monitors the temporal order of patches to obtain the input embedding $\boldsymbol{h}_i \in \mathbb{R}^{n\times C_{k'}\times L\times D}$:

$$\boldsymbol{h}_i = \text{Concat}(\tilde{\boldsymbol{h}}_i, \tilde{\boldsymbol{h}}_i^{\text{F}}) + \text{Broadcast}(\mathbf{W}_{\text{pos}}), \tag{1}$$

where the Broadcast$(\cdot)$ operator broadcasts $\mathbf{W}_{\text{pos}}$ to the same shape as Concat$(\tilde{\boldsymbol{h}}_i, \tilde{\boldsymbol{h}}_i^{\text{F}})$. Finally, we make a clone of the input embedding $\boldsymbol{h}_i$ and obtain $\hat{\boldsymbol{h}}_i$, preparing for the subsequent encoding process.

Encoder. In order to generalize to various scenarios, we incorporate both mask-prediction and forecasting tasks during pre-training to learn representations with rich semantic information. For the purpose of simultaneously accomplishing these two pre-training tasks, we design a multi-feed-forward (multi-FFN) Transformer block, as illustrated in the top right corner of Fig. 2. The block contains two FFNs, where one is used for signal reconstruction, and the other is employed for forecasting. We utilize a temporal encoder to capture time dependencies and a spatial encoder to capture channel correlations, both of which are composed of stacked multi-FFN Transformer blocks.

For temporal encoding, we model series of patches of length L: $\boldsymbol{h}_{i,j} \in \boldsymbol{h}_i$, $\hat{\boldsymbol{h}}_{i,j} \in \hat{\boldsymbol{h}}_i$, $j = 1, 2, ..., n \times C_{k'}$, where $\boldsymbol{h}_{i,j}$, $\hat{\boldsymbol{h}}_{i,j} \in \mathbb{R}^{L\times D}$. We denote the outputs of the $l-1$-th layer of the temporal encoder as $\boldsymbol{o}_{i,j}^{l-1}$, $\hat{\boldsymbol{o}}_{i,j}^{l-1} \in \mathbb{R}^{L\times D}$, where $\boldsymbol{o}_{i,j}^{0} = \boldsymbol{h}_{i,j}$, $\hat{\boldsymbol{o}}_{i,j}^{0} = \hat{\boldsymbol{h}}_{i,j}$. For the l'th layer, outputs from the last layer $\boldsymbol{o}_{i,j}^{l-1}$, $\hat{\boldsymbol{o}}_{i,j}^{l-1}$ go through the same multi-head attention [46] followed by a residual addition and a normalization to obtain the attention outputs $\boldsymbol{a}_{i,j}^{l}$, $\hat{\boldsymbol{a}}_{i,j}^{l} \in \mathbb{R}^{L\times D}$,

which will be separately fed into two FFNs (denoted as $\text{FFN}^l_{\text{mask}}$ and $\text{FFN}^l_{\text{fcst}}$). Due to the incomplete data resulting from the mask operation during pre-training, we utilize the information reconstructed by $\text{FFN}^l_{\text{mask}}$ to assist forecasting. Therefore, we establish a residual connection without gradients between the two attention outputs $\boldsymbol{a}^l_{i,j}$ and $\hat{\boldsymbol{a}}^l_{i,j}$:

$$\boldsymbol{f}^l_{i,j} = \text{FFN}^l_{\text{mask}}(\boldsymbol{a}^l_{i,j}), \tag{2}$$

$$\hat{\boldsymbol{f}}^l_{i,j} = \text{FFN}^l_{\text{fcst}}(\hat{\boldsymbol{a}}^l_{i,j} + \text{Detach}(\boldsymbol{a}^l_{i,j}))), \tag{3}$$

where $\text{Detach}(\cdot)$ operator returns a clone of the original input without gradients. Followed by a residual addition and a normalization, we derive the outputs of the l-th layer $\boldsymbol{o}^l_{i,j}$, $\hat{\boldsymbol{o}}^l_{i,j} \in \mathbb{R}^{L\times D}$. The whole outputs of the temporal encoder are denoted as $\boldsymbol{t}_i$, $\hat{\boldsymbol{t}}_i \in \mathbb{R}^{n\times C_{k'}\times L\times D}$. For spatial encoding, we take the outputs of the temporal encoder and model sets of patches of length $C_{k'}$ from different channels: $\boldsymbol{t}_{i,j} \in \boldsymbol{t}_i$, $\hat{\boldsymbol{t}}_{i,j} \in \hat{\boldsymbol{t}}_i$, $j = 1, 2, ..., n \times L$, where $\boldsymbol{t}_{i,j}$, $\hat{\boldsymbol{t}}_{i,j} \in \mathbb{R}^{C_{k'}\times D}$. The encoding process of the spatial encoder is the same as the temporal encoder which is described above. The outputs of the spatial encoder $\boldsymbol{z}_i$, $\hat{\boldsymbol{z}}_i \in \mathbb{R}^{n\times C_{k'}\times L\times D}$ are served as the latent representations of *Brant-2*.

2.2 Pre-training and Fine-tuning

Pre-training. We adopt mask-prediction and forecasting tasks during pre-training to fully extract rich semantic information to adapt to different downstream tasks. The mask-prediction allows the model to understand the patterns within a certain segment of the signal. On the other hand, the forecasting task enables the model to learn future trend changes from the current observed series. We utilize two linear heads $\mathbf{W}_{\text{rec}} \in \mathbb{R}^{D\times P_k}, \mathbf{W}_{\text{fcst}} \in \mathbb{R}^{D\times F_k}$ to map the latent representations to the original signals. During pre-training, the model conducts a patch-level reconstruction and a series-level forecasting:

$$\boldsymbol{p}^{\text{rec}}_i = \boldsymbol{z}_i\mathbf{W}_{\text{rec}}, \tag{4}$$

$$\boldsymbol{x}^{\text{fcst}}_i = \text{MeanPool}(\hat{\boldsymbol{z}}_i)\mathbf{W}_{\text{fcst}}, \tag{5}$$

where $\boldsymbol{p}^{\text{rec}}_i \in \mathbb{R}^{n\times C_{k'}\times L\times P_k}, \boldsymbol{x}^{\text{fcst}}_i \in \mathbb{R}^{n\times C_{k'}\times F_k}$, the $\text{MeanPool}(\cdot)$ operation aggregates each L consecutive patches in $\hat{\boldsymbol{z}}_i$. Then following the masked modeling and forecasting paradigms, *Brant-2* is supervised by two MSE losses in the pre-training stage:

$$\mathcal{L}_{rec} = \sum_{i=1}^{N} \|\tilde{\boldsymbol{p}}_i - \boldsymbol{p}^{\text{rec}}_i\|_2^2, \tag{6}$$

$$\mathcal{L}_{fcst} = \sum_{i=1}^{N} \|\boldsymbol{x}^{\text{fut}}_i - \boldsymbol{x}^{\text{fcst}}_i\|_2^2, \tag{7}$$

where N is the number of training samples. The objective of joint optimization is obtained by adding the losses $\mathcal{L}_{rec}$ and $\mathcal{L}_{fcst}$.

