SELF SUPERVISED LEARNING OF EEG (ELECTROENCEPHALOGRAM) RAW DATA TO LEARN THE HIDDEN PATTERNS OF HUMAN BRAIN ACTIVITIES.

Thambawita Maddumage Tharindu Akalanka Gunarathna
209327H

Master of Science in Computer Science Specialising in Data Science Engineering and Analytics

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

October 2022

SELF SUPERVISED LEARNING OF EEG (ELECTROENCEPHALOGRAM) RAW DATA TO LEARN THE HIDDEN PATTERNS OF HUMAN BRAIN ACTIVITIES.

Thambawita Maddumage Tharindu Akalanka Gunarathna
209327H

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science Specialising in Data Science Engineering and Analytics

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

October 2022

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Candidate Name: Gunarathna T. M. T. A.

……

Signature of Candidate Date:

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Supervisor Name: Dr. Thanuja D. Ambegoda

……

Signature of Supervisor Date:

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my supervisor, Dr. Thanuja D. Ambegoda for his invaluable advice, continuous support, and patience during my MSc academic research study. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I would also like to thank Dr. Sapumal Ahangama for introducing me to my supervisor. I would like to thank all the members in the Computer Science and Engineering Department. It is their kind help and support that have made my study and life in the University of Moratuwa a wonderful time. Finally, I would like to express my gratitude to my parents, my sisters and their families and my friends. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Abstract

EEG is a non-invasive neuroimaging modality that operates by measuring changes in electrical voltage on the scalp that are induced by cortical activity. In this research, we propose a method for self-supervised learning of EEG raw data to learn the hidden patterns of human brain activities. This work was performed through a pipeline consisting of five phases. Each of the phase's output will be the input for the next phase. Phase 1 is for pre-processing raw EEG sequences into EEG representations that catch the spacial and temporal properties in the original raw EEG sequences. We have followed a relatively less complex method to pre-process raw EEG sequences. In phase 2, pre-processed raw EEG sequences will be learnt by self-supervised representation learning. For that self-supervised vision transformers with DINO will be used. These vision transformers models are computationally more demanding and require more training data therefore more computational resources and training data will be needed. So that at the presence of more training data and computational processing power, self-supervised vision transformer architectures will be expected to produce the best results while outperforming supervised learning architectures. Then at the phase 3, sequences of prototypes for each raw EEG data sequence of the dataset will be generated. To evaluate the prototypes that are generated from raw EEG data, phase 4 and 5 have been used as the downstream task for the self-supervised learning task. For phase 4 and 5, we again used a transformer architecture, that is a BERT based model called RoBERTa to learn the synthetic language generated by phase 3 or to learn the context and the language of generated prototype sequences and by performing a multi class prototype sequence classification, prototype generation for each representation at specific time stamp of raw EEG data sequence can be evaluated. We believe that since the models are computationally demanding and require more training data, the latter explained pipeline of five phases should be improved with more training and performing hyperparameter tuning at a high computational resources and data rich environment.

Keywords: Electroencephalogram, Self-Supervised Learning, Vision Transformers, Natural Language Processing

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
EEG	Electroencephalogram
ECG	Electrocardiogram
EMG	Electromyogram
MRI	Magnetic Resonance Imaging
BCI	Brain Computer Interface
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
SSL	Self Supervised Learning
DINO	Self-distillation with no labels
BERT	Bidirectional Encoder Representations from Transformers

1. INTRODUCTION

1.1 Research background

EEG is a non-invasive neuroimaging modality that operates by measuring changes in electrical voltage on the scalp that are induced by cortical activity. It provides a wealth of physiological, psychological, and pathological information about the subject. There has been a large number of neurophysiological and psychological studies conducted that have discovered that the generation and activity of human emotions is highly correlated with the activity of the cerebral cortex. It has been shown that different cognitive and emotional activities in humans can result in different EEG signals [28]. The processing and recognition of EEG signals is a difficult task that requires a lot of effort. First and foremost, the EEG signal has a low signal-to-noise ratio and is susceptible to being interfered with by other noise sources. A sensitive EEG recording device, for example, is very susceptible to being influenced by the surrounding environment, and muscle activity, eye movement, and blinking can all contribute to unwanted noise. For the second time, people are frequently interested in EEG signals relating to specific brain activities that are always submerged in background noise, and it is difficult to distinguish these signals from the background noise in most cases.

The electroencephalogram consists of multiple time series corresponding to measurements taken at various spatial locations throughout the cerebral cortex. When it comes to audio signals, the frequency domain is where the most noticeable characteristics can be found [1]. In most existing works, EEG is either treated as a series of chain-like sequences[5,] with complex dependencies between adjacent signals[6], or it is represented as raw data[2]-[4], with important spatial information being lost. The electroencephalogram (EEG) is considered to be one of the most practical approaches to the development of brain-computer interface (BCI) systems as well as medical diagnostics. While it is possible to successfully identify the parts of the brain that are associated with cognitive events, doing so precisely remains a difficult task [6].

1.2 Research problem

The findings of this research would include a more precise identification of the brain regions that are activated by cognitive events. However, despite the extensive research on EEG that has been conducted in recent years, it is still difficult to interpret EEG signals effectively because of the large amount of noise present in EEG signals and the difficulty in capturing the subtle relationships between EEG signals and certain brain activities [5]. As a result, the findings of this research could be used to improve the existing EEG-based BCI in real-world applications by involving precisely identified areas of the brain and providing support for experiments involving neurodegenerative diseases such as Parkinson's disease.

By combining self-supervised representation learning with sequence classification to precisely identify the areas of the brain that are associated with cognitive events, we hope to uncover hidden patterns in EEG raw datasets while preserving the spatial information contained in EEG recordings.

1.3 Research objectives

1. Introduce a concept of brain language to generalize the handling of raw EEG data.
2. Support on applying EEG data acquired with different hardware (e.g. with different number of electrodes) by merging various EEG datasets together.
3. Identify precise areas of the brain where specific cognitive events are involved to address the uncertainties associated with EEG-based BCIs.
4. Support classification of unlabeled EEG raw data by self-supervised vision transformers.
5. Derive and compare the efficiency of different self-supervised learning methods proposed in the literature with respect to the proposed self-supervised vision transformers with DINO.

2. LITERATURE REVIEW

In this section, we will go over the relevant literature that has been cited so far in this paper. EEG representation, self-supervised learning with EEG, and self-supervised learning with imagery will all be covered in this section, which will be divided into three main subsections.

2.1 EEG representations

As deep learning and artificial intelligence technologies have advanced in recent years,emotion recognition has emerged as a popular research topic in the fields of human-computer interaction and affective computing. Emotion recognition is the process of determining the type of emotion that has been expressed by a given individual. The development of efficient and robust human emotion recognition algorithms will have a significant impact on wearable human-computer interaction in the near future, according to industry experts. It is possible to significantly improve the quality of the user experience by incorporating automatic emotion recognition into human-computer interaction applications. This can be accomplished by increasing the number of wearable devices that can detect emotional EEG activity and by realizing many different control functions that are based on emotion perception and regulation. A non-invasive brain imaging technique known as electroencephalography (EEG) measures the electrophysiological activities of the brain using scalp electrodes. EEG can reveal a great deal about a person's physiological, psychological, and pathological state. There has been a large number of neurophysiological and psychological studies conducted that have discovered that the generation and activity of human emotions are highly correlated with the activity of the cerebral cortex. It is possible for human beings to produce different EEG signals depending on their cognitive and emotional activities. The goal of effective brain-computer interface (BCI) control can be achieved with the help of effective feature extraction and classification techniques. Human facial expression, voice, electrocardiogram (ECG), electromyogram (EMG), magnetic resonance imaging (MRI), and other physiological signals are being replaced by EEG for the recognition of human emotion. This is due to the advantages of EEG, which include its strong objectivity, the fact that it is difficult to forge, the ease with which it can be acquired using a wearable EEG headset, and the ease with which it can be operated.

The processing and recognition of EEG signals is a difficult task that requires a lot of effort. First and foremost, the EEG signal has a low signal-to-noise ratio and is susceptible to interference from other sources of noise. For example, sensitive EEG recording equipment is very susceptible to being influenced by the surrounding environment, and muscle activity, eye movement, and blinking can all contribute to undesired noise in the recordings. For the second time, people are frequently interested in EEG signals relating to specific brain activities that are always submerged in background noise, and it is difficult to distinguish these signals from the background

noise in most cases. Although EEG signals recorded on the scalp have a low spatial resolution (in milliseconds), their high time resolution (in milliseconds) allows them to record changes in brain activity that are slow or rapid in nature. Consequently, both spatial and temporal EEG correlation must be taken into account in order to extract features associated with specific brain emotion dynamics.

It is proposed in the paper [8] that a new method of representing EEG data, which converts 1D chain-like EEG vector sequences into 2D mesh-like matrix sequences, be used for emotion recognition. The matrix structure simply maps the brain area distribution of 32 EEG electrodes' locations, which may be a more accurate representation of the spatial correlation of multiple electrodes that are physically adjacent to one another. In the following step, the sliding window is used to segment the 2D matrix sequences into segments that contain equal time points. Each segment is referred to as an EEG sample or epoch, and it incorporates all of the spatial and temporal information available. To predict the emotion category of each EEG sample, recurrent neural networks with cascaded and parallel convolution convolution convolution are proposed. CNN is used in these two hybrid networks to learn the high-level spatial correlation between physically adjacent EEG electrodes, and LSTM-based RNN is used to learn the subtle temporal dependency between time points in each network. On a large-scale DEAP dataset (32 subjects, 9,830,400 EEG recordings), we conducted extensive binary emotion classification experiments in valence and arousal to test the efficacy of our proposed methods and models. Using our spatial-temporal EEG representations, the experimental results demonstrate that the classification accuracies of both proposed hybrid networks achieve over 93 percent, outperforming the most recent baseline methods and other deep learning models, and yielding accuracy increases of 5.1 and 6.6 percent in valence and arousal, respectively, in a within-subject validation scenario.

An interface between the brain and the computer (brain-computer interface, BCI) converts neural activities into electrical signals, allowing researchers to investigate the relationship between brain activities and behavior. BCI systems are widely used in the field of rehabilitation medicine, where they are used to recognize the intentions of patients, among other things. Because of its low cost, ease of operation, and high time resolution, electroencephalography (EEG), the most common signal source for BCI systems, is widely used in intention recognition. EEG signals acquired by the EEG headset show different fluctuation patterns as subjects perform motor imagery or carry out specific actions, depending on the task at hand. When a subject's EEG signals are decoded, the BCI system kicks in. In recent years, researchers have concentrated on the recognition of intentions using electroencephalography (EEG). The company has created numerous BCI applications, including the control character system based on the P300, cursor control systems based on the mu rhythm and beta rhythm, and robot arm control, among other things.

