Enhancing Sentiment Analysis on SEED-IV Dataset with Vision Transformers: A Comparative Study

Imad Eddine TIBERMACINE
tibermacine@diag.uniroma1.it
Department of Computer, Automation and Management Engineering,
Sapienza University of Rome
Rome, Italy

Ahmed TIBERMACINE
Department of Computer Science,
University of Biskra
Biskra, Italy
ahmed.tibermacine@univ-biskra.dz

Walid GUETTALA
Department of Computer Science,
University of Biskra
Biskra, Algeria
walid.guettala@gmail.com

Christian NAPOLI
Department of Computer, Automation and Management Engineering,
Sapienza University of Rome
Rome, Italy
Institute for Systems Analysis and Computer Science, Italian National Research Council
Rome, Italy
c.napoli@diag.uniroma1.it

Samuele Russo
Department of Psychology, Sapienza University of Rome
Rome, Italy
samuele.russo@uniroma1.it

ABSTRACT

This paper introduces a new approach to emotion classification utilising deep learning models, specifically the Vision Transformer (ViT) model, in the analysis of electroencephalogram (EEG) signals. A dual-feature extraction approach was implemented in our study, utilising Power Spectral Density and Differential Entropy, to analyse the SEED IV dataset. This methodology resulted in the detailed classification of four distinct emotional states. The ViT model, which was originally designed for image processing, has been successfully applied to EEG signal analysis. It demonstrated remarkable performance by attaining a test accuracy of 99.02% with little variance. Notably, it outperformed conventional models like GRUs, LSTMs, and CNNs in this context. The findings of our study indicate that the ViT model has a high level of effectiveness in accurately identifying complex patterns present in EEG data. Specifically, the precision and recall rates achieved by the model surpass 98%, while the F1 score is estimated to be about 98.9%. The results of this study not only demonstrate the efficacy of transformer-based models in analysing cognitive states, but also indicate their considerable potential in improving systems for sympathetic human-computer interaction.

KEYWORDS

Deep Learning, Vision Transformer, EEG, Signal Analysis, Classification

This work is licensed under a Creative Commons Attribution International 4.0 License.

ICIT 2023, December 14–17, 2023, Kyoto, Japan
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0904-3/23/12.
https://doi.org/10.1145/3638985.3639024

ACM Reference Format:
Imad Eddine TIBERMACINE, Ahmed TIBERMACINE, Walid GUETTALA, Christian NAPOLI, and Samuele Russo. 2023. Enhancing Sentiment Analysis on SEED-IV Dataset with Vision Transformers: A Comparative Study. In *2023 The 11th International Conference on Information Technology: IoT and Smart City (ICIT 2023), December 14–17, 2023, Kyoto, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3638985.3639024