Fine-tuning. When fine-tuning the model, we first use a mean pooling operation to gather each L consecutive patches of the latent representations $\boldsymbol{z}_i$, $\hat{\boldsymbol{z}}_i \in \mathbb{R}^{n\times C_{k'}\times L\times D}$, then the representations are aggregated by a weighted sum:

$$\boldsymbol{r}_i = \lambda\text{MeanPool}(\boldsymbol{z}_i) + (1-\lambda)\text{MeanPool}(\hat{\boldsymbol{z}}_i), \tag{8}$$

where $\boldsymbol{r}_i \in \mathbb{R}^{n\times C_{k'}\times D}$ and λ is a learnable parameter. The aggregated representation $\boldsymbol{r}_i$ will be fed into a linear or non-linear head for the downstream tasks.

3 EXPERIMENTS

3.1 Pre-training Setup

Pre-training datasets. The pre-training corpus of *Brant-2* incorporates a mixed dataset of SEEG and EEG data from over 15k subjects with a total data size approaching 4 TB. The SEEG corpus used for pre-training contains intracranial neural data recorded from 26 subjects. The original corpus has a total size of 12 TB. After removing unused channels and applying preprocessing like denoising and filtering, we obtain 2.3 TB of SEEG data for pre-training. The surgical procedure involves the implantation of invasive electrodes with 47 to 238 channels. The corpus contains SEEG data at 1000Hz, 2000Hz, and 4000Hz sampling rates. The EEG corpus utilized in the pre-training is a publicly available dataset, TUEG[19], which comprises 1,643 GB of clinical recordings from 14,987 individuals with a total of 27,063 hours of data. The dataset contains over 40 different channel configurations, in which approximately 95% of the data includes a 10/20 configuration as a subset of the available channels. The sampling rate of the recordings varies between 250Hz and 1024Hz.

Pre-training details. In the encoder block of *Brant-2*, we apply RMSNorm[55] and use the Swish activation function[38]. We set the context length L of *Brant-2* as 16 patches, the masking ratio as 40%, and the forecasting length as 1/4 of the context length (The hyperparameter analysis of the masking ratio and forecasting length is shown in App.B). The adjustment factor of the sampling rate is uniformly chosen from $\mathcal{M} = \{0.25, 0.5, 1.0, 2.0\}$ and the reorganized channel number is chosen from $C = \{1, 2, 4, 8, 16, 32, 64, 128\}$. *Brant-2* is trained using the AdamW optimizer[28], with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $eps = 10^{-5}$. For the learning rate scheduling, we use a linear warmup of 1k steps to reach a peak learning rate of 1.0×10^{-5}, followed by a cosine decay of 150k steps to decay the final learning rate to 0. During the pre-training process, the model parameters are updated for a total of 105k steps. Our models are trained on a Linux system with 2 CPUs (AMD EPYC 9654 96-Core Processor) and 4 GPUs (NVIDIA Tesla A100 80G). *Brant-2* contains 1 billion parameters, which takes over 100 hours to pre-train.

3.2 Evaluation Setup

We conduct evaluation experiments on nine diverse SEEG and EEG datasets, encompassing five downstream tasks: seizure detection, seizure prediction, sleep stage classification, emotion recognition, and motor imagery classification. We divide each dataset into several non-overlapping groups and conduct n-fold cross-validation for all groups. Each time, we set one group for evaluation and the others for fine-tuning. The fine-tuning process mainly involves updating the last two layers of the temporal encoder and the classification head, while freezing the remaining parameters of the *Brant-2* (more details are shown in App. D.2).

Seizure detection. Accurate seizure detection is crucial for diagnosing and treating individuals with epilepsy and other seizure-related disorders. Seizure detection aims to identify and classify instances of seizures in brain signals recorded from epilepsy patients,

which is formalized as a binary classification to classify between physiological and pathological samples. We employ 2 SEEG and 2 EEG datasets to evaluate the model performance on seizure detection. The SEEG datasets, MAYO and FNUSA [32], contain 5000Hz SEEG recordings from 18 and 13 subjects, respectively. The subjects are divided into six groups for each dataset, and the data is segmented into 3-second data clips. We preserve the physiological and pathological activities and remove the artifacts and power line noise. CHB-MIT [18, 40] consists of 23-channel EEG recordings with a sampling rate of 256Hz from 22 subjects with intractable seizures. We divide the subjects into four groups and segment the signals into 8-second data clips. Siena [11, 12] consists of 27-channel EEG recordings of 14 patients with a sampling rate of 512Hz. The subjects are divided into five groups, and the signals are segmented into 4-second data clips. We use precision, recall, F1 and F2 scores as evaluation metrics. In the scenario of epilepsy, F2 is more valued than F1 since ignoring any seizure is costly in diagnosis.

Seizure prediction. Different from seizure detection, seizure prediction is conducted under a more challenging setting where the task is to predict the likelihood of future seizures based on the current observations. Seizure prediction is crucial for providing early warnings and alerts for individuals with epilepsy. We utilize a clinical SEEG dataset from a first-class hospital labeled by professional neurosurgeons. The dataset contains 5 subjects with a sampling rate of 1000Hz, and we adopt a 5-fold cross-validation. We sample 16-second segments and predict whether a seizure occurs within the next 1 minute. We adopt the same evaluation metrics as in seizure detection (i.e., precision, recall, F1 and F2 score) to measure the performance of the models.

Sleep stage classification. In sleep health research, sleep staging plays a critical role in enhancing our understanding of sleep states and patterns, contributing to the prevention and diagnosis of sleep-related disorders [35]. The American Academy of Sleep Medicine (AASM) manual defines sleep into five stages: wake, N1, N2, N3, and REM [5]. Thus, sleep stage classification is formalized as a 5-class classification. We choose 2 EEG datasets, SleepEDF [23] and Haaglanden Medisch Centrum sleep staging database [2, 3](referred to as HMC), to verify the model performance on sleep stage classification. For SleepEDF, we adopt the SleepEDF-78 dataset which contains 153 whole-night polysomnographic sleep recordings from 78 subjects during sleep cassette studies. The EEG data with a sampling rate of 100Hz contains 1 EEG channel, which is segmented into 30-second epochs. We randomly divide the subjects into five groups. HMC is a collection of 151 whole-night polysomnographic (PSG) sleep recordings from 151 subjects. The data is sampled at 256Hz and contains 4 EEG channels. The subjects are split into five groups, and the signals are segmented into 30-second epochs. As for the evaluation metrics, we utilize accuracy, sensitivity, specificity, macro F1 score, and Cohen's kappa κ.

Emotion recognition. Emotion recognition using EEG is becoming an interesting topic among researchers, which has made advancements in various domains, including biomedical research, brain-computer interfaces (BCIs), etc [22]. The SEED dataset [15, 60] contains 62-channel EEG data from 15 subjects while watching film clips. The film clips are carefully selected to induce different types of emotion (positive, negative, and neutral). Thus, we conduct discrete emotion recognition formalized as a 3-class classification. The data is down-sampled to 200Hz, segmented into 5-second segments, and split into five groups. The evaluation metrics include accuracy and macro F1 score.

Motor imagery classification. Motor imagery classification is to classify brain activity patterns related to imagined movements, which has gained significant attention due to its potential applications in BCIs, rehabilitation therapies, and assistive technologies. We select EEG Motor Movement/Imagery [17, 39](referred to as Motor Imagery) as the dataset for this task. Motor Imagery consists of over 1500 one- and two-minute 64-channel EEG recordings with a 160Hz sampling rate obtained from 109 volunteers. For each subject, a target appears on either the left or the right side of the screen, and the subject imagines opening and closing the corresponding fist until the target disappears. The model aims to differentiate whether the subject is imagining opening and closing their left fist or right fist based on the collected EEG signals. We split the subjects into five groups and clip the data into 6.4-second segments. The evaluation metrics include accuracy and F1 score.