The inherent limitations of EEG continue to pose numerous challenges to EEG intention decoding, despite the rapid development and satisfactory results obtained to

date. Random non-stationary behavior, a low signal-to-noise ratio, and high susceptibility to interference are all characteristics of the EEG signal. Interferences between the Electrooculogram (ECG), Electromyography (EMG), and Electrocardiogram (ECG) make it impossible to determine a relationship between the actions or emotions of the subject and the EEG signals, and vice versa. So feature extraction necessitates complex signal preprocessing, which varies according to the requirements of different research fields. Effective feature extraction is also critical for decoding motor intentions, and a variety of algorithms have been developed and tested in this area.

The paper [8] employs the novel Grad-CAM method for channel selection, which combines feature visualization technology with channel selection to achieve superior performance and channel selection. Although deep learning models lack transparency, their method not only explains the decision-making process of deep learning models but also explains the decision-making process of deep learning models. It also allows for the visualization of the channel selection process during the selection process. Their method, in order to maintain model performance, reduces the number of channels, resulting in a reduction in system resources. Furthermore, they demonstrate that their method achieves the best possible trade-off between model performance and the number of available channels. Also proposed is an EEG intention recognition network structure based on a recurrent-convolutional neural network structure that preserves and captures spatial information by converting the original one-dimensional (1D) EEG data vector into a two-dimensional (2D) EEG data matrix. The 2D EEG data matrix segments obtained by using the time sliding window contain both spatial and temporal information and thus are considered to be complete. In this paper, they propose a CNN-GRU structure for learning the decomposed space-time representation. When compared to the CNN-LSTM structure, their model performs significantly better in terms of complexity and training time. In cross-subject and multiclass scenarios, the researchers' method outperforms both the comparative state-of-the-art models and the baseline models, according to the results of their experiments.

Using the publicly available motor imagery EEG dataset EEGMMIDB, they carried out an experiment. The experimental results demonstrate that their method achieves an accuracy of 97.36 percent at the full channel, outperforming many state-of-the-art models as well as baseline models in the process. Their model has fewer parameters and requires less time to train, despite the fact that its decoding rate is the same as the best model tested. Following the channel selection, their model maintains the intention decoding performance of 92.31 percent while reducing the number of channels by nearly half and saving system resources, according to the authors. In the case of EEG intention decoding, their method achieves the best possible trade-off between performance and the number of electrode channels.

With the help of a brain-computer interface (BCI), users can directly communicate with the outside world or control instruments solely through their thoughts, providing

an alternative and practical way to assist people who are suffering from severe motor disabilities, such as stroke or paralysis. Recent research has also discovered applications for healthy users, such as brain-computer interface (BCI) games in the entertainment industry. It is believed that scalp-recording electroencephalography (EEG) is the most practical method of realizing brain-computer interface (BCI) systems because of its portability and ease of implementation of the acquisition system. It has been observed that when someone imagines moving different parts of his body or giving different controlling commands to an instrument, the EEG signals from his scalp fluctuate in various modes. The EEG signals can be analyzed in this manner to determine the intentions of the subject. It has been gaining popularity, and various studies have attempted to incorporate EEG-based BCI into real-world applications such as mind-controlled wheelchairs, prosthetics, and exoskeletons, among other things.

In real-world situations, on the other hand, due to a variety of open challenges, EEG-based BCI systems are still in their infancy. EEG signals are typically characterized by a significant amount of background noise. The EEG signals, in addition to the common noises that affect sensory systems, such as power line interference and improper electrode connections, also contain some unique and inevitably noisy components. EEG signals with high signal-to-noise ratios are adversely affected by physiological activities such as eye blinks, muscle activity, and heartbeat during the recording process. Ensuring that participants remain focused on the tasks at hand throughout the duration of an experiment is difficult. Furthermore, a typical EEG-based BCI system typically has 8 to 128 signal channels, resulting in limited signal resolution when compared to tasks requiring image or video processing and recognition capabilities. For the second time, there is some uncertainty about the correlations between EEG signals and the brain intentions that correspond to them in deeper structures. Unlike body movements, which can be easily explained by monitoring accelerometers or gyroscopes, inferring the intentions of the brain by directly observing EEG signals is not straightforward. In addition, widely used brain intention recognition methods heavily rely on handcrafted features, requiring extensive preprocessing prior to making a prediction in order to be accurate. Signal de-noising and feature selection steps are performed in some methods before a final recognition model is developed. It is inconvenient to train and implement a two-stage model in this manner, and the entire process is time-consuming and highly dependent on professional knowledge in this area. In conclusion, current research is focused primarily on either intra-subject (where test data and train data are collected from the same subject), or binary EEG signal classification scenarios. It is important to note that very little research has been done on scenarios involving more than one subject or class. But for the purpose of implementing real-world applications, cross-subject and multi-class scenarios are highly desired. Aside from that, many existing works exhibit poor performance even in intra-subject or binary classification scenarios, with accuracy levels hovering around 80%.

In order to overcome the aforementioned challenges in the development of EEG-based BCIs, the paper [5] proposes two types of convolutional recurrent neural networks, which they refer to as cascade and parallel models, for detecting human intentions by learning the effective compositional Spatio-temporal dynamics from raw EEG streaming signals without preprocessing, respectively. To be more specific, they create a mesh-like raw EEG signal hierarchy from 1D chain-like EEG vectors by mapping the EEG recordings with the spatial information of the EEG acquisition electrodes, which allows them to align the correlations between neighboring EEG signals and the corresponding brain regions. Afterward, both cascade and parallel convolutional recurrent network models are developed to decode robust EEG representations from both space and time dimensions in either a sequential or a parallel manner, as appropriate. In this paper, they propose models that are unified end-to-end trainable models that learn robust feature representations while also classifying EEG raw signals to detect movement or instruction intentions at the same time. It is possible to generalize the proposed models to more complex and practical scenarios (both cross-subject and multi-class). When it comes to movement intention recognition, both the cascade and parallel models achieve high accuracy of close to 98.3 percent, outperforming the current state-of-the-art methods by approximately 18 percent. Also included is an evaluation of our models on a real-world BCI system, with results showing that our models achieve an acceptable accuracy of 93 percent when recognizing five instruction intentions with limited EEG channels. Thus, their proposed models demonstrate robust abilities to recognize a wide range of human intentions when used in conjunction with various BCI systems.

When it comes to human daily life, emotion is extremely important, as it reflects the emotions that people have toward various things. People's interpersonal interactions and decision-making are influenced by their mental health status. It has been suggested that the emotional states detected in patients can be used as an indicator for certain functional emotional disorders, such as posttraumatic stress disorder and major depression, in the medical fields of psychiatry and neurology. In a recent study, EEG signals were used to compare the emotional characteristics of a group of people who overused their smartphones to a healthy group.

Facial expressions, speech, eye blinking, and physiological signals can all be used to determine how someone is feeling. The first three approaches, on the other hand, are susceptible to subjective influences from the participants, which means that participants can purposefully conceal their emotions in these approaches. In contrast, physiological signals such as electroencephalograms (EEGs), electrooculography (EOGs), and blood volume pressure (BVPs) are produced by the human body on its own initiative and without the assistance of a machine. As a result, physiological signals are more objective and reliable when it comes to capturing the true emotional states of humans. Of all of these physiological signals, the EEG signal is the only one that originates in the human brain, which means that changes in EEG signals can be

used to directly correlate changes in human emotional states. As a result, researchers intend to use EEG signals to investigate human emotion.

Automatic real-time emotion recognition based on multi-channel EEG signals is becoming an increasingly important computer-aided method for diagnosing emotional disorders in neurology and psychiatry, despite the fact that it is a difficult pattern recognition task. Machine learning approaches that are based on comprehensive domain knowledge are required to design and extract various features from single or multiple channels in order to be effective. As a result, these approaches may present a challenge for those who are not domain experts. Rather, deep learning approaches have been successfully applied in many recent publications to learn features and categorize different types of data, with varying degrees of success. Specifically, in the paper [9], baseline signals are taken into consideration, and a simple but effective pre-processing method is proposed in order to improve the recognition accuracy of the network. The use of a hybrid neural network, which incorporates elements of both "Convolutional Neural Network (CNN)" and "Recurrent Neural Network (RNN), has been demonstrated to effectively classify human emotion states by effectively learning compositional spatial-temporal representations of raw EEG streams. It is necessary to convert the chain-like EEG sequence into a 2D-like frame sequence before using the CNN module to mine the inter-channel correlation among physically adjacent EEG signals in order to perform this task. The LSTM module is used to extract contextual information from data. A segment-level emotion identification task is used in the experiments, which are carried out on the DEAP benchmarking dataset. Their experimental results indicate that the proposed pre-processing method can increase emotion recognition accuracy by approximately 32 percent and that the model achieves high performance, with a mean accuracy of 90.80 percent on valence classification tasks and 91.03 percent on arousal classification tasks, respectively, on valence and arousal classification tasks.

Another challenge in analyzing and modeling cognitive events from electroencephalogram (EEG) data is identifying representations that are invariant to individual differences as well as to the inherent noise associated with EEG data collection. According to the paper [1], the authors propose a novel approach for learning such representations from multichannel EEG time series and demonstrate its benefits in the context of mental load classification tasks. For starters, they convert EEG activities into a sequence of topology-preserving multi-spectral images, as opposed to standard EEG analysis techniques, which ignore spatial information in favor of spectral information. They then train a deep recurrent convolutional network, which is inspired by current video classification techniques, in order to learn robust representations from the sequence of images they have collected. When using the proposed approach, the spatial, spectral, and temporal structure of EEG is preserved, which allows for the identification of features that are less sensitive to variations and distortions within each of the three dimensions. The results of an empirical evaluation of the cognitive load classification task revealed significant improvements in

classification accuracy when compared to current state-of-the-art approaches in this field.

Neurodegenerative diseases such as Parkinson's disease are associated with REM Behavior Disorder (RBD), which is a serious risk factor (PD). Deep learning methods for idiopathic rapid eye movement behavior disorder (RBD) prognosis classification from electroencephalography are described in the paper [2] (EEG). They make use of a few minutes of resting-state EEG data collected from patients with idiopathic RBD (121) and healthy controls (120). (HC, 91). A subset of RBD patients eventually developed either Parkinsonism (19) or Dementia with Lewy bodies (DLB) over the course of their follow-up (mean of 4 2 years), while the remainder remained idiopathic RBD. They first describe a deep convolutional neural network (DCNN) trained with stacked multi-channel spectrograms, treating the data in the same way that deep classifiers have proven highly successful in audio and image problems, where deep classifiers have exploited compositional and translationally invariant features in the data to great effect. The performance of a small DCNN network can typically achieve 80 percent classification accuracy by utilizing a multi-layer architecture that combines filtering and pooling.Using this approach, The researchers found that using a single channel, they were able to achieve an area under the curve (AUC) of 87 percent in the HC vs PD outcome problem. The trained classifier can also be used to generate synthetic spectrograms in order to investigate which aspects of the spectrogram are relevant to classification, highlighting the presence of theta bursts and a decrease in power in the alpha band in the future Parkinson's disease or dementia with Lewy bodies (DLB) patients. To provide a point of comparison, the researchers investigate a deep recurrent neural network that employs either stacked long short term memory network (LSTM) cells or gated-recurrent unit (GRU) cells, with results that are similar. They come to the conclusion that, despite the limitations of this first study's scope, deep classifiers may be an important technology for analyzing EEG dynamics from small datasets and identifying new biomarkers.