1 INTRODUCTION

The role of emotion is crucial in all aspects of human life, including communication, decision-making, and interactions between humans and machines. Emotion has a significant impact on a wide range of daily activities, such as interpersonal interactions, learning, and work. The domain of affective computing, namely the identification of human emotions, has garnered significant attention in scholarly investigations owing to its crucial implications in several domains such as affective brain-computer interaction, emotion regulation, and the diagnosis of illnesses associated with emotions. The notion aims to advance the development of systems that possess the ability to identify, comprehend, analyse, and replicate human emotions [16][15]. In the present day, there has been a notable shift in the field towards the use of multimodal approaches for emotion recognition. This involves the integration of physiological data obtained from electroencephalography (EEG) with eye movement features, resulting in the development of more resilient and reliable models. The utilisation of this integration takes advantage of the inherent, automatic reactions of the central nervous system, which are more resistant to manipulation in comparison to deliberate manifestations like facial expressions or vocal intonations [11][20][27] [6]. The utilisation of EEG signals has gained popularity due to the increased dependability facilitated by the introduction of consumer-grade, non-invasive, and cost-effective wearable sensors. This has resulted in their preference over outward manifestations, which are susceptible to manipulation and influenced by external factors [21]. The comprehension of emotions is a challenge due to its intricate nature, encompassing subjective perception, outward manifestations, and physiological responses.[17] Attempting to capture emotions solely through signals of a single modality proves to be insufficient. The research of multimodal data has been prompted by its ability to provide a full viewpoint on emotional changes, therefore enabling the development of emotion detection algorithms that are more precise and dependable[28][20]. The incorporation of multimodal techniques, encompassing physiological inputs that pose difficulties for users to consciously manipulate, has demonstrated considerable potential as a technique for emotion identification[12][19]. Previous research has provided evidence about the effectiveness of electroencephalography (EEG) in the classification of emotions, with attempts made to reduce the number of electrodes used while still achieving a high level of accuracy in emotion recognition [8][25]. In addition, the relationship between emotions and eye movement, specifically alterations in pupil size, highlights the possibility of integrating behavioural modalities with physiological information to gain a more comprehensive comprehension of emotional states[26]. The present direction of research in affective computing is marked by a collective endeavour to use the combined potential of multimodal data. The objective of this endeavour is to develop emotion identification algorithms that exhibit both high accuracy and resilience in the face of the natural diversity observed in human emotional expression[5]. The potential of these systems is in their utilisation across several domains, encompassing the personalization of user interactions and the progression of mental health assessments, signifying the advent of a novel era in the realm of emotionally intelligent computing[1].

The discipline of emotion recognition, which encompasses the study of cognitive processes and psychophysiological alterations, is making progress by using multimodal data. The intricate nature of human emotions requires the utilisation of various modalities in order to attain a comprehensive comprehension. Multimodal emotion identification systems, which integrate several data including EEG and eye movement, have demonstrated potential in improving precision and dependability[3][22]. The integration of these signals is of utmost importance, as several methods such as feature-level concatenation and decision-level fusion have shown the synergistic relationship between distinct modalities[28].

Within the domain of multimodal representations, the utilisation of joint and coordinated techniques presents unique methodologies for the purpose of emotion recognition. The phenomenon described has been seen in the research conducted by [11], who employed EEG and eye movement data to identify emotions. This line of inquiry has been further developed by [10] and [23], who utilised deep learning models such as Bimodal Deep AutoEncoders. Notwithstanding these advancements, there is still ample room for additional investigation in the realm of coordinated representations, wherein signals are processed separately but with an emphasis on inter-modal similarity.

The majority of EEG-based emotion recognition research has focused on healthy datasets like AMIGOS, DEAP, DREAMER, and SEED-IV. Deep learning algorithms for EEG data-based emotion mapping in Parkinson's disease (PD) patients are understudied, creating a substantial research opportunity. It's difficult to collect and label EEG data for emotion identification in PD patients. To investigate the stability of deep learning models like CRNN and ELM-based networks for clinical emotion recognition, our study uses a private dataset of PD patients to trigger emotions [13]. This revolutionary application enhances multimodal emotion identification systems and psychologically profiles cognitive problem patients, revealing their emotional landscape.

In this paper, we advance the methodology of emotion classification by integrating it as a feedback loop within a deep reinforcement learning system. This feedback loop is pivotal for providing insights into the human emotional state as derived from simulated scenarios. To facilitate this, a diverse array of models was harnessed to segregate brain signals into four discrete emotional categories: anger, sadness, happiness, and neutrality. The suite of models utilized encompasses a broad spectrum of computational techniques, including Convolutional Neural Networks (CNN), Vision Transformers (ViT), Gated Recurrent Units (GRUs), Long Short-Term Memory networks (LSTMs), EEGNET, FBCNet, and FBCCNN. The crux of our feature extraction process lies in the application of Power Spectral Density (PSD) and Differential Entropy (DE) to the EEG data. PSD is employed to analyze the power distribution over five distinct frequency bands, while DE is leveraged to compute the entropy differentials within each EEG segment and across the bands. This meticulous feature extraction is a cornerstone of our approach, enabling the nuanced classification of emotional states.