3.3 Baselines

We extensively compare our model with 12 advanced methods, which are divided into three categories, including 1) 3 methods aimed at time series universal modeling, 2) 3 methods based on self-supervised pre-training on brain signals, and 3) 6 methods specifically designed for each task. The methods in the first two categories are evaluated on all downstream tasks, while the methods in the third category are only evaluated on the specific tasks.

To be precise, we choose TF-C [58] SimMTM [14] and One Fits All [61] as universal modeling methods for time series. For the pre-training works on brain signals, we choose BrainBERT [47], Brant [56] and MBrain [7]. In addition, we select PPi [54] (seizure detection on SEEG data), ScatterFormer [59] (seizure detection on EEG data), Lopes et al. [27] (seizure prediction), SleepHGNN [20] (sleep stage classification), EEG Conformer [42] (emotion recognition) and TSFF-Net [29] (motor imagery classification) as the task-specific methods for each downstream task. More details of the baselines are shown in App. C.

3.4 Evaluation Results

Main Results. Fig. 3 summarizes the overall results of *Brant-2* compared with the baseline methods on all the downstream tasks. The radar chart shows that *Brant-2* outperforms all universal time series modeling methods and pre-training methods on brain signals, surpassing a majority of task-specific methods, indicating that our method exhibits strong generalization ability across various scenarios of brain signals. Detailed statistics and comparisons of each task will be discussed in the following paragraphs, where in all the tables, we mark values ranking the first (**v**), second (<u>v</u>), and third (v*) in each column.

Seizure Detection. Tab. 1 shows the results of seizure detection on SEEG and EEG datasets. In the results of MAYO and FNUSA, *Brant-2* achieves the best recall and F2 score over other models, demonstrating our model's ability in seizure detection on SEEG data. Regarding the F2 score, PPi secures the second position, which can

Table 1: Average performance on seizure detection.

Methods \ Metrics	MAYO				FNUSA			
	Pre.	Rec.	F1	F2	Pre.	Rec.	F1	F2
TF-C[58]	55.12 ±10.37	70.72 ±14.49	55.84 ±8.89	61.48 ±9.43	69.80±4.91	81.46 ±7.63	74.76 ±6.30	78.50 ±8.16
SimMTM[14]	51.03 ±10.75	78.97 ±9.31	56.66 ±10.44	67.13 ±10.04	69.33*±5.71	82.44 ±12.05	74.74 ±6.03	79.00 ±9.39
One Fits All[61]	59.60±11.21	77.62 ±10.71	66.26*±3.23	72.22 ±4.02	63.65 ±11.03	82.18 ±9.22	70.90 ±6.00	76.95 ±5.02
BrainBERT[47]	54.47 ±11.25	89.70*±11.43	64.65 ±7.54	76.02 ±6.31	58.11 ±8.60	87.62*±8.79	67.91 ±6.46	77.67 ±4.95
Brant [56]	49.24 ±12.56	93.30±7.84	63.38 ±8.22	77.92*±2.36	48.15 ±5.13	94.72±4.94	63.61 ±3.72	79.07 ±3.31
MBrain[7]	55.39 ±10.46	80.08 ±10.97	61.81 ±4.55	70.32 ±10.12	67.88 ±5.43	85.89 ±9.51	75.25*±6.78	81.10*±10.68
PPi[54]	**68.94**±12.14	88.73 ±8.49	**73.77**±6.24	80.74±4.63	**72.77**±5.06	86.60 ±13.67	78.39±8.29	82.93±9.49
Brant-2	55.41*±9.74	**95.88**±3.86	69.90±8.45	**83.30**±6.07	68.78 ±9.86	**95.96**±4.43	**79.76**±7.61	**88.59**±5.21
Methods \ Metrics	CHB-MIT				Siena			
	Pre.	Rec.	F1	F2	Pre.	Rec.	F1	F2
TF-C[58]	17.82 ±8.44	61.36 ±10.66	18.01 ±6.72	27.61 ±6.52	7.98 ±1.34	61.63 ±10.92	13.68 ±3.20	24.92 ±3.78
SimMTM[14]	54.31 ±8.12	36.77 ±8.71	42.82 ±8.95	38.84 ±8.79	32.30 ±10.51	43.37 ±11.72	27.47 ±6.10	32.97 ±9.35
One Fits All[61]	51.14 ±12.20	56.58 ±5.23	51.21 ±9.95	53.69 ±5.47	47.27 ±10.26	43.68 ±8.55	43.00 ±5.42	42.97 ±5.29
BrainBERT[47]	52.47 ±8.04	59.44 ±12.25	55.09 ±8.95	57.41 ±10.06	47.33*±9.81	60.61 ±10.15	48.54*±6.20	53.52 ±6.92
Brant [56]	**57.42**±8.42	61.41*±10.07	57.99*±7.63	59.65*±8.09	43.07 ±11.05	66.04*±6.81	48.36 ±8.42	55.27*±4.82
MBrain[7]	53.39 ±14.42	52.87 ±8.08	49.35 ±7.40	50.38 ±6.18	38.94 ±8.83	60.99 ±5.61	45.88 ±7.06	53.32 ±6.62
ScatterFormer[59]	55.73*±8.03	64.92±10.14	59.81±8.80	62.72*±9.52	50.33±11.38	67.49±7.21	53.49±7.04	58.98±4.15
Brant-2	56.44±7.53	**73.04**±7.83	**63.42**±7.48	**68.78**±7.65	**50.35**±8.52	**70.42**±4.53	**58.14**±4.77	**64.70**±1.92

Table 2: Average performance on sleep stage classification.

Methods \ Metrics	SleepEDFx					HMC				
	Acc.	Sens.	Spec.	Macro F1	Kappa	Acc.	Sens.	Spec.	Macro F1	Kappa
TF-C[58]	65.96 ±1.28	51.42 ±2.37	90.26 ±0.47	49.37 ±1.92	50.73 ±2.04	48.04 ±4.89	35.30 ±4.67	84.35 ±1.31	30.15 ±6.67	23.90 ±7.08
SimMTM[14]	63.85 ±2.30	36.29 ±1.54	88.97 ±0.68	31.37 ±2.62	44.40 ±3.54	44.64 ±2.01	31.52 ±2.48	83.18 ±0.76	27.27 ±3.51	17.76 ±3.90
One Fits All[61]	68.45 ±1.95	56.32 ±3.04	91.20 ±0.71	54.77 ±1.70	55.03 ±2.75	58.64 ±1.55	51.12 ±2.85	88.42 ±0.70	50.54 ±2.97	43.52 ±3.00
BrainBERT[47]	69.56 ±1.85	59.40*±2.40	91.80 ±0.55	58.66*±1.66	57.13 ±2.51	60.69 ±1.67	53.06 ±2.13	89.04 ±0.61	51.95*±2.01	46.48 ±2.58
Brant [56]	69.06 ±2.69	58.25 ±3.62	91.63 ±0.83	56.84 ±3.32	56.55 ±3.77	51.02 ±3.15	41.90 ±2.20	85.89 ±0.60	38.12 ±3.74	30.42 ±2.90
MBrain[7]	71.91*±0.98	58.35 ±1.99	92.16*±0.38	58.22 ±3.63	59.83*±2.10	62.33*±1.24	53.97*±1.87	89.50*±0.14	51.48 ±2.94	48.65*±0.88
SleepHGNN[20]	**77.56**±2.06	**70.38**±2.57	**94.18**±0.60	**69.79**±3.98	**69.72**±3.79	64.87±2.34	57.27±2.48	90.24±0.56	56.93±3.50	52.21±3.56
Brant-2	77.15±1.39	67.01±2.51	93.90±0.43	67.20±2.42	68.05±2.22	**68.76**±2.41	**63.74**±1.74	**91.52**±0.43	**63.87**±1.94	**58.20**±2.65

be attributed to the fact that PPi contains a specifically designed pre-training framework for seizure detection and dedicated techniques to address inter-subject variability. From the results of CHB-MIT and Siena, we can observe that *Brant-2* ranks the first in almost all performance metrics, showing a strong ability in EEG-based seizure detection.