2.2 Self-supervised learning with EEG

Many applications, both inside and outside of the clinical domain, have been made possible by electroencephalography (EEG) and other biosignal modalities, such as the study of sleep patterns and their disruption, the monitoring of seizures, and the interfacing of the brain with computers. These devices' availability and portability have increased dramatically in recent years, effectively democratizing their use and unlocking the potential for them to have a positive impact on people's daily lives. Application areas such as at-home sleep staging and apnea detection, pathological EEG detection, mental workload monitoring, and so on are now entirely feasible. Examples include

All of these scenarios involve the use of monitoring modalities, which generate an ever-increasing amount of data that must be interpreted. It is necessary, therefore, to develop predictive models that can classify, detect, and ultimately "understand"

physiological data in real-time. Traditionally, this type of modeling has relied heavily on supervised approaches, which necessitate the use of large datasets of annotated examples in order to train models that are capable of high performance.

Accurate annotations of physiological data, on the other hand, can be difficult, expensive, and time-consuming to obtain. Adding annotations to sleep recordings, for example, requires trained technicians to visually scan hours of data and label 30-second windows one at a time, which can take several hours. Neurologists must review clinical recordings, such as those used to diagnose epilepsy or brain lesions, and they may not be available at all times. The complexity of brain processes of interest, combined with noise in the data, makes it difficult to interpret and annotate EEG signals. This can result in high inter-rater variability, also known as label noise. Furthermore, it can be difficult to determine exactly what participants were thinking or doing in cognitive neuroscience experiments in some cases, making it difficult to obtain accurate labels for the participants. If the subjects are performing imagery tasks, for example, it is possible that they are not following instructions or that the process under investigation is difficult to objectively measure (e.g., meditation, emotions). As a result, in order to make use of large unlabeled sets of recordings such as those generated in the scenarios described above, a new paradigm that does not rely primarily on supervised learning is needed. Traditional unsupervised learning approaches, such as clustering and latent factor models, do not provide completely satisfactory answers because their performance is more difficult to quantify and interpret than that of supervised learning approaches.

As stated in this paper [10], supervised learning paradigms are frequently constrained by the amount of labeled data that is available to the researchers. This phenomenon is particularly troublesome in clinically relevant data, such as electroencephalography (EEG), where labeling can be time-consuming and expensive in terms of specialized expertise and human processing time, as shown in the figure below. As a result, deep learning architectures designed to learn from EEG data have produced models that are relatively shallow and have performance that is at best comparable to that of traditional feature-based approaches at best. However, in the majority of cases, unlabeled data is readily available in large quantities. Using deep neural networks to extract information from this unlabeled data, it may be possible to achieve competitive performance despite having limited access to labeled data in the future. When it came to learning representations of EEG signals, the researchers looked into self-supervised learning (SSL), a promising technique for discovering structure in unlabeled data. Specific tasks based on temporal context prediction and contrastive predictive coding were investigated on two clinically-relevant problems: EEG-based sleep staging and pathology detection, respectively. They carried out experiments on two large public datasets containing thousands of recordings and made baseline comparisons between approaches that were purely supervised and those that were hand-engineered. Linear classifiers trained on SSL-learned features outperformed purely supervised deep neural networks on a consistent basis in low-labeled data regimes, while also

achieving competitive performance in high-labeled data regimes. The embeddings learned with each method also revealed clear latent structures associated with physiological and clinical phenomena, such as age effects. As a result, they demonstrate the effectiveness of self-supervised learning approaches on EEG data. Their findings suggest that SSL may pave the way for a more widespread application of deep learning models on EEG data in the future.

In every action of our daily lives, our emotions are manifested in a variety of ways. It is one of the most important aspects of human development and growth to be able to recognize and understand one's own emotions, and it plays an important role in the emulation of human intelligence. As a result, effective computing and automatic emotion recognition are critical for the advancement of artificial intelligence and all of the research fields that stem from it. Electroencephalography (EEG) is a technique that measures oscillations in the brain that are caused by the synchronized activity of neuronal networks. It is hypothesized that changes in these oscillations are correlated with the cognitive process and that they can be used to reveal important information about human emotional states in certain situations. When considered as a type of physiological signal, EEG has the advantage of being difficult to conceal or concealment. The time resolution of this signal is excellent when compared to other physiological signals, and it is similar to the nuanced changes in emotional states on a temporal scale. Increasing attention has been drawn to EEG-based emotion recognition as a result of the rapid development of non-invasive, simple to use, and inexpensive recording devices. This has resulted in an increase in both research and application interest in this area.

Although EEG has several advantages, it also has some disadvantages. First and foremost, because it is an aggregate signal derived from the activity of millions of neurons, the EEG has a low signal-to-noise ratio (SNR). Second, EEG is generally recorded using tens to hundreds of electrodes at the same time, and the sampling time in each trial is typically greater than a few seconds per electrode. As a result, the original feature dimension of an EEG sample is not very small. On the other hand, a typical dataset for cognitive neuroscience tasks, usually only contains a few hundred to a few thousand samples, which is not very many (i.e., experimental trials). Because of this, the initial sample to feature ratio is extremely low. Third, the EEG is a non-stationary signal, which means that its statistics change over the course of time. EEG analyses cannot be generalized across subjects because of the inherent variability in brain anatomy, head size, and dynamics across trials/subjects. This is especially true when comparing EEG analyses across subjects performing a single task. The second limitation significantly complicates the application of machine learning models, and the other two limitations only serve to exacerbate this difficulty.

Using machine learning algorithms, particularly deep learning models, to solve the data scarcity problem in Electroencephalography (EEG)-based affective computing results in difficulty in developing an effective model with high accuracy and stability. It has recently been demonstrated that data augmentation can significantly improve

deep learning model performance by increasing accuracy and stability while also decreasing over-fitting (see Figure 1). Researchers have proposed a novel data augmentation framework, referred to as GANSER (Generative Adversarial Network-based Self-supervised Data Augmentation Framework for EEG-based Emotion Recognition), in their paper [11] (GANSER). The proposed framework, which is the first to combine adversarial training with self-supervised learning for EEG-based emotion recognition, can generate high-quality and high-diversity simulated EEG samples, making it a world first. The researchers in this study, in particular, used adversarial training to learn an EEG generator and forced the generated EEG signals to approximate the distribution of real samples in order to ensure the quality of augmented samples.For a wide variety of samples, a transformation function is used to mask portions of EEG signals and force the generator to synthesize potential EEG signals based on the portions that remain visible. It is proposed to use the masking possibility during transformation as prior knowledge to guide the extraction of distinguishable features from simulated EEG signals and the generalization of the classifier to the augmented sample space, respectively. Finally, extensive experiments demonstrate that their proposed method can aid in the recognition of emotions for the purpose of performance improvement while also achieving state-of-the-art results.

It is usually straightforward to obtain EEG signals, but it is more difficult to label them accurately. In the field of EEG signal analysis, supervised learning has been widely employed; however, the generalization performance of this technique is limited by the amount of annotated data available. SSL, a popular learning paradigm in computer vision (CV) and natural language processing (NLP), can make use of unlabeled data to compensate for the lack of labeled data in supervised learning, according to the researchers. Using EEG signals for sleep stage classification, the researchers propose a self-supervised contrastive learning method for EEG signals in their paper [12]. During the training process, they set up a pretext task for the network to complete in order to match the appropriate transformation pairs generated from EEG signals to the correct transformation pairs. This way, the network learns the general characteristics of EEG signals and improves its ability to represent them. The network's robustness is also enhanced when dealing with a variety of data types, which is to say when extracting constant features from constantly changing data sources. Detailing the network's performance, the amount of unlabeled data that is used in the training process of self-supervised learning, as well as the transformations used, can be found in the paper. Experiments with the Sleep-edf dataset demonstrate that their method has competitive performance on sleep staging (88.16 percent accuracy and 81.96 percent F1 score) and that the SSL strategy is effective for EEG signal analysis in limited labeled data regimes.

In many cases, the cost - and sometimes the impracticality - of data collection and labeling in multiple domains limits the application of the supervised learning paradigm. Self-supervised learning, a paradigm that takes advantage of the structure of unlabeled data to generate learning problems that can be solved with standard

supervised approaches, has shown great promise in fields such as computer vision and time series processing as a pre-training or feature learning approach. A self-supervision strategy for learning informative representations from multivariate time series is presented in this paper [13], and it can be used to learn informative representations from electroencephalography signals. One approach that has proven successful is based on predicting whether time windows are sampled from the same temporal context or whether they are not. The researchers demonstrated that their approach outperforms a purely supervised approach on a clinically relevant task (sleep scoring) and with two electroencephalography datasets in low data regimes, while also capturing important physiological information without access to labels.

Generally speaking, deep neural networks (DNNs) used for brain-computer interface (BCI) classification are expected to learn general features when trained across a variety of contexts, with the expectation that these features can then be fine-tuned to specific contexts. While some success has been achieved using this approach, the researcher[14] believes that it is limited and that an alternative approach would better utilize the newly (publicly) available massive electroencephalography (EEG) datasets. They consider how to adapt techniques and architectures used for language modeling (LM) that appear capable of ingesting enormous amounts of data toward the development of encephalography modeling with deep neural networks (DNNs) in the same vein as language modeling.

For this purpose, they specifically adopt an approach that has proven successful for automatic speech recognition and that, like LMs, employs a self-supervised training objective in order to learn compressed representations of raw data signals. With EEG adaptation, they discover that a single pre-trained model is capable of modeling completely novel raw EEG sequences recorded with different hardware and different subjects completing a variety of tasks with the help of different training methods. The internal representations of this model as well as its overall architecture can be tailored to a variety of downstream BCI and EEG classification tasks, outperforming previous work in more task-specific classification tasks (sleep stage classification) self-supervision

The amount of labeled data that can be used in supervised learning paradigms is frequently limited. As a result, this phenomenon is particularly problematic when dealing with clinically relevant data, such as electroencephalography (EEG), where labeling can be time-consuming and expensive in terms of specialized expertise and human processing time. As a result, deep learning architectures designed to learn from EEG data have produced models that are relatively shallow and have performance that is at best comparable to that of traditional feature-based approaches, at worst they are worse. Unlabeled data, on the other hand, is readily available in the majority of situations. It may be possible to achieve competitive performance with deep neural networks despite having limited access to labeled data if information can be extracted from this unlabeled dataset. To learn EEG signal representations, the researchers used self-supervised learning (SSL), a promising technique for discovering structure in

unlabeled data. The findings were published in the paper [15]. To be more specific, they take a contrastive approach and present results on two clinically relevant problems: EEG-based sleep staging and the detection of pathology in the brain. Overall, the results show that linear classifiers trained on SSL-learned features consistently outperform pure-supervised deep neural networks in low-labeled data regimes and can achieve competitive performance when all labels are available. Their findings suggest that self-supervision may pave the way for a more widespread application of deep learning models on EEG data in the future, according to the researchers.