The main contributions of our research are summarized as follows:

1) We propose a novel approach that uses two different feature modalities, Frequency-Domain and Entropy, to recognize emotions on SEEDIV dataset.

2) Our proposed model is a novel approach to include attention mechanisms in the context of multiple model time series data tasks.

3) We achieved a higher accuracy compared to state of the art work.

2 DATASET

The SEED-IV dataset is an expanded version of the original SEED series, providing a comprehensive collection of EEG and eye movement signals that can be utilised for the development of sophisticated affective computing models. The purpose of this study is to accurately measure and analyse four distinct emotional states: happy, sorrow, neutrality, and fear. This was achieved through a meticulously planned and executed controlled experiment, which involved the participation of 15 individuals (comprising 7 males and 8 females). These participants were actively engaged in the study and supplied their replies throughout three separate sessions. Each experimental session consisted of a total of 24 trials. Each trial involved the presentation of one of the 72 meticulously chosen video clips, each lasting roughly 2 minutes. A 5-second baseline period was included before each film clip to establish a reference point. The primary objective of these film clips was to evoke specific emotions in the participants. The electroencephalogram (EEG) data was first collected at a sampling rate of 1000 Hz using a 62-channel ESI Neuro scan system. Subsequently, the data was downsampled to 200 Hz in order to decrease complexity while maintaining the integrity of the data. In order to maintain compatibility with other datasets such as AMIGOS and PD, the data was resampled to a frequency of 128 Hz. The participants in the study engaged in self-annotation of their emotional states by utilising the Positive and Negative Affect Schedule (PANAS) measures. This process contributed an additional subjective component to the dataset. The comprehensive and multi-dimensional approach to emotion recognition in SEED-IV, which encompasses EEG and eye movement data, facilitates the integration of neuroscience and artificial intelligence, hence strengthening their junction. The SEED-IV dataset is subject to ethical criteria and usage agreements, as described in the foundational publication[29]. Its applications span a wide range, including the evaluation of emotion recognition algorithms and the enhancement of human-computer interaction systems[7].

Figure 1: Different visual emotional stimulations used to collect the dataset

3 EXPERIMENTAL SETUP

3.1 Preprocessing

The first step in the preprocessing pipeline consisted of downsampling the EEG data to a sampling rate of 200 Hz. In more accessible language, downsampling might be likened to reducing the frequency of capturing images of the brain's electrical activity. The implementation of a reduced quantity of snapshots is a pragmatic strategy aimed at mitigating the burden of data, hence facilitating more efficient and expedited processing, but yet retaining essential details pertaining to the brain's electrical patterns that are pivotal for our analytical endeavours. After performing downsampling, we proceeded to apply a bandpass filter to the electroencephalogram (EEG) signals. Conceptualise the electrical activity within the brain as a symphony characterised by a diverse array of auditory stimuli, wherein each distinct note corresponds to a specific frequency. Certain auditory stimuli in our investigation exhibit frequencies that are either excessively low or excessively high, hence lacking relevance. For instance, the profound reverberation of thunder, characterised by low-frequency noise, or the shrill, high-pitched chirping emitted by a cricket, denoting high-frequency noise. A bandpass filter might be likened to a discerning listener that exclusively focuses on the specific range of sounds (frequencies) that hold significance. By configuring our filter to permit the transmission of frequencies ranging from 1 Hz to 75 Hz, we effectively disregarded extraneous noises and directed our attention onto the frequencies that hold the most significance in relation to emotional processing within the brain.

The preprocessing stages play a critical role in the cleansing of EEG data, hence enhancing the accuracy and fidelity of the subsequently extracted information pertaining to the actual brain activity.

3.2 Feature Extraction

Once the EEG data has undergone preprocessing and cleaning procedures, the subsequent step involves feature extraction, which can be regarded as the central aspect of our study. In this context, meticulous selection and precise calculation of certain metrics are conducted on EEG data to provide valuable insights into the neural activity associated with emotional states.