Seizure Prediction. Tab. 3 shows the average performance of seizure prediction task on the clinical dataset. In Tab. 3, *Brant-2* achieves the first place in terms of F1 and F2 scores, which are improved by 37.97% and 32.02% compared to Brant, indicating that the pre-training forecasting task enhances the predictive capability. Lopes et al. [27] achieves the second-best F1 and F2 scores, demonstrating the effectiveness of combining original signals and handcrafted features. However, as a fully supervised method, Lopes et al. [27] relies heavily on labels, which are often scarce in clinical settings. We will investigate the impact of scarce labels on model performance in Sec. 4.3. Apart from *Brant-2* and Lopes et al. [27], One Fits All achieves the highest F2 score, which could be attributed to its utilization of a pre-trained GPT-2 [37] as the backbone with strong predictive abilities.

Sleep Stage Classification. The results of sleep stage classification on SleepEDFx and HMC are shown in Tab. 2. Overall, *Brant-2* and SleepHGNN exhibit comparable performance, with *Brant-2* outperforming SleepHGNN on HMC and SleepHGNN having a slight edge on SleepEDFx. As a specialized model designed for

* Here we calculate the relative improvement.

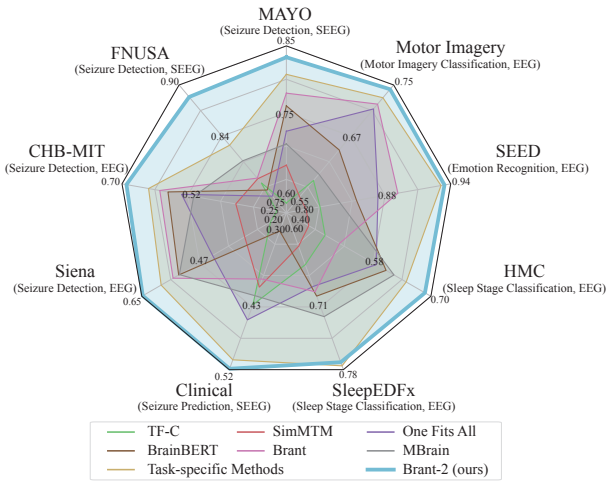

Figure 3: The overall performance comparison of our model and other baseline methods on all downstream datasets.

Table 3: Average performance on seizure prediction.

Methods \ Metrics	Clinical			
	Pre.	Rec.	F1	F2
TF-C[58]	34.92 ±6.61	49.35±13.53	37.28 ±9.50	42.88 ±11.10
SimMTM[14]	60.39 ±10.76	37.94 ±6.72	45.37 ±7.69	40.42 ±6.99
One Fits All[61]	55.84 ±8.48	43.41 ±10.39	48.04*±9.42	45.01*±10.02
BrainBERT[47]	61.89*±13.92	29.38 ±12.47	39.05 ±13.98	32.55 ±13.05
Brant [56]	55.39 ±11.28	40.89 ±9.97	40.01 ±8.80	39.29 ±9.29
MBrain[7]	54.75 ±13.29	44.26 ±11.86	41.79 ±9.59	41.67 ±9.82
Lopes et al. [27]	**62.82**±9.17	46.86*±10.21	50.84±8.46	47.85±10.36
Brant-2	62.67±7.25	**49.94**±8.04	**55.20**±7.93	**51.87**±8.05

Table 4: Average performance on emotion recognition.

Methods \ Metrics	SEED	
	Acc.	Macro F1
TF-C[58]	82.87 ±5.21	82.13 ±5.66
SimMTM[14]	81.69 ±7.06	81.26 ±7.20
One Fits All[61]	87.80 ±3.35	87.67 ±3.38
BrainBERT[47]	85.98 ±5.46	85.81 ±5.98
Brant [56]	89.50*±3.57	89.43*±3.71
MBrain[7]	84.60 ±7.47	84.52 ±7.45
EEG Conformer[42]	93.17±4.20	93.10±4.22
Brant-2	**93.47**±3.09	**93.42**±3.08

sleep stage classification, SleepHGNN incorporates EEG signals and synchronously collected EOG, ECG, and EMG signals, thereby leveraging multiple modalities for improved performance. *Brant-2* achieves comparable performance using only EEG signals, highlighting the effectiveness of our large-scale pre-training.

Table 5: Average performance on motor imagery classification.

Methods \ Metrics	Motor Imagery	
	Acc.	F1
TF-C[58]	60.06 ±1.62	57.79 ±3.00
SimMTM[14]	57.48 ±1.82	57.37 ±2.80
One Fits All[61]	71.25 ±3.50	72.56*±3.06
BrainBERT[47]	64.84 ±4.19	70.32 ±3.55
Brant [56]	72.00*±1.93	71.84 ±2.42
MBrain[7]	61.06 ±2.09	60.42 ±4.65
TSFF-Net[29]	73.00±4.32	73.87±2.10
Brant-2	**74.33**±3.61	**74.30**±3.83

Emotion Recognition. Tab. 4 contains the results of emotion recognition on SEED dataset. Our model obtains the best results and EEG Conformer achieves the second place. Like *Brant-2*, EEG Conformer also considers the temporal dependency and spatial correlations by designing a convolution module with temporal and spatial convolutional layers.

Motor Imagery Classification. The performance of motor imagery classification is shown in Tab. 5. The achievement of the highest accuracy and F1 score by *Brant-2* demonstrates the effectiveness of our model in motor imagery classification. The utilization of time-frequency spectrograms in TSFF-Net, along with its second-best accuracy and F1 score, highlights the significance of time-frequency domain information in this scenario.

4 ANALYSIS

4.1 Scalability Analysis

Large language and vision models[4, 44, 45, 50] have shown strong scalability behavior. As a large model in brain signals, we aim to investigate the scaling behavior of our model in terms of the pre-training loss and downstream task performance.

Setup. In addition to *Brant-2*, we pre-trained two smaller versions with 200 million, 460 million parameters, following the same training configurations described in Sec.3.1. Then we evaluate the models on all five downstream tasks, with each task utilizing one dataset (CHB-MIT for seizure detection, Clinical for seizure prediction, SleepEDFx for sleep stage classification, SEED for emotion recognition and Motor Imagery for motor imagery classification).