It is possible to significantly improve seizure diagnosis and treatment by using automated seizure detection and classification from electroencephalography (EEG). In previous automated seizure detection and classification studies, several modeling challenges went unaddressed, including (1) representing non-Euclidean data structure in EEGs, (2) accurately classifying rare seizure types, and (3) the lack of a quantitative interpretability approach to measure the model ability to localize seizures. Using graph neural networks (GNNs) to represent spatiotemporal dependencies in EEGs and proposing two EEG graph structures that capture electrode geometry or dynamic brain connectivity, the researchers attempt to address these challenges by (1) developing a self-supervised pre-training method that predicts preprocessed signals for the next time period to further improve the accuracy of EEG analysis, and (2) proposing a self-supervised pre-training method that predicts preprocessed signals for the next time period to further improve the accuracy of Using a large public dataset (5,499 EEGs), they discover that their GNN with self-supervised pre-training achieves a 0.875 Area Under the Receiver Operating Characteristic Curve on seizure detection and a 0.749 weighted F1-score on seizure classification, outperforming previous methods for both detection and classification of seizure activity. Furthermore, their [16] self-supervised pre-training strategy significantly improves the classification of rare seizure types, which is of particular importance (e.g. 47 points increase in combined tonic seizure accuracy over baselines). The researchers also discovered that their GNN with self-supervised pretraining accurately localizes 25.4 percent of focal seizures, which represents a 21.9 point improvement over existing CNNs. After all, is said and done, their approach, which involves superimposing the identified seizure locations on both raw EEG signals and EEG graphs, has the potential to provide clinicians with an intuitive visualization of localized seizure regions.

Using self-supervised learning, the authors of this paper [17] attempt to learn efficient representations from raw electroencephalogram (EEG) signals for sleep stage classification using predictive and discriminatory contrastive coding (SSL). Despite the fact that supervised methods have demonstrated superior performance, they are heavily reliant on manually labeled datasets. SSL has recently demonstrated that it can achieve comparable performance to fully supervised methods despite having only a small amount of labeled data by extracting high-level semantic representations. They propose SleepDPC, a novel sleep stage classification algorithm based on SSL, in order

to alleviate the over-reliance on labels that have been observed. SleepDPC improves on the efficiency with which it discovers underlying semantics from raw EEG signals by incorporating two dedicated predictive and discriminative learning principles into its algorithm. They thoroughly evaluate the performance of their proposed method on two publicly available datasets, which they developed themselves. The experimental results demonstrate that their method not only learns meaningful representations, but it also produces superior performance when compared to various competing methods, despite the fact that they only have access to labeled data.

Among the most common neurological diseases in humans, epilepsy is one of the most prevalent, and electroencephalography (EEG) is the most widely used method for clinicians to detect epileptic seizures. On the contrary, manually observing EEG data is error-prone, and labeling epilepsy data is an expensive and time-consuming process that requires a large number of resources. For the purpose of anomaly detection on electroencephalography signals, a new self-supervised learning method is proposed in this study [18]. This method does not require any epileptic EEG data and only uses normal EEGs to detect anomalies in electroencephalography signals. The original EEG data is subjected to a series of scaling transformations in order to produce self-labeled scaled EEG data, where different labels correspond to different scaling transformations. In particular, Then, using the self-labeled normal EEG dataset, a multi-class classifier can be trained to accurately predict the scaling transformations on new normal EEG data, but not accurately on abnormal (epileptic) EEG data. Following that, the degree of inconsistency between predicted scaling transformations and ground truth scaling transformations can then be used to determine the degree of abnormality in a newly acquired EEG dataset.In-depth experimental evaluations demonstrate that the proposed self-supervised method outperforms traditional anomaly detection methods such as the one-class support vector machine (SVM) and autoencoders. Experiments with different classifier structures and variations in relevant hyper-parameters have also demonstrated the robustness of the proposed method.

2.3 Self supervised learning with imagery

Convolutional neural networks are being used to pre-train general-purpose visual features that do not require annotations, which is a difficult and important task. Unsupervised feature learning has recently been focused on either small or highly curated datasets like ImageNet, with non-curated raw datasets being found to reduce feature quality when evaluated on a transfer task. It is the authors' goal in this paper [19] to bridge the performance gap between unsupervised methods trained on curated data, which is difficult to obtain, and massive raw datasets, which are readily available. This is accomplished through the use of a novel unsupervised approach that takes advantage of self-supervision and clustering to extract complementary statistics from large-scale data. Using 96 million images from the YFCC100M dataset, they demonstrate the effectiveness of their approach by achieving state-of-the-art results

among unsupervised methods on standard benchmarks. This demonstrates the potential of unsupervised learning even when only non-curated raw data is available. It is also demonstrated that pre-training with their method results in 74.9 percent top-1 classification accuracy on the ImageNet validation set, which is an improvement of +0.8 percent over the same network trained from scratch, as demonstrated in their paper.

The paper [20] introduces SimCLR: a simple framework for contrastive learning of visual representations. SimCLR is a simple framework for contrastive learning of visual representations. They make it easier to use recently proposed contrastive self-supervised learning algorithms because they do not require specialized architectures or a memory bank, as was previously required. They conduct a systematic study of the major components of their framework in order to gain a better understanding of what allows the contrastive prediction tasks to learn useful representations. They demonstrate that (1) the composition of data augmentations is critical in the definition of effective predictive tasks, (2) the introduction of a learnable nonlinear transformation between the representation and the contrastive loss significantly improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps when compared to supervised learning, respectively. The researchers have discovered a way to combine their findings in order to outperform previous methods for self-supervised and semi-supervised learning on ImageNet by a significant margin. A linear classifier trained on self-supervised representations learned by SimCLR achieves top-1 accuracy of 76.5 percent, which represents a 7% improvement over the previous state-of-the-art and is comparable to the performance of a supervised ResNet-50 in terms of accuracy and precision. When they are fine-tuned on only 1% of the labels, they achieve top-5 accuracy of 85.8%, outperforming AlexNet, despite having 100 times fewer labels to work with.

Recently developed contrastive learning methods, in particular, have made significant progress in closing the gap between unsupervised and supervised image representations. It is computationally challenging to perform a large number of explicit pairwise feature comparisons in real-time with these contrastive methods, which are typically used online. In this paper [21], the authors propose an online algorithm, SwAV, that takes advantage of contrastive methods without the need to compute pairwise comparisons between the input and output images. As an example, instead of directly comparing features as in contrastive learning, their method simultaneously clusters the data while enforcing consistency between cluster assignments generated for different augmentations (or "views") of the same image. Put another way, they employ a "swapped" prediction mechanism, in which they predict the code of a view based on its representation in another view. A large and small batch of data can be used to train their method, and it can handle an unlimited amount of data. Because it does not necessitate the use of a large memory bank or a special momentum network, their method is more memory efficient than previously used contrastive methods. In

addition, they propose a new data augmentation strategy called multi-crop, which employs a mix of views with different resolutions in place of two full-resolution views, without increasing the memory or compute requirements of the system in question. It is demonstrated that they achieved 75.3 percent top-1 accuracy on ImageNet with ResNet-50, as well as surpassing supervised pretraining on all of the transfer tasks considered in the study.

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning is a new approach to self-supervised image representation learning introduced in the paper [22]. Essentially, BYOL is comprised of two neural networks, referred to as the online and target networks, that communicate with and learn from one another. In order to predict the target network representation of an image under a different augmented view, they train the online network to predict the target network representation of the same image under the same augmented view. Meanwhile, they update the target network with a slow-moving average of the online network, which is updated at the same time. BYOL achieves a new state of the art without using negative pairs, in contrast to current state-of-the-art methods that do so. The BYOL system achieves 74.3 percent top-1 classification accuracy on ImageNet when evaluated using a linear evaluation and a ResNet-50 architecture, and 79.6 percent when evaluated using a larger ResNet. They demonstrate that BYOL is on par with or better than the current state of the art on both transfer and semi-supervised benchmarks, according to the researchers.

Momentum Contrast (MoCo) for unsupervised visual representation learning is presented in the paper [23] (in English only). They develop a dynamic dictionary with a queue and a moving-averaged encoder from the perspective of contrastive learning as a dictionary lookup. This allows for the construction of a large and consistent dictionary on the fly, which aids in the facilitation of contrastive unsupervised learning. On ImageNet classification, MoCo produces results that are competitive with those obtained using the common linear protocol. More importantly, the representations learned by MoCo are easily transferable to other tasks in the future. Using PASCAL VOC, COCO, and other datasets, MoCo can outperform its supervised pre-training counterpart in 7 detection/segmentation tasks, with some results exceeding the supervised counterpart by significant margins. This suggests that the gap between unsupervised and supervised representation learning has been substantially narrowed in many vision tasks over the past few years.

Contrastive unsupervised learning has made significant strides in recent years, as demonstrated by the Momentum Contrast (MoCo) and SimCLR experiments. In this paper [24], the authors test the efficacy of two SimCLR design improvements by incorporating them into the MoCo framework and verifying their effectiveness. They establish stronger baselines that outperform SimCLR with simple modifications to MoCo, such as the use of an MLP projection head and more data augmentation, and they do so without the need for large training batches of data. They hope that by doing so, they will be able to make cutting-edge unsupervised learning research more

accessible.

In the paper [25], the researchers investigate whether self-supervised learning can impart new properties to Vision Transformers (ViTs) that are distinguishable from those found in convolutional neural networks (CNNs) (convnets). Aside from the fact that adapting self-supervised methods to this architecture works exceptionally well, they make the following observations about the architecture: The self-supervised ViT features, for starters, contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs or convnets, and second, self-supervised ViT features are more efficient than supervised ViT features. First and foremost, these characteristics make for excellent k-NN classifiers, with top-1 performance on ImageNet reaching 78.3 percent with a low ViT. Also highlighted in their research is the importance of using a momentum encoder, multi-crop training, and the use of small patches when working with virtual insects (VITs). DINO is a simple self-supervised method developed by the researchers, which they interpret as a form of self-distillation that does not require any labels. Using ViT-Base, they demonstrate the synergy between DINO and ViTs by achieving an 80.1 percent top-1 on ImageNet in linear evaluation.

3. METHODOLOGY

According to the previous literature about self-supervised learning of EEG, they have been mainly focused on EEG as a multi-channel EEG time-series that ignores the spatial features on them. In addition, those approaches require comprehensive domain knowledge and that might be an obstacle for non-domain experts in the field. To mitigate those challenges, in this research, we propose an approach that considers the spatial features of EEG raw data by applying a data representation, which transforms 1D chain-like EEG vector sequences into 2D mesh-like sequences and applies state-of-the-art approaches like Self Supervised Learning with Vision Transforms (ViT) for unsupervised representation learning on preprocessed 2D meshes and BERT (Bidirectional Encoder Representations from Transformers) based models for evaluation of unsupervised representations by prototype sequence classification.