One of the key elements that is extracted is referred to as Power Spectral Density (PSD). In essence, the power spectral density (PSD) enables us to comprehend the magnitude of power, or energy, exhibited by the brain's electrical activity across various frequencies, or rates of brain wave oscillations. Similar to how various musical instruments are capable of producing notes at different pitches, the brain exhibits distinct waves that oscillate at varying frequencies. Power spectral density (PSD) provides quantitative information about the amplitude or strength of various brain wave frequencies. As an illustration, a frequency range known as 'Delta' (1-4 Hz) may manifest as a low-frequency resonance, signifying a state of profound slumber or relaxation. Conversely, a frequency range referred to as 'Beta' (14-31 Hz) could be characterised by a higher tempo, denoting heightened mental activity or a state of unease, such as intense cognitive processing or anxiety.

Another important aspect that we consider is Differential Entropy (DE). The concept of power spectral density (PSD) pertains to the magnitude of neural signals in the brain, while differential entropy (DE) focuses on the intricacy or complexity of these signals. This provides insight into the degree of predictability or unpredictability exhibited by the brain's neural activity. A very foreseeable and regular sequence would exhibit a diminished level of differential entropy, akin to the rhythmic and unvarying sound produced by a continuous drumming. Conversely, a highly intricate and uncertain arrangement, such as a jazz solo, would exhibit a substantial degree of differential entropy. By examining the differential entropy (DE) over several frequency bands, one can gain insights on the level of organisation or randomness in the brain's activity during distinct emotional states.

In the present investigation, we employed the aforementioned feature extraction techniques on the SEED IV dataset, a dataset particularly curated for the purpose of analysing emotions through electroencephalography (EEG). PSD and DE features were derived using the data included within the 'eeg_feature_smooth' directory of the given dataset. This process facilitated the conversion of the unprocessed EEG data into a structured format suitable for our models to acquire knowledge about the distinctive patterns linked to various emotional states.

The objective of this study is to construct a model that can effectively discern the emotional state of an individual by analysing the brain's electrical activity, with a particular emphasis on the identified qualities.

3.3 Vision Transformer (ViT)

The Vision Transformer (ViT) [4] has emerged as a groundbreaking architecture in the field of computer vision, drawing inspiration from the transformative success of the Transformer model in natural language processing [24]. This innovative approach reimagines image classification by treating images as sequences of discrete elements, analogous to the parsing of text into a series of words. In the case of ViT, an image is partitioned into a grid of fixed-size patches, each of which is subsequently flattened and transformed into a one-dimensional token through linear projection[24]. These tokens are then concatenated with positional embeddings, a critical step that embeds the spatial coordinates of each patch, thereby preserving the two-dimensional topological information within the one-dimensional sequence.

The ensemble of patch embeddings, now analogous to a sentence composed of words, is processed through the Transformer encoder—a sophisticated stack of layers renowned for its self-attention mechanism[14]. This mechanism, central to the Transformer's architecture, enables the model to weigh the influence of all other patches when encoding a particular patch, thereby capturing the global context of the image in a manner that is dynamically contingent on the content of the image itself. The mathematical underpinning of this process is encapsulated in the self-attention formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where the matrices Q,K, and V correspond to queries, keys, and values, respectively, and Dk represents the dimensionality of the key vectors. This formula reflects the computation of attention weights, which are used to scale the value vectors, effectively allowing the model to focus on the most salient parts of the image[4].

Further enhancing the model's capacity to capture diverse aspects of the image, the multi-head attention mechanism divides the attention process into multiple 'heads', each of which attends to different parts of the patch sequence. This is mathematically represented as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

The outputs of these heads are then concatenated and linearly transformed, providing a rich, composite representation of the image's features.

Subsequent to the attention layers, the Transformer employs MLP blocks, which consist of dense layers with non-linear activation functions, such as the Gaussian Error Linear Unit (GELU). These blocks serve to introduce non-linearity into the model, allowing for the capture of complex patterns within the data. The operation of an MLP block is captured by the following equation:

$$\text{MLP}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \tag{4}$$

Here, x signifies the input to the MLP block, W1 and W2 are the weight matrices of the first and second linear transformations, respectively, and b1 and b2 are the bias vectors. The final stage of the ViT is the classification head, which typically comprises a simple linear layer that projects the Transformer encoder's output to the label space, thus yielding the final class predictions for the image[4].