Results. The training losses of the two pre-training objectives (the losses are calculated every 5k steps) are shown in Fig.4(a). One can observe that as training progresses: 1) the training losses of the models, regardless of their size, continue to decrease; 2) as we increase the model size, the losses decrease faster. These observations indicate that *Brant-2* shows scalability behavior during pre-training. Furthermore, as shown in Fig.4(b), larger models attain better performance across all tasks, showcasing that our scalable overall performance transfers to a range of downstream tasks.

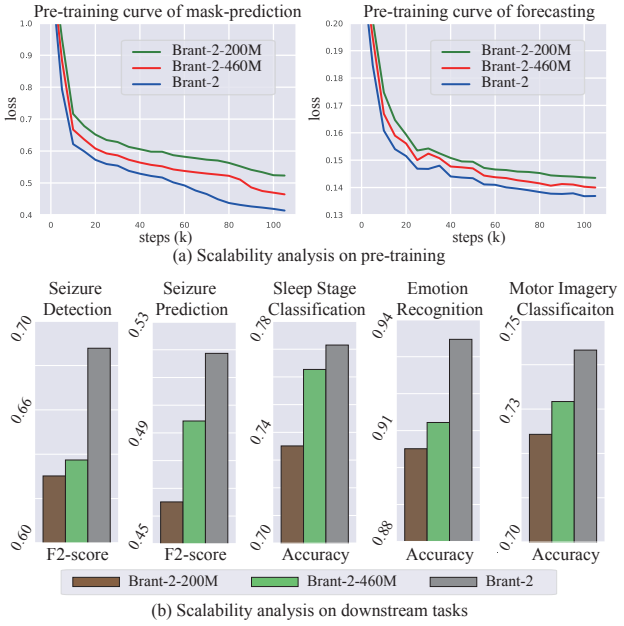

Figure 4: The results of scalability analysis.

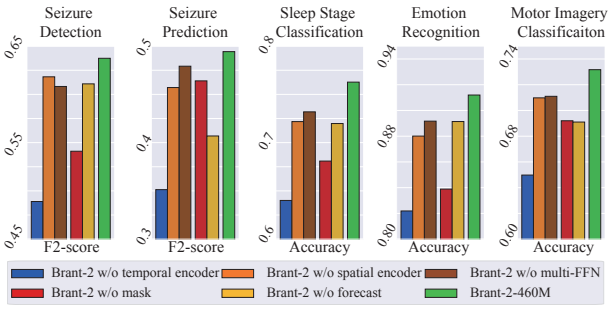

Figure 5: The results of ablation study.

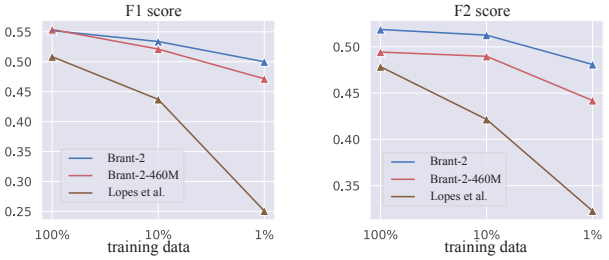

Figure 6: the performance changes of the model as the training labels decrease.

4.2 Ablation Study

We perform ablation experiments to assess the effectiveness of the architectural design of the model and the pre-training tasks.

Setup. We set three model variants to validate the effectiveness of our architectural design: 1) *Brant-2* w/o temporal encoder: remove the temporal encoder; 2) *Brant-2* w/o spatial encoder: remove the spatial encoder; 3) *Brant-2* w/o multi-FFN: replace the multi-FFN Transformer encoder block of *Brant-2* with the vanilla Transformer encoder block [46]. For each model variant, we control the parameter count of the models to be approximately the same to ensure fair comparison. To illustrate the usefulness of both pre-training tasks, we perform two sets of experiments: 4) *Brant-2* w/o mask: pre-train with forecasting; 5) *Brant-2* w/o forecast: pre-train with mask-prediction. *Brant-2* and the above five variants are evaluated on all the five downstream tasks, with each task utilizing the same dataset as the one used in the scalability analysis (i.e., CHB-MIT for seizure detection, Clinical for seizure prediction, SleepEDFx for sleep stage classification, SEED for emotion recognition and Motor Imagery for motor imagery classification). Since each model variant requires pre-training and such process for a 1-billion scale model alone takes over 100 hours as described in Sec. 3.1, all the experiments in the ablation study are based on *Brant-2*-460M.

Results. The ablation results are shown in Fig.5, in which *Brant-2* outperforms the other variants across all five downstream tasks, demonstrating the effectiveness of each component of our work. *Brant-2* w/o temporal encoder exhibits overall poor performance in these downstream tasks, highlighting the crucial importance of temporal dependency for brain signals. In certain tasks (e.g., seizure detection, emotion recognition), *Brant-2* w/o mask outperforms *Brant-2* w/o forecast, indicating that these tasks require a better understanding of patterns within a signal segment. On the other hand, in some tasks (e.g., seizure prediction), *Brant-2* w/o forecast performs better, demonstrating that these tasks prioritize predicting future changes based on the current observed series. Therefore, joint training of the two pre-training tasks enhances the adaptive ability to different downstream tasks.

4.3 Label Scarcity Scenario Exploration

The results in Sec.3.4 have demonstrated that *Brant-2* can generalize well to various tasks. As a foundation model, we also aim to investigate whether our model can address the issue of over-reliance on labels and be applicable to scenarios with scarce labels.

Setup. We choose to conduct experiments on the Clinical dataset originating from real-world clinical scenario of epilepsy, where the annotation cost is high. By choosing this dataset, we intend to simulate real-world scenarios closely and address the challenges associated with expensive annotations in clinical settings. We compare our model with the best-performed baseline method Lopes et al. [27], which is fully supervised. We conduct three sets of experiments on each model with 100%, 10%, and 1% of training data.

Results. The variation in model performance with decreasing training labels is shown in Fig.6. Overall, as the training labels decrease, the performance exhibits a certain degree of decline. When transitioning from 100% to 1% labels, *Brant-2* and *Brant-2*-460M show F1 and F2 scores decreases of less than 10% and 15%, respectively, In contrast, the F1 and F2 scores of Lopes et al. [27] decline 50.6% and 32.6%, respectively. The results indicate that *Brant-2* can reduce reliance on labels, thereby ensuring performance in scenarios with scarce labels.

5 CONCLUSION

We propose a foundation model *Brant-2*, the first off-the-shelf model that can be applied to scenarios of both SEEG and EEG. *Brant-2* is able to handle significant data variations and generate powerful representations of brain signals from a broad range of application scenarios. We experiment on five downstream tasks to illustrate the generalization ability of *Brant-2*. In addition, *Brant-2* shows a scalability behavior in both pre-training and downstream tasks. Furthermore, we explore the change of model performance in low-resource labeled scenarios, in which the performance of *Brant-2* remains much more stable than the supervised SOTA method designed for the scenario, indicating that our model alleviates the issue of label efficiency. The field of brain signals is continuously evolving, with emerging research directions and scenarios. In the future, we aim to train our model on a more diverse and extensive corpus, enabling its application in more research areas and scenarios.

REFERENCES

[1] Fahd A. Alturki, Khalil AlSharabi, Akram M. Abdurraqeeb, and Majid Aljalal. 2020. EEG Signal Analysis for Diagnosing Neurological Disorders Using Discrete Wavelet Transform and Intelligent Techniques. *Sensors* 20 (2020).