In the following subsections, the procedure which is going to be followed will be discussed and will be divided into five main phases, phase 1 - preprocessing of EEG raw data, phase 2 - unsupervised representation learning by self-supervised learning of 2D meshes, phase 3 - transforming raw EEG sequences into prototype sequences, phase 4 - training a masked language model and finally phase 5 - pretraining a prototype sequence classifier as a downstream task.

3.1 Spatio-temporal preserving representations for raw EEG data

To conduct our research, we make use of the publicly available motor imagery dataset EEGMMIDB [28], which comes from PhysioNet. An international standard 10–10 system was used to acquire the data, with a sampling frequency of 160 Hz, on BCI2000 equipment, which had a total of 64 electrode channels and utilized an international standard 10–10 system to do so. Eyes closed, imagining opening/closing the left fist, imagining opening/closing the right fist, imagining opening and closing both feet are among the five brain activities included in the dataset, which was collected from 109 subjects. Each subject participated in 14 experiments, with the baseline starting with the eyes closed and the rest of the imagery being repeated three times, for a total of more than 1500 recordings on each subject's behalf. Because the data of subject 89 is completely different from the data of the other subjects, only one motor intention was performed in the experiment, and as a result, the data of subject 89 were excluded from consideration in this investigation. Therefore, a dataset consisting of 108 participants is used in this paper.

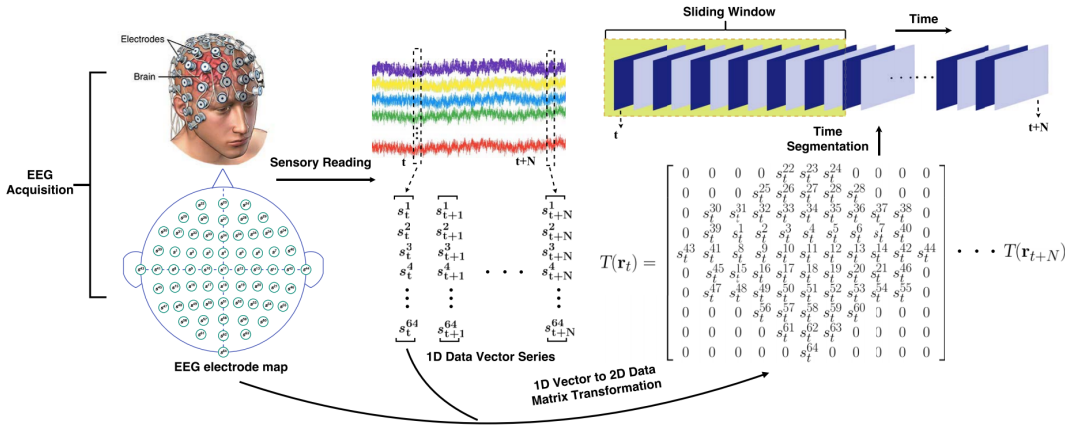

Figure 1. Acquisition and preprocessing of electroencephalogram (EEG) data In order to record EEG signals as time-series data vectors, a BCI headset with multiple electrodes is first used to capture EEG signals. In accordance with the electrode map of the BCI headset, these data vectors are then converted to 2D data meshes for further processing. Sliding window techniques are used to segment the converted 2D meshes into clips at the end of the process.

The overall EEG data acquisition and preprocessing flowchart of our proposed method are shown in Figure 1. The EEG-based BCI system uses a wearable headset with multiple electrodes to capture the EEG signals. When a subject imagines performing a certain instruction, the electrodes of the headset acquire the fluctuations of the voltages from the scalp. The EEG electrode map in Figure 1 depicts the electrodes placement of an example BCI headset. The electrode map varies from different BCI systems according to the different number of recording channels. The sensory readings from the EEG acquisition system represent time-series data at the acquiring frequency. Typically, the raw data from the EEG signal acquisition system at time index t is a one-dimensional (1D) data vector $r_t = [s^1{}_t, s^2{}_t, s^i{}_t \dots, s^n{}_t]^T$, where $s^i{}_t$ is the reading data of the ith electrode channel at timestamp t. The acquisition system totally contains n channels. For the observation period [t, t + N], there are (N + 1) 1D data vectors, each of which contains n elements corresponding to n electrodes of the acquisition headset.

From the EEG electrode map, it is observed that each electrode is physically neighboring multiple electrodes which measure the EEG signals in a certain area of the brain, while the elements of the chain-like 1D EEG data vectors are restricted to two neighbors. Furthermore, different brain regions correspond to different brain activities. From this conceptualization, we convert the 1D EEG data vectors to 2D EEG data meshes according to the spatial information of the electrode distribution of the acquisition system. The transformation function of the 1D data vector r_t at timestamp t for its corresponding 2D data mesh m_t is denoted as follows:

$$T(\mathbf{r}_t) = \begin{bmatrix} 0 & 0 & 0 & 0 & s_t^{22} & s_t^{23} & s_t^{24} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & s_t^{25} & s_t^{26} & s_t^{27} & s_t^{28} & s_t^{28} & 0 & 0 & 0 \\ 0 & s_t^{30} & s_t^{31} & s_t^{32} & s_t^{33} & s_t^{34} & s_t^{35} & s_t^{36} & s_t^{37} & s_t^{38} & 0 \\ 0 & s_t^{39} & s_t^{1} & s_t^{2} & s_t^{3} & s_t^{4} & s_t^{5} & s_t^{6} & s_t^{7} & s_t^{40} & 0 \\ s_t^{43} & s_t^{41} & s_t^{8} & s_t^{9} & s_t^{10} & s_t^{11} & s_t^{12} & s_t^{13} & s_t^{14} & s_t^{42} & s_t^{44} \\ 0 & s_t^{45} & s_t^{15} & s_t^{16} & s_t^{17} & s_t^{18} & s_t^{19} & s_t^{20} & s_t^{21} & s_t^{46} & 0 \\ 0 & s_t^{47} & s_t^{48} & s_t^{49} & s_t^{50} & s_t^{51} & s_t^{52} & s_t^{53} & s_t^{54} & s_t^{55} & 0 \\ 0 & 0 & 0 & s_t^{56} & s_t^{57} & s_t^{58} & s_t^{59} & s_t^{60} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & s_t^{61} & s_t^{62} & s_t^{63} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & s_t^{64} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where the positions of the null electrodes are padded with zeros. Through this transformation, the raw 1D data vector series [r_t, r_{t+1} ... r_{t+N}] is converted to the 2D data mesh series [m_t, m_{t+1} ... m_{t+N}]. During observation duration [t, t+N], the number of 2D data meshes is still (N +1). After 2D data mesh transformation, the data mesh is normalized across the non-zero elements using Z-score normalization. Each of the resulted 2D data meshes contains the spatial information of the brain activity at its recording time. During the recording process, some EEG readings are variably missing largely due to issues of electrical conductivity and subjects' movement, resulting in all channels recording zeros. This issue is unavoidable in sensor-based systems, and it might not be tolerated by BCIs. From the application point of view, smooth manipulation of the BCI system provides an improved user experience. For this reason, a BCI system should preferably translate brain activities to the output information continuously without interruption. As missing information is a clinical reality, in this work we preserve the incomplete recordings which are discarded in previous work to maintain the integrity of EEG signals. The experimental results show our 2D EEG meshes perform well in dealing with the “missing readings”.

Up to this point, we apply the sliding window approach to divide the streaming 2D meshes into individual clips as shown in the last step of Figure 1. Each clip has a fixed length of time series 2D data meshes with 50% overlapping between continuous neighbors. The data meshes segment S_j is created as follows:

$$S_j = [m_t, m_{t+1} \ldots m_{t+S-1}]$$

where S is the window size and j = 1, 2, ..., q with q segments during the observation period. Our goal is to develop an effective model to recognize a set of human intentions A = [a_1, a_2 ... a_K] ^{T} from each windowed data meshes segment S_j . The recognition approach tries to predict the human intention $Y_t \in$ A performed during this windowed period.

3.2 Self-supervised learning on preprocessed raw EEG data

In our work, a self-supervised vision transforms with DINO (*self-**di**stillation with **no** labels*) is used for unsupervised visual representation learning with images. The DINO will be fed with preprocessed 2D mesh-like matrices of size 224×224. Even though the previous phase returns 2D mesh-like matrices of size 10×11, we appended an extra row of zeros to the top of the 2D mesh-like matrix model to balance the number of rows and columns in the matrix and that are being resized into three-channel images of size 330×330. The DINO has been trained and evaluated with the ImageNet dataset and since the sizes of ImageNet dataset images are considered 224×224, we also considered the same sized images by taking resized crops while taking image augmentations according to global and local crop scales.

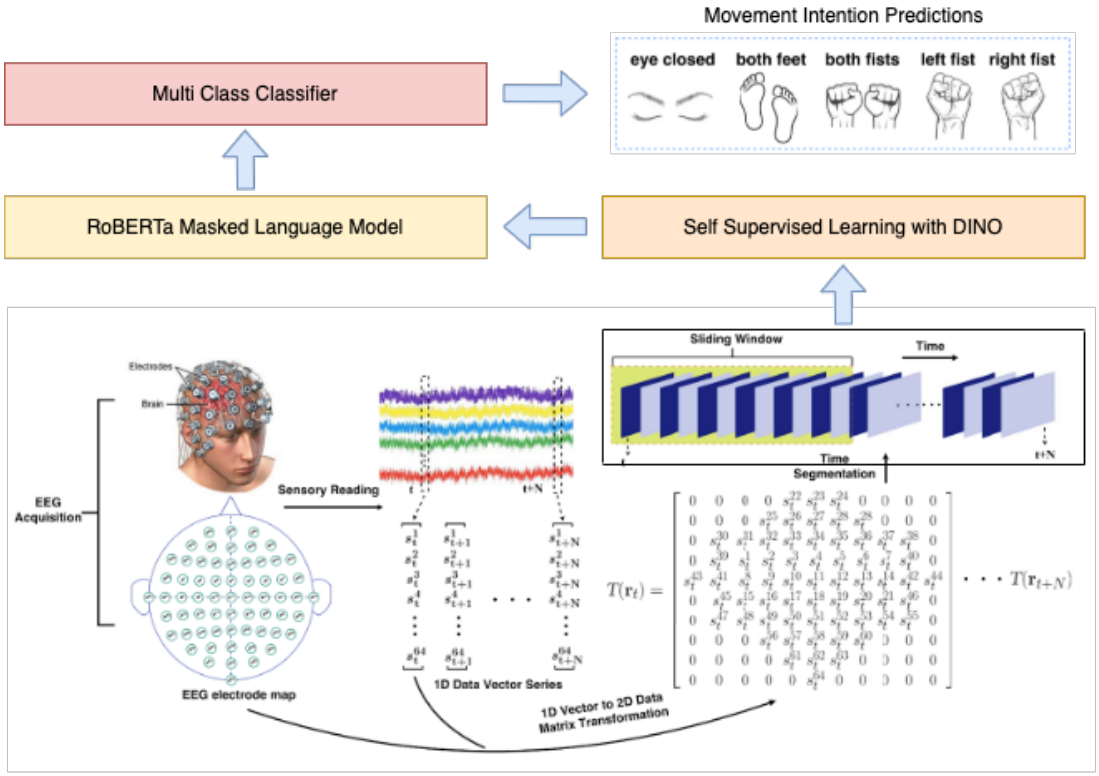

Figure 2. System architecture.