The original exposition of the Vision Transformer by [4] in their seminal paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" not only elucidates the architecture in meticulous detail but also validates its performance against established benchmarks, showcasing its superiority in various image classification tasks. The adaptability of ViT to a broad spectrum of applications, including the classification of EEG signals, is particularly noteworthy. By converting EEG data into a format amenable to the ViT, such as time-frequency representations, researchers can leverage the model's ability to discern intricate patterns associated with different cognitive and emotional states, thereby advancing the frontiers of both neuroscience and artificial intelligence.

Incorporating the Vision Transformer into EEG signal classification endeavors holds the promise of harnessing its powerful attention-driven mechanism to unravel the complex spatial-temporal dynamics inherent in brain activity data. The potential of ViT in this domain is under active exploration, with researchers seeking to adapt and optimize its architecture to accommodate the unique characteristics of EEG signals, thereby opening new avenues for the interpretation and classification of neural patterns[4].

Table 1: Hyperparameters of the used architecture

Component	Value
Input Chunk Size	20
Grid Size	9×9
Temporal Patch Size	10
Number of Classes	4
Loss Function	CrossEntropyLoss
Batch Size	128
Epochs	100
Number of Trials	3
Device	GPU

The Vision Transformer (ViT) processes high-dimensional, temporal EEG data well in the suggested design. Given the temporal granularity needed to capture EEG signal changes, the model uses input data divided into 20 chunks. The grid configuration of 9x9 accommodates the spatial interdependence of EEG data, mimicking the common architecture of EEG electrodes. These chunks contain temporal patches with 10 time points each to teach the model about EEG signal temporal evolution. The design is ready for a classification job with four discrete output classes, which represent cognitive states or event-related potentials examined in neuroscience. The standard CrossEntropyLoss function is used for category training. In batches of 128 data over 100 epochs, the model is trained to optimise learning and computing efficiency. The training routine is repeated three times to test the model's robustness and generalizability across data subsets. The model's

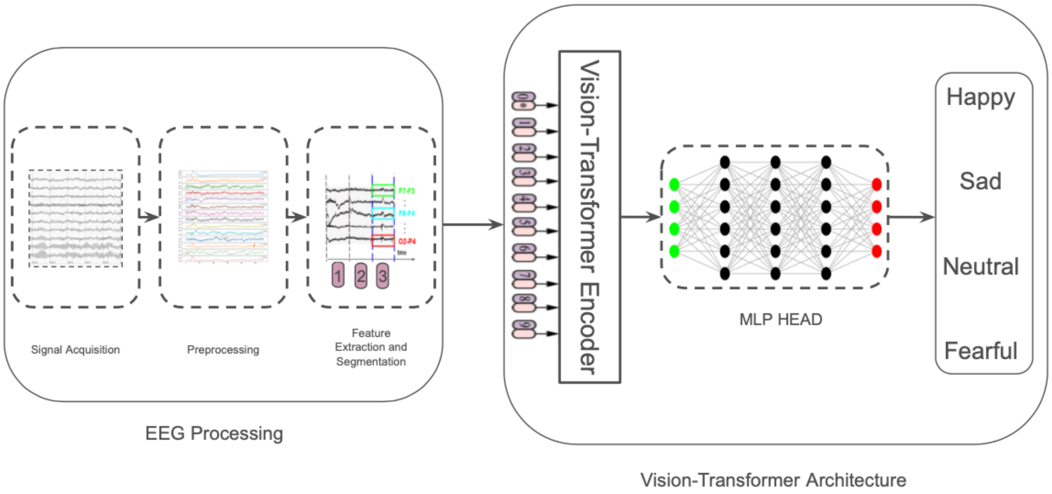
Signal Acquisition
Preprocessing
Feature Extraction and Segmentation
1
2
3
EEG Processing
0
1
2
3
4
5
6
7
8
9
Vision-Transformer Encoder
MLP HEAD
Happy
Sad
Neutral
Fearful
Vision-Transformer Architecture

Figure 2: The pipeline of classification using ViT Model

computational flexibility allows GPU deployment depending on computational capabilities. This architecture is carefully designed to use the transformer model's ability to detect complicated patterns in multidimensional EEG data for advanced brain-computer interface classification tasks.