[2] Diego Alvarez-Estevez and Roselyne Rijsman. 2022. Haaglanden Medisch Centrum sleep staging database (version 1.1). https://doi.org/10.13026/t79q-fr32.

[3] Diego Alvarez-Estevez and Roselyne M Rijsman. 2021. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one* 16 (2021).

[4] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. 2023. Sequential Modeling Enables Scalable Learning for Large Vision Models. arXiv:2312.00785 [cs.CV]

[5] Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. 2017. AASM scoring manual updates for 2017 (version 2.4).

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG]

[7] Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. 2023. MBrain: A Multi-Channel Self-Supervised Learning Framework for Brain Signals. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[8] Vairis Caune, Juris Zagars, and Radu Ranta. 2012. EEG/SEEG Signal Modelling using Frequency and Fractal Analysis.. In *BIOSIGNALS*.

[9] Junru Chen, Yang Yang, Tao Yu, Yingying Fan, Xiaolong Mo, and Carl Yang. 2022. 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*.

[10] Debjani Dasgupta, Anna Miserocchi, Andrew W McEvoy, and John S Duncan. 2022. Previous, current, and future stereotactic EEG techniques for localising epileptic foci. *Expert Review of Medical Devices* 19 (2022), 571–580.

[11] Paolo Detti. 2020. Siena Scalp EEG Database (version 1.0.0). https://doi.org/10.13026/5d4a-j060.

[12] Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. 2020. EEG Synchronization Analysis for Seizure Prediction: A Study on Data of Noninvasive Recordings. *Processes* 8 (2020).

[13] Marina Diachenko, Simon J Houtman, Erika L Juarez-Martinez, Jennifer R Ramautar, Robin Weiler, Huibert D Mansvelder, Hilgo Bruining, Peter Bloem, and Klaus Linkenkaer-Hansen. 2022. Improved manual annotation of EEG signals through convolutional neural network guidance. *Eneuro* 9 (2022).

[14] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. 2023. SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[15] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*.

[16] Jay Gavvala, Muhammad Zafar, Saurabh R. Sinha, Giridhar Kalamangalam, and Stephan Schuele. 2022. Stereotactic EEG Practices: A Survey of United States Tertiary Referral Epilepsy Centers. *Journal of Clinical Neurophysiology* 39 (2022).

[17] Ary L. Goldberger, Luís A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, ..., and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101 (2000).

[18] John Guttag. 2010. CHB-MIT Scalp EEG Database. PhysioNet. https://doi.org/10.13026/C2K01R

[19] A Harati, S Lopez, I Obeid, J Picone, MP Jacobson, and S Tobochnik. 2014. The TUH EEG CORPUS: A big data resource for automated EEG interpretation. In *2014 IEEE signal processing in medicine and biology symposium (SPMB)*.

[20] Ziyu Jia, Youfang Lin, Yuhan Zhou, Xiyang Cai, Peng Zheng, Qiang Li, and Jing Wang. 2023. Exploiting Interactivity and Heterogeneity for Sleep Stage Classification Via Heterogeneous Graph Neural Network. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[21] M K Kalaivani, V. Kalaivani, V. Anusuya Devi, Ismail M Gursoy, André L. V. Coelho, Clodoaldo A. M. Lima, André L. V. Coelho, Deng Wang, Duoqian Miao, Kai-Cheng Hsu andSung Nien, Reza Boostani, and Ahmad Ghanizadeh. 2014. Analysis of EEG Signal for the Detection of Brain Abnormalities.

[22] Kranti Kamble and Joydeep Sengupta. 2023. A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals. *Multimedia Tools and Applications* (2023), 1–36.

[23] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering* 47 (2000).

[24] Sándor Kovács, Márton Tóth, József Janszky, Tamás Dóczi, Dániel Fabó, István Boncz, Lajos Botz, and Antal Zemplényi. 2021. Cost-effectiveness analysis of invasive EEG monitoring in drug-resistant epilepsy. *Epilepsy & Behavior* 114 (2021).

[25] Alycia Lee, Brando Miranda, Sudharsan Sundar, and Sanmi Koyejo. 2023. Beyond Scale: the Diversity Coefficient as a Data Quality Metric Demonstrates LLMs are Pre-trained on Formally Diverse Data. arXiv:2306.13840 [cs.CL]

[26] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. arXiv:2308.12032 [cs.CL]

[27] Fábio Lopes, Adriana Leal, Mauro F. Pinto, António Dourado, Andreas Schulze-Bonhage, Matthias Dümpelmann, and César Teixeira. 2023. Removing artefacts and periodically retraining improve performance of neural network-based seizure prediction models. *Scientific Reports* (2023).

[28] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* abs/1711.05101 (2017).

[29] Zhengqing Miao and Meirong Zhao. 2023. Time-space-frequency feature Fusion for 3-channel motor imagery classification. *arXiv preprint arXiv:2304.01461* abs/2304.01461 (2023).

[30] Zhengqing Miao and Meirong Zhao. 2023. Time-space-frequency feature Fusion for 3-channel motor imagery classification. arXiv:2304.01461 [cs.LG]

[31] Santiago Morales and Maureen Bowers. 2022. Time-frequency analysis methods and their application in developmental EEG data. *Developmental Cognitive Neuroscience* 54 (2022).

[32] Petr Nejedly, Vaclav Kremen, Vladimir Sladky, Jan Cimbalnik, Petr Klimes, Filip Plesinger, Filip Mivalt, Vojtech Travnicek, Ivo Viscor, Martin Pail, et al. 2020. Multicenter intracranial EEG dataset for classification of graphoelements and artifactual signals. *Scientific data* 7 (2020).

[33] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*.

[34] Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2021. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2021).

[35] Huy Phan and Kaare Mikkelsen. 2022. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement* 43 (2022).

[36] Huy P Phan, Kristian P. Lorenzen, Elisabeth Roxane Marie Heremans, Oliver Y. Ch'en, Minh C. Tran, Philipp Koch, Alfred Mertins, Mathias Baumert, Kaare B. Mikkelsen, and Marina De Vos. 2023. L-SeqSleepNet: Whole-cycle Long Sequence Modeling for Automatic Sleep Staging. *IEEE Journal of Biomedical and Health Informatics* 27 (2023).

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1 (2019).

[38] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2017. Searching for Activation Functions. *CoRR* abs/1710.05941 (2017).

[39] Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. 2004. BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. *IEEE Transactions on Biomedical Engineering* 51 (2004).

[40] Ali Hossam Shoeb. 2009. *Application of machine learning to epileptic seizure onset detection and treatment.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[41] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2020. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing* 11 (2020).

[42] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. 2023. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization.

IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2023).

[43] Akara Supratak and Yike Guo. 2020. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*.

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[47] Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. 2022. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*.

[48] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. 2023. InternImage: Exploring Large-Scale Vision Foundation Models With Deformable Convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[49] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*.

[50] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open Large-scale Language Models. arXiv:2309.10305 [cs.CL]

[51] Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. 2023. Learning Topology-Agnostic EEG Representations with Geometry-Aware Modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[52] Richard N. Youngworth, Benjamin B. Gallagher, and Brian L. Stamper. 2005. An overview of power spectral density (PSD) calculations. In *Optical Manufacturing and Testing VI*.

[53] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A New Foundation Model for Computer Vision. *CoRR* abs/2111.11432 (2021).