Our DINO training was performed in three attempts. For our first attempt of DINO training, we applied almost the same set of hyper-parameters (Table 1) that have been used for training ViT-S/8 with 21 million parameters except for the hyper-parameters, output dimension, batch size per GPU, number of epochs, global crop scale, and local crop scale. In the original DINO training, 65536 was taken as the output dimension but for our case, we have taken the output dimension as 300. In addition, in the original DINO training, global and local crop sizes (figure 3) have been taken as [0.4, 1.0] and [0.05, 0.4] respectively, but for our work, we have taken them as [0.75, 1.0] and [0.5,

0.75] respectively. For the DINO training, augmentations of different sizes according to global and local crop scales should be taken and since we are training the DINO with synthetic images that are having vague visual representations, augmentations were taken that covers a larger portion of an image and the DINO original research paper, two globle and ten local crops were taken to train this specific varient of DINO, therefore we have also taken the same number of crops to train our DINO model. According to [11], by taking augmentations of images of EEG representations (2 global resized crops and 10 local resized crops), the noises associated with EEG representations can be easily eliminated. It is because when training DINO, the model will give attention to the obvious visual representations while ignoring the noise associated with them.

In attempt 1, DINO training was performed with a single GPU for up to 100 epochs along with the batch size of 8 per GPU while having the limitations accompanied by the training environment. At the end of the DINO training, the model will return a prototype, a number in-between the range of output dimension. In this attempt, we were taken output dimension as 300, so that, the model will return a prototype from 0 to 299 for each input image. Actually, the output dimension is the feature dimension or the probability distribution of the output layer of DINO and prototypes will be derived by taking the index of the maximum arg value in the probability distribution.

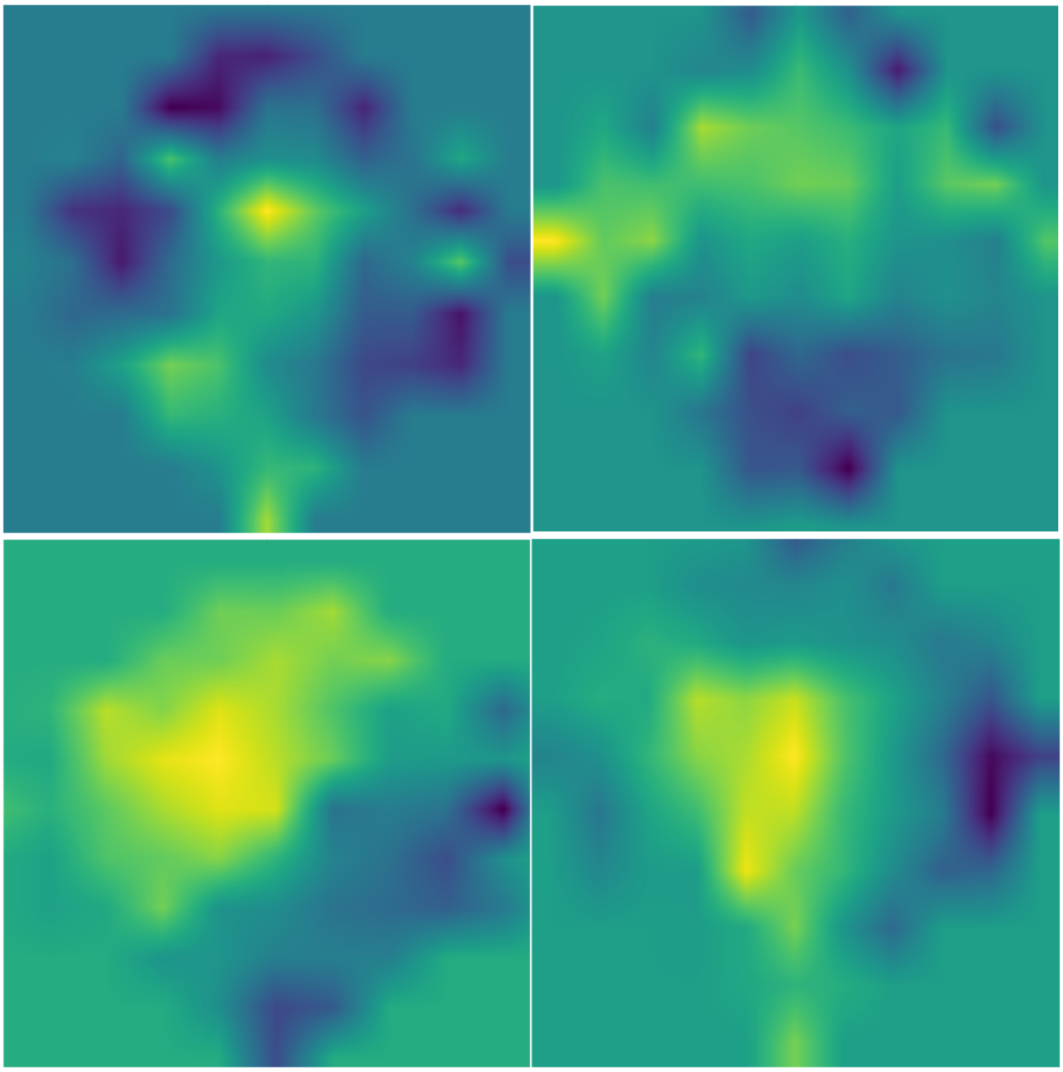

Figure 3. Randomly generated global resized crops.

Table 1. Comparison of default and our DINO hyper-parameters at attempt 1.

Hyper-parameter name	Default hyper-parameter	Our hyper-parameter
arch	vit_small	vit_small
patch_size	8	8
out_dim	65536	300
norm_last_layer	false	false
warmup_teacher_temp	0.04	0.04
teacher_temp	0.07	0.07
warmup_teacher_temp_epochs	30	30

use_fp16	false	false
weight_decay	0.04	0.04
weight_decay_end	0.4	0.4
clip_grad	3.0	3.0
batch_size_per_gpu	16	8
epochs	800	100
freeze_last_layer	1	1
lr	0.0005	0.0005
warmup_epochs	10	10
min_lr	1e-06	1e-06
global_crops_scale	[0.4, 1.0]	[0.75, 1.0]
local_crops_scale	[0.05, 0.4]	[0.5, 0.75]
local_crops_number	10	10
seed	0	0
num_workers	10	10
world_size	64	64
ngpus	8	8
nodes	8	8
optimizer	adamw	adamw
momentum_teacher	0.996	0.996
use_bn_in_head	false	false
drop_path_rate	0.1	0.1

In our second attempt of DINO training, almost all the hyperparameters (Table 2) used in the previous attempt were considered but the architecture used for student and teacher models and number of epochs were changed. According to the DINO original paper, they used three flavors of vision transformer architectures for their training. Those three flavors are ViT/tiny, ViT/small and ViT/base. For this attempt, we used the smallest architecture among later mentioned three flavors of archictures, that was ViT/tiny. The training with ViT/small at the previous attempt, it was taken much processing time with its model complexity and at this attempt with ViT/tiny, we could

be able to reduce the processing time considerably while training the model with 300 epochs.

Table 2. Comparison of default and our DINO hyper-parameters at attempt 2.

Hyper-parameter name	Default hyper-parameter	Our hyper-parameter
arch	vit_small	vit_tiny
patch_size	8	8
out_dim	65536	300
norm_last_layer	false	false
warmup_teacher_temp	0.04	0.04
teacher_temp	0.07	0.07
warmup_teacher_temp_epochs	30	30
use_fp16	false	false
weight_decay	0.04	0.04
weight_decay_end	0.4	0.4
clip_grad	3.0	3.0
batch_size_per_gpu	16	8
epochs	800	300
freeze_last_layer	1	1
lr	0.0005	0.0005
warmup_epochs	10	10
min_lr	1e-06	1e-06
global_crops_scale	[0.4, 1.0]	[0.75, 1.0]
local_crops_scale	[0.05, 0.4]	[0.5, 0.75]
local_crops_number	10	10
seed	0	0
num_workers	10	10
world_size	64	64
ngpus	8	8

nodes	8	8
optimizer	adamw	adamw
momentum_teacher	0.996	0.996
use_bn_in_head	false	false
drop_path_rate	0.1	0.1

In our third attempt of DINO training, almost all the hyperparameters (Table 3) used in the previous attempts (attempt 1 and attempt 2) were considered but the output dimension and number of epochs were changed. In our previous attempts, we considered 300 as the output dimension of student and teacher models but for this attempt, it was used as 10 and the training was conducted by 100 epochs. But the trained model of this attempt was not used in the further phases of the methodology pipeline because when we are reducing the feature dimensions of the student and teacher models, according to that, the feature learning of the models will be reduced as the models are not learning the details of the input images.

Table 3. Comparison of default and our DINO hyper-parameters at attempt 3.

Hyper-parameter name	**Default hyper-parameter**	**Our hyper-parameter**
arch	vit_small	vit_tiny
patch_size	8	8
out_dim	65536	300
norm_last_layer	false	false
warmup_teacher_temp	0.04	0.04
teacher_temp	0.07	0.07
warmup_teacher_temp_epochs	30	30
use_fp16	false	false
weight_decay	0.04	0.04
weight_decay_end	0.4	0.4
clip_grad	3.0	3.0
batch_size_per_gpu	16	8
epochs	800	100

freeze_last_layer	1	1
lr	0.0005	0.0005
warmup_epochs	10	10
min_lr	1e-06	1e-06
global_crops_scale	[0.4, 1.0]	[0.75, 1.0]
local_crops_scale	[0.05, 0.4]	[0.5, 0.75]
local_crops_number	10	10
seed	0	0
num_workers	10	10
world_size	64	64
ngpus	8	8
nodes	8	8
optimizer	adamw	adamw
momentum_teacher	0.996	0.996
use_bn_in_head	false	false
drop_path_rate	0.1	0.1

3.3 Prototype sequence generation

In this phase, raw EEG sequences will be transformed into prototype sequences. In our work, we have considered the output dimension as 300 (as the output dimension of attempt 1 and attempt 2) and the prototype sequences that are generated will be in the range of 0 - 300. Figure 4 is a snapshot of the generated prototype sequences for raw EEG sequence that belongs to subject 1 (out of 109 subjects) recode 3 (out of 14 experiments) which is referring to the experiment of opening/closing right fist or opening/closing left fist at attempt 1. At this phase, by using the respective trained DINO model which belongs to its attempt, prototype sequences will be generated for each raw EEG data recording in the dataset which we considered. Therefore, at the end of this phase, we have both the raw EEG data recordings with their respective prototype sequences.