[54] Zhizhang Yuan, Daoze Zhang, Yang Yang, Junru Chen, and Yafeng Li. 2023. PPi: Pretraining Brain Signal Model for Patient-independent Seizure Detection. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[55] Biao Zhang and Rico Sennrich. 2019. Root Mean Square Layer Normalization. *CoRR* abs/1910.07467 (2019).

[56] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. 2023. Brant: Foundation Model for Intracranial Neural Signal. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[57] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. McAlpine, and Y. Zhang. 2021. A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *Journal of Neural Engineering* 18 (2021).

[58] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems* 35 (2022).

[59] Ruizhe Zheng, Jun Li, Yi Wang, Tian Luo, and Yuguo Yu. 2023. ScatterFormer: Locally-Invariant Scattering Transformer for Patient-Independent Multispectral Detection of Epileptiform Discharges. arXiv:2304.14919 [eess.SP]

[60] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development* 7, 3 (2015).

[61] Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Thirty-seventh Conference on Neural Information Processing Systems*.

ETHICS STATEMENT

The data collection and experiments conducted in our work on the private datasets (i.e., the pre-training SEEG corpus and the clinical dataset for seizure prediction) have been approved by the Institutional Review Board (IRB) and passed ethical review. All participants have signed informed consent forms. All publicly available datasets used in this paper are not associated with any privacy or security concerns. Furthermore, we have followed guidelines on responsible use specified by the primary authors of the datasets used in the current work.

A RELATED WORK

Scenario-specific methods for brain signals. In view of the diverse application scenarios of brain signals, researchers have designed various methods specifically tailored for these contexts. Yuan et al. [54] propose a self-supervised learning framework and two techniques for SEEG-based patient-independent seizure detection. Zheng et al. [59] propose a model, ScatterFormer, for patient-independent seizure detection on EEG data, which is an invariant scattering transform-based hierarchical Transformer that specifically pays attention to subtle features. Lopes et al. [27] conduct seizure prediction by encoding the original signals and hand-crafted features with a deep and shallow networks, respectively. Jia et al. [20] propose a novel Sleep Heterogeneous Graph Neural Network (SleepHGNN) to capture the heterogeneity and the interactivity of physiological signals for sleep stage classification. Song et al. [42] design a compact Convolutional Transformer, named EEG Conformer, to encapsulate local and global features for emotional recognition and motor imagery classification. Miao and Zhao [30] propose a shallow, lightweight decoding architecture (TSFF-img) based on time-frequency spectrograms for motor imagery classification.

Universal modeling for brain signals. Universal modeling exhibits great advantages by learning highly general representations to enable customization for various applications. Developing such techniques for brain signals with a broad range of applications is suitable. Wang et al. [47] propose an off-the-shelf model, BrainBERT, that provides embeddings for intracranial recordings. Zhang et al. [56] propose a foundation model Brant for SEEG modeling, which is the largest model for intracranial recordings. Both BrainBERT and Brant are limited to SEEG data with a relative narrow range of application scenarios. Cai et al. [7] design a unified self-supervised learning framework for brain signals which can be utilized on either SEEG or EEG data. However, their work cannot model different kinds of brain signals simultaneously.

B HYPERPARAMETER ANALYSIS

The pre-training performance of a foundation model is of utmost importance as it significantly impacts the performance on downstream tasks. Therefore, it is necessary to select the optimal masking ratio and forecasting length which are two crucial hyperparameters of *Brant-2*. However, due to limited computational resources, it is impractical to perform a grid search for these two parameters on a model with over 1 billion parameters (as the pre-training time for *Brant-2* exceeds 100 hours). Therefore, we conduct experiments on two smaller-scale models (100M and 200M) to observe their performance on downstream tasks and determine the optimal hyperparameters, which are then applied to larger-scale models. For the searching strategy, since the two pre-training tasks are relatively independent, we adopted the following strategy to determine the two optimal hyperparameters instead of grid search: First, we solely focus on the mask-prediction task during pre-training and select the optimal masking ratio. Next, with the chosen masking ratio fixed, we determine the optimal forecasting length.

Setup. To comprehensively evaluate the model to determine the optimal hyperparameters, we pre-trained two models with 100M and 200M parameters and conducted experiments on all five downstream tasks. Each task is evaluated on the same dataset as the one used in the scalability analysis (i.e., CHB-MIT for seizure detection, Clinical for seizure prediction, SleepEDFx for sleep stage classification, SEED for emotion recognition and Motor Imagery for motor imagery classification) described in Sec. 4.1. We follow the same pre-training configurations described in Sec.3.1. For the masking ratio, we experiment with settings of 20%, 40%, 60%, and 80%. Regarding the forecasting length, we try lengths of 1/16 L, 1/4 L, 1/2 L, and L, where L is the context length.

Results. Fig.7 illustrates the results of the hyperparameter analysis, in which the top five graphs represent the analysis for the masking ratio, and the bottom five graphs depict the analysis for the forecasting length. It can be observed that with a masking ratio of 40% and a forecasting length of 1/4 L, the models achieve the best overall performance among the five downstream tasks.

Since brain signals are natural signals with heavy redundancy, a missing patch can be recovered from neighboring patches with little high-level understanding of the semantic information. Thus, masking with a relative low ratio (20% or lower) or conducting very short-term forecasting (e.g. 1/16 L) may lead to ineffectiveness of high-level representation learning, which is crucial for time series classification[49]. However, as non-stationary time series, the values and associations between variables in brain signals significantly change over time. Therefore, a excessively high masking ratio (e.g., 80%) poses great challenges for the model to reconstruct the original signals. Similarly, when the forecasting length is equal to the context length, the prediction becomes extremely challenging, both of which may unstabilize the pre-training process and lead to a decline in performance on downstream tasks.

C DETAILS OF BASELINES

We extensively compare our model with 12 advanced methods which are divided into three categories, including 1) 3 methods aimed at time series universal modeling; 2) 3 methods based on self-supervised pre-training on brain signals and 3) 6 methods specifically designed for each downstream task. The detailed information of these methods are described below.

For the first category:

- TF-C [58]: A decomposable pre-training model for general time series modeling, where the self-supervised signal is provided by the distance between time and frequency components.
- SimMTM [14]: A pre-training framework on time series to recover masked time points by the weighted aggregation of multiple neighbors outside the manifold.

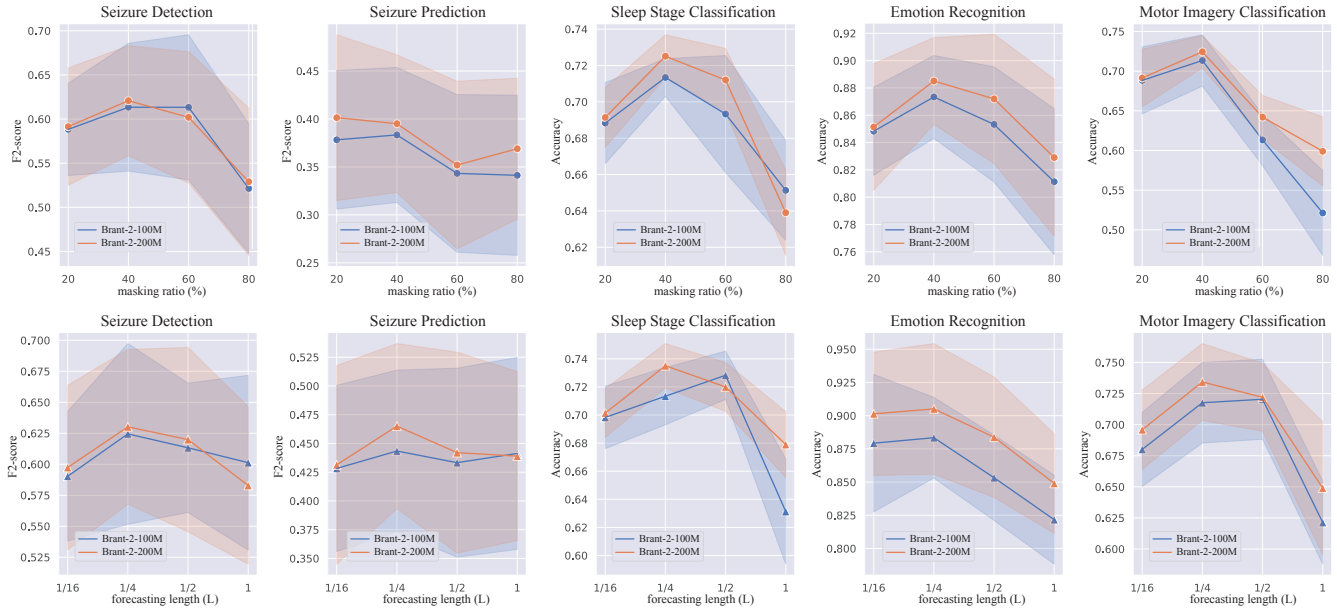

Figure 7: The hyperparameter analysis.

- One Fits All [61]: A unified model that leverages language models for time series analysis, leading to a comparable or SOTA performance in all main time series analysis tasks.

For the second category:

- BrainBERT[47]: A reusable transformer for intracranial field potential recordings enables classifying complex concepts and decoding neural data.
- Brant[56]: A foundation model for intracranial neural recordings which is a large-scale, off-the-shelf model for medicine.
- MBrain[7]: A multi-channel self-supervised learning framework which explicitly capture the spatial and temporal correlations of brain signals to learn a unique representation for each channel.

For the third category:

- PPi[54]: A pre-training-based model for patient-independent seizure detection on SEEG data, which contains two novel self-supervised tasks to extract rich information from abundant SEEG data and two techniques to tackle the domain shift problem.
- ScatterFormer[59]: An invariant scattering transform-based hierarchical Transformer that specifically pays attention to subtle features which is designed for patient-independent detection of epileptic based on visual spectral representation of continuous EEG.
- Lopes et al. [27]: A deep convolutional neural network-based EEG artefact removal model designed for seizure prediction using a deep convolutional neural network connected to a bidirectional long short-term memory layer (CNN-BiLSTM) using time series as input and a shallow artificial neural network trained using established handcrafted features.
- SleepHGNN[20]: A novel sleep heterogeneous graph neural network designed to capture interactivity and heterogeneity of physiological signals for accurate sleep stage classification.
- EEG Conformer[42]: A compact convolutional Transformer to encapsulate local and global features in a unified EEG classification framework for motor imagery and emotion recognition.
- TSFF-Net[29]: A novel network architecture designed for motor imagery classification that integrates time-space-frequency features, effectively compensating for the limitations of single-mode feature extraction networks based on time-series or time-frequency modalities.

For TF-C, SimMTM, BrainBERT, Brant and MBrain which need to be pre-trainined and applied on all the downstream tasks, we utilize the same pre-training corpus of *Brant-2* to pre-train these baselines for fair comparison. During fine-tuning, we fine-tune all the parameters of these baselines.

D DETAILS OF EXPERIMENTAL SETUP

D.1 Evaluation Metrics

In the seizure detection and prediction tasks, following the existing works[7, 9, 54], we adopt precision, recall, F1 score and F2 score as the evaluation metrics. For the sleep stage classification, following the existing works [34, 36, 43], we use accuracy, sensitivity, specificity, macro F1 score, and Cohen's kappa κ as evaluation metrics. For emotion recognition and motor imagery classification, we use accuracy and F1 score as evaluation metrics. Detailed information of these metrics are given as follows:

- Precision: Also known as positive predictive value (PPV), precision is the proportional accuracy of correctly identified positive outcomes out of all predicted positive outcomes. It's a crucial metric when the cost of a false positive is high. The higher the value, the more relevant the results returned by the model. A lower

value would mean that the model returns more false positives.

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

where TP is the number of true positives and FP is the number of false positives.

- Recall (Sensitivity): Also known as the true positive rate or sensitivity, recall measures the proportion of actual positive observations that are correctly identified as such. It helps us understand the predictive capacity of the model concerning the positive class. The higher the sensitivity, the fewer real positive cases the model will miss. A value of 1 means the model has perfect sensitivity and is not missing any real positives.

$$Sensitivity = Recall = \frac{TP}{TP + FN}, \tag{10}$$

where TP is the number of true positives and FN is the number of false negatives.

- Specificity: Also known as the true negative rate, specificity measures the proportion of actual negatives that are correctly identified. This provides insight into the predictive capacity of the model for the negative class. A higher specificity value means that the model is good at avoiding false positives, whilst a lower specificity indicates that the model often predicts a positive outcome when it's actually negative.

$$Specificity = \frac{TN}{TN + FP}, \tag{11}$$

where TN is the number of true negatives and FP is the number of false positives.

- F-measure: The F-measure is a metric defined as the weighted harmonic mean of precision and recall, with the following equation:

$$\mathrm{F}\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}. \tag{12}$$

In the scenario of epilepsy, F2 is more valued than F1, since ignoring any seizure is costly in diagnosis. While in other scenarios like sleep or emotion, F1 score is more valuable as a reference as it seeks a balance between precision and recall. For multi-class problems, macro Fβ score calculates Fβ for each class independently and then averages them. A higher macro Fβ score indicates that the classifier has both good precision and good recall.

- Cohen's Kappa: Kappa κ is a statistic that measures inter-rater agreement for qualitative items. It generally measures how well the model is performing over the random prediction. The value lies between -1 to 1. A high positive value (close to 1) signifies that the model's predictions align well with the actual results beyond what would be expected by chance, a value of 0 indicates alignment similar to random chance, and a negative value indicates agreement less than chance.

D.2 Fine-tuning Details

For most of the datasets (MAYO, FNUSA, CHB-MIT, Siena, Clinical, SEED and Motor Imagery), we find that only fine-tuning the last two layers of the temporal encoder and freezing the remaining parameters of the *Brant-2* encoder, we can achieve satisfactory results. Thus, for the evaluation on the above datasets, we only fine-tune the last two layers of the temporal encoder with a learning rate of 1.0×10^{-5} and the classification head with a learning rate of 1.0×10^{-3}. For datasets SleepEDFx and HMC, we fine-tune all the parameters of *Brant-2* with a learning rate of 1.0×10^{-6}.

D.3 Model Configurations

Table 6: Configurations of *Brant-2* with different sizes.

Model \ Config	Temporal/Spatial Encoder Layer	Model Dimension	Inner Dimension	Parameter Count
Brant-2-100M	10/2	768	2304	115M
Brant-2-200M	10/2	1024	3072	204M
Brant-2-460M	10/2	1536	4608	459M
Brant-2	8/2	2560	7680	1065M

The model configurations of *Brant-2* with different sizes are shown in Tab. 6.