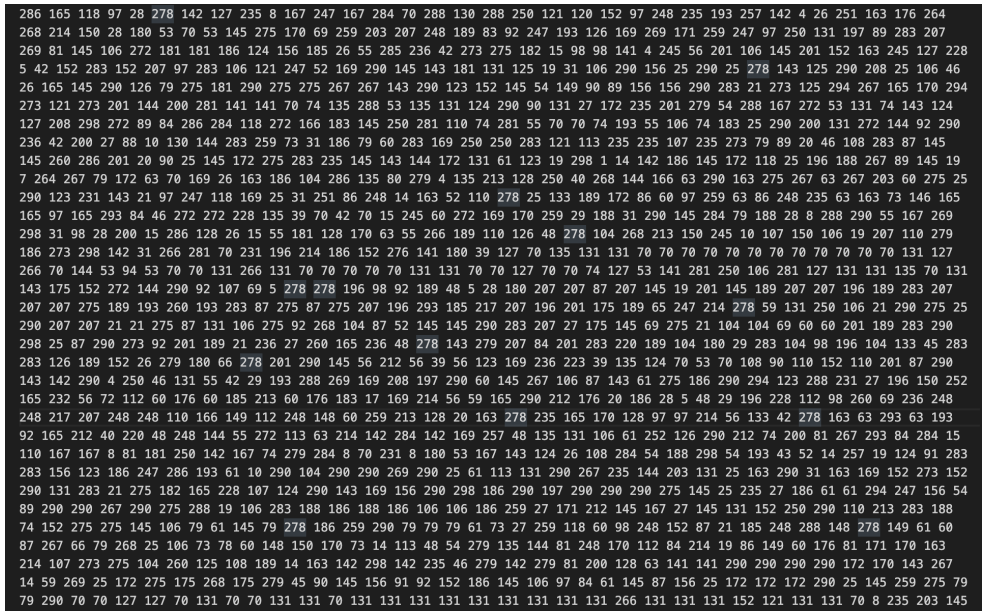

Figure 4. Snapshot of generated prototype sequence for subject 1 (out of 109 subjects) recode 3 (out of 14 experiments) at attempt 1

3.4 Masked language model on generated prototype sequences

In our work, phase 3 generates prototype sequences for given raw EEG sequences. So when considering the prototype sequences generated in phase 3, they can be considered as a brand new language that can be considered a language synthesized by the human brain. Therefore, we trained a new masked language model that is a variant of the BERT (Bidirectional Encoder Representations from Transformers)[26] called RoBERTa (A Robustly Optimized BERT Pretraining Approach)[27]. In this phase, a RoBERTa tokenizer and a RoBERTa masked language model have trained with the configurations of vocabulary size as 52000, 15 percent mask randomness, optimizer as AdamW, and learning rate of 1e-4.

3.5 EEG prototype sequence classification

This is the final phase of our work, which consists of training a multiclass prototype sequence classifier using RoBERTa. The tokenizer and the masked language model that was trained in the previous phase was used in this phase, and a size four linear layer of size was appended to the top of the stack to derive the probability distributions for classifications.

4. MAIN RESULTS

The main purpose of this research is to learn the hidden patterns of human brain activities by EEG raw data. For that, we trained self-supervised vision transformers with DINO to generate prototype sequences, and to validate the process along the phases 1, 2 and 3 which explaind at the methodology section above, we developed and trained a brand new masked language model and a sequence classifier as a downstream task.

The training DINO in phase 2 was performed by 40,000 images. Those images were taken by preprocessing 3rd and 5th (out of 14 experiments) recodes of subject 1 (out of 109 subjects) and the training was done at three attempts. Figure 5, 6 and 7 show the performance of DINO training at attempt 1, 2 and 3 respectively. Since we are following self supervised learning at phase 2, we do not have a way to validate the performance of the trained DINO models using validation datasets at each attempt.

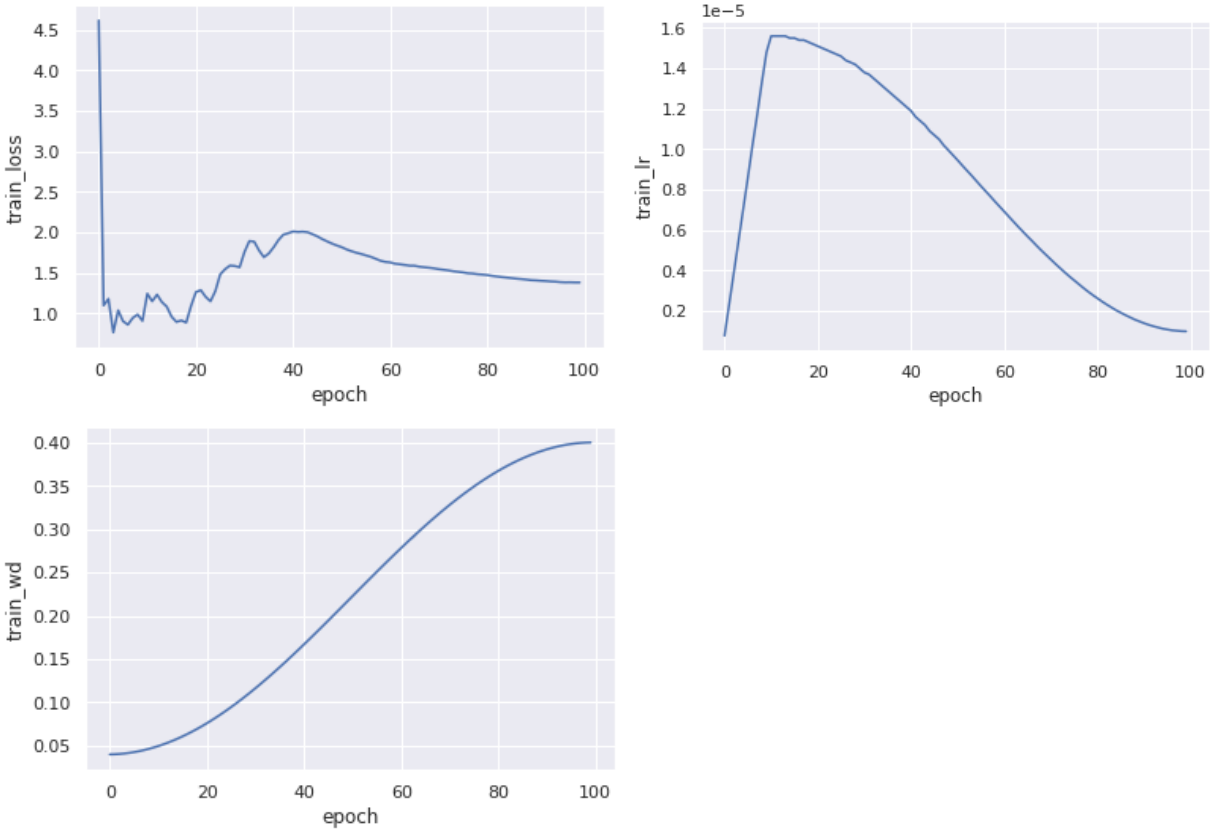

Figure 5. Behaviour of train loss, train learning rate and train weight decay of DINO training at attempt 1

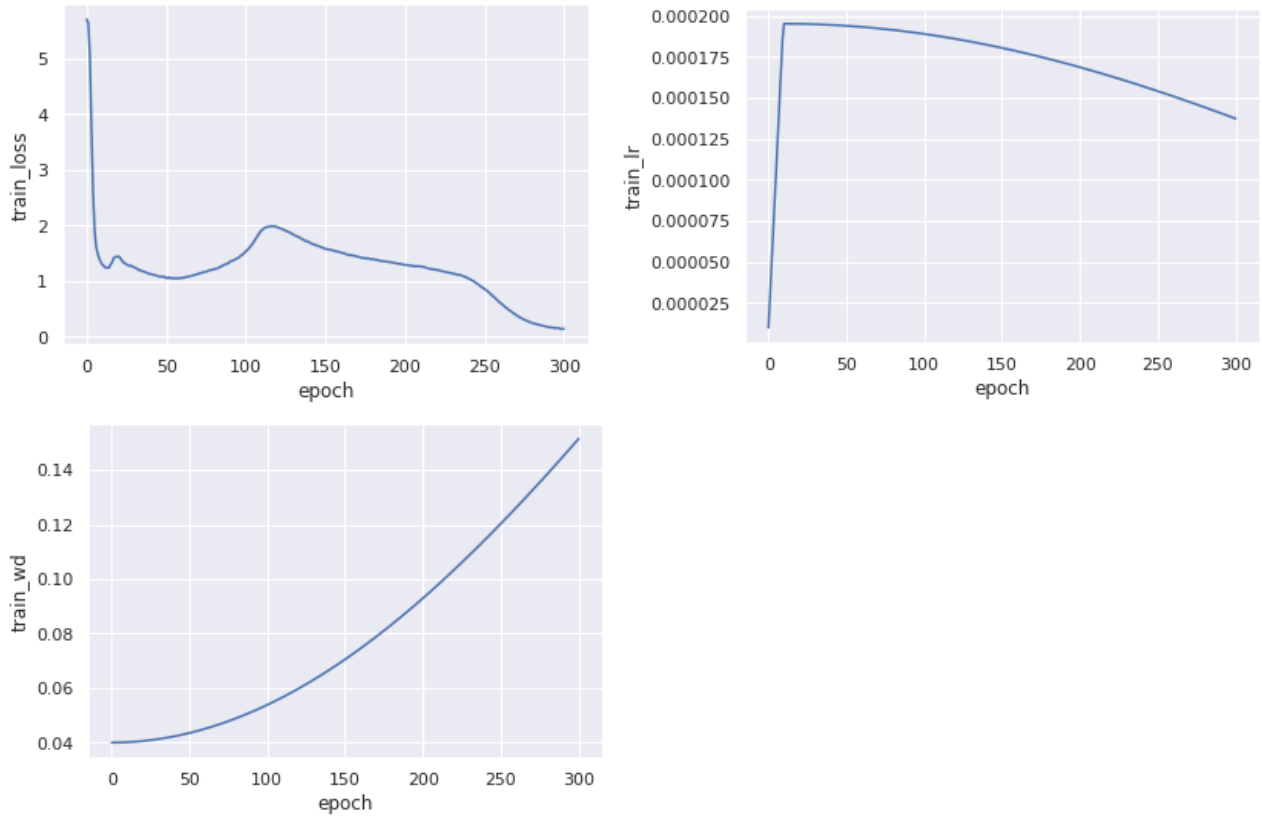

Figure 6. Behaviour of train loss, train learning rate and train weight decay of DINO training at attempt 2

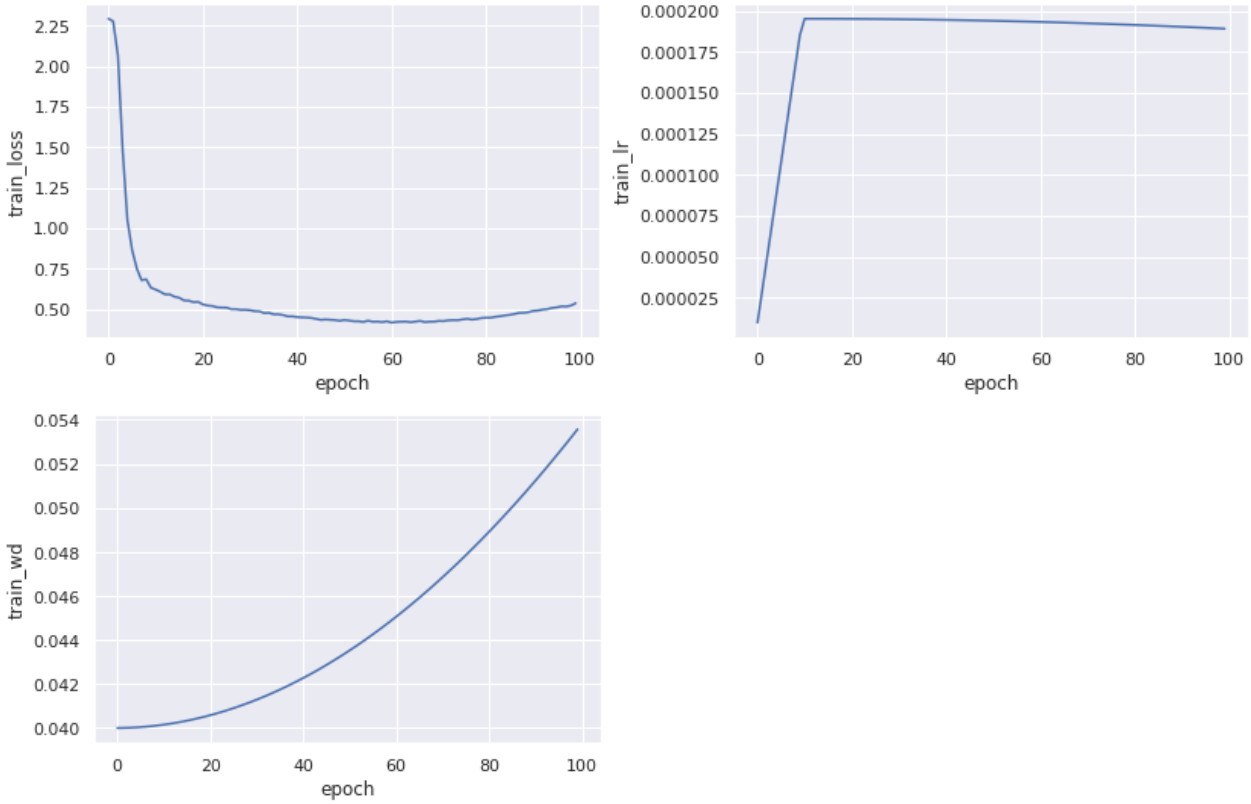

Figure 7. Behaviour of train loss, train learning rate and train weight decay of DINO training at attempt 3

Therefore, to validate our methodology (from phase 1 to phase 3 along the methodology pipeline), after training for 100 and 300 epochs respectively at attempt 1 and 2, prototype sequences were generated and they were used to train a brand new masked language model and the multiclass prototype sequence classifier for respective attempts (only attempt 1 and 2 have been considered). Figure 8 and 9 shows the performance of the multiclass prototype sequence classifiers.

Figure 8. Training and validating multiclass prototype sequence classifier at attempt 1.

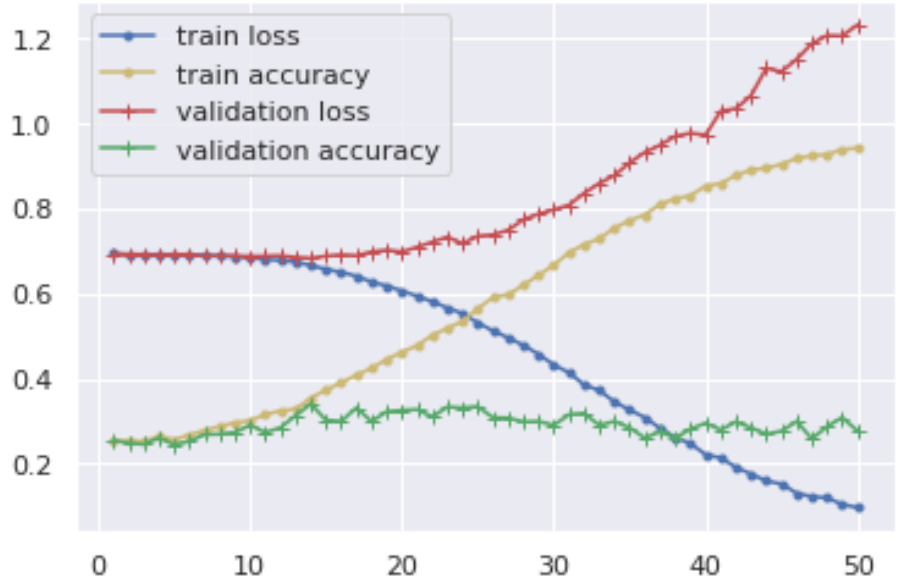

Figure 9. Training and validating multiclass prototype sequence classifier at attempt 2.

According to figure 8 and 9, BERT based trained multiclass prototype sequence classifier models got overfitted. That can be observed from the train loss was reduced and the train accuracy was increased by the training but the validation loss was increased and the validation accuracy showed constant behavior at both of the attempts (attempt 1 and 2).

5. CONCLUSION

In this research, we propose a method for self supervised learning of EEG raw data to learn the hidden patterns of human brain activities. This work was performed through a pipeline consisting of five phases. Each of the phase's output will be the input for the next phase. Phase 1 is for preprocessing raw EEG sequences into EEG representations that catch the spacial and temporal properties in the original raw EEG sequences. We have followed a relatively less complex method to preprocess raw EEG sequences. In phase 2, preprocessed raw EEG sequences will be learnt by self supervised representation learning. For that self supervised vision transformers with DINO will be used. These vision transformers models are computationally more demanding and require more training data therefore more computational resources and training data will be needed. So that at the presence of more training data and computational processing power, self supervised vision transformer architectures will be expected to produce the best results while outperforming supervised learning architectures. Then at the phase 3, sequences of prototypes for each raw EEG data sequence of the dataset will be generated. To evaluate the prototypes that are generated from raw EEG data, phase 4 and 5 have been used as the downstream task for the self supervised learning task. For phase 4 and 5, we agained used a transformer architecture, that is a BERT based model called RoBERTa to learn the synthetic language generated by phase 3 or to learn the context and the language of generated prototype sequences and by performing a muliti class prototype sequence classification, prototype generation for each representation at specific time stamp of raw EEG data sequence can be evaluated. We believe that since the models are computationally demanding and require more training data, the latter explained pipeline of five phases should be improved with more training and performing hyperparameter tuning at a high computational resources and data rich environment. As future works, phase 2 could be trained with different output dimension which ranges from smaller to larger feature dimensions as a task of hyperparameter tuning and that may output prototypes which are more and more sensitive to raw EEG data sequences. In addition, our concept could be evaluted to test out whether different EEG datasets with different number of electrodes are supported with the involvement of spatio-temporal preserving representation of raw EEG data sequences. Finally, generalized prototypes which preserves spacial property can be used to represent any or domain specific raw EEG data sequences to support different tasks that are related to raw EEG data sequence feature learnings.

REFERENCES

1. Bashivan, P., Rish, I., Yeasin, M. and Codella, N., 2015. Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.
2. Ruffini, G., Ibañez, D., Kroupi, E., Gagnon, J.F., Montplaisir, J., Castellano, M. and Soria-Frisch, A., 2018. Deep learning using EEG spectrograms for prognosis in idiopathic rapid eye movement behavior disorder (RBD). *bioRxiv*, p.240267.
3. Pailla, T., Miller, K.J. and Gilja, V., 2019. Autoencoders for learning template spectrograms in electrocorticographic signals. *Journal of neural engineering*, *16*(1), p.016025.
4. Ranieri, C.M., Moioli, R.C., Romero, R.A., de Araújo, M.F., De Santana, M.B., Pimentel, J.M. and Vargas, P.A., 2020, July. Unveiling Parkinson's Disease Features from a Primate Model with Deep Neural Networks. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
5. Zhang, D., Yao, L., Zhang, X., Wang, S., Chen, W., Boots, R. and Benatallah, B., 2018, April. Cascade and Parallel Convolutional Recurrent Neural Networks on EEG-based Intention Recognition for Brain Computer Interface. In *AAAI* (pp. 1703-1710).
6. Zhang, D., Yao, L., Chen, K., Wang, S., Chang, X. and Liu, Y., 2019. Making sense of spatio-temporal preserving representations for EEG-based human intention recognition. *IEEE transactions on cybernetics*.
7. Chen, J., Jiang, D., Zhang, Y. and Zhang, P., 2020. Emotion recognition from spatiotemporal EEG representations with hybrid convolutional recurrent neural networks via wearable multi-channel headset. *Computer Communications*, *154*, pp.58-65.
8. Li, Y., Yang, H., Li, J., Chen, D. and Du, M., 2020. EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM. *Neurocomputing*, *415*, pp.225-233.
9. Yang, Y., Wu, Q., Qiu, M., Wang, Y. and Chen, X., 2018, July. Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.
10. Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.A. and Gramfort, A., 2021. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, *18*(4), p.046020.
11. Zhang, Z., Zhong, S.H. and Liu, Y., 2021. GANSER: A Self-supervised Data Augmentation Framework for EEG-based Emotion Recognition. *arXiv preprint arXiv:2109.03124*.
12. Jiang, X., Zhao, J., Du, B. and Yuan, Z., 2021, July. Self-supervised Contrastive Learning for EEG-based Sleep Staging. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

13. Banville, H., Albuquerque, I., Hyvärinen, A., Moffat, G., Engemann, D.A. and Gramfort, A., 2019, October. Self-supervised representation learning from electroencephalography signals. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.
14. Kostas, D., Aroca-Ouellette, S. and Rudzicz, F., 2021. BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, *15*.
15. Gramfort, A., Banville, H., Chehab, O., Hyvärinen, A. and Engemann, D., 2021, February. Learning with self-supervision on EEG data. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)* (pp. 1-2). IEEE.
16. Tang, S., Dunnmon, J., Saab, K.K., Zhang, X., Huang, Q., Dubost, F., Rubin, D. and Lee-Messer, C., 2021, September. Self-Supervised Graph Neural Networks for Improved Electroencephalographic Seizure Analysis. In *International Conference on Learning Representations*.
17. Xiao, Q., Wang, J., Ye, J., Zhang, H., Bu, Y., Zhang, Y. and Wu, H., 2021, June. Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1290-1294). IEEE.
18. Xu, J., Zheng, Y., Mao, Y., Wang, R. and Zheng, W.S., 2020, December. Anomaly Detection on Electroencephalography with Self-supervised Learning. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 363-368). IEEE.
19. Caron, M., Bojanowski, P., Mairal, J. and Joulin, A., 2019. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2959-2968).
20. Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., 2020, November. A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.
21. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. and Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, *33*, pp.9912-9924.
22. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. and Piot, B., 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, pp.21271-21284.
23. He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729-9738).

24. Chen, X., Fan, H., Girshick, R. and He, K., 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
25. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9650-9660).
26. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
28. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.