4 RESULTS AND DISCUSSION

In the presented results [Table 1], the performance metrics of the models trained on the SEED IV dataset are meticulously tabulated. Each model underwent a tripartite training regimen, with the ensuing outcomes expressed as mean and variance of the accuracy scores. The Vision Transformer (ViT) emerged as the most accurate model, consistently surpassing the 90% threshold.

The ViT model demonstrated exceptional proficiency, with an average training accuracy of 99.67% coupled with a validation accuracy of 98.60%. This remarkable efficacy is likely a consequence of the model's inherent architectural features. Originally conceived for image classification, the ViT leverages self-attention mechanisms to discern global interdependencies within the input data. This attribute is particularly beneficial for identifying and extracting salient patterns and features pertinent to the SEED IV dataset, which is critical for emotion recognition tasks.

Table 2: Training Loss and Accuracy

Model	Train Loss	Train Accuracy (%)
GRU*	$0.0063 \pm 4 \times 10^{-8}$	$67.15 \pm 8 \times 10^{-8}$
LSTM*	$0.0062 \pm 2 \times 10^{-8}$	$68.02 \pm 6 \times 10^{-8}$
CNN*	$0.0108 \pm 5 \times 10^{-8}$	$65.9 \pm 3 \times 10^{-6}$
FBCNet*	$0.0098 \pm 6 \times 10^{-8}$	$27.54 \pm 3 \times 10^{-8}$
FBCCNN*	$0.0001 \pm 5 \times 10^{-8}$	$39.645 \pm 6 \times 10^{-8}$
ViT*	$0.0003 \pm 5 \times 10^{-8}$	$99.67 \pm 5 \times 10^{-8}$

Table 3: Validation and Test Accuracy

Model	Validation Accuracy (%)	Test Accuracy (%)
GRU*	$65.05 \pm 1 \times 10^{-7}$	$65.9 \pm 3 \times 10^{-6}$
LSTM*	$65.33 \pm 1 \times 10^{-7}$	$68.2 \pm 2 \times 10^{-6}$
CNN*	$65.9 \pm 3 \times 10^{-6}$	$65.9 \pm 3 \times 10^{-6}$
FBCNet*	$26.87 \pm 2 \times 10^{-7}$	$28.32 \pm 3 \times 10^{-6}$
FBCCNN*	$38.645 \pm 2 \times 10^{-7}$	$38.5 \pm 5 \times 10^{-6}$
ViT*	$98.60 \pm 2 \times 10^{-7}$	$99.02 \pm 2 \times 10^{-6}$
CAN [2]	/	87.71 ± 9.74
SVM [18]	/	75.88 ± 16.14
PR-PL [30]	/	85.56 ± 4.78
DCCA [9]	/	87.45 ± 9.23

The empirical evidence suggests that the ViT model's superior performance is a direct result of its sophisticated architecture, which is adept at capturing and representing the intricate EEG signal patterns associated with emotional states. Consequently, the ViT model has demonstrated a pronounced advantage over competing models in terms of accuracy on the SEED IV dataset, underscoring the significance of its design in understanding complex data relationships.

The comparative analysis delineated in the accompanying table juxtaposes the accuracy of models developed in-house (denoted by an asterisk) against those cited from extant literature. The tabulation elucidates the impact of distinct features on the resultant accuracies, providing a basis for a detailed examination of performance metrics and feature efficacy.

Within the cohort of models we trained, the Vision Transformer (ViT) model was preeminent, achieving a test accuracy of 99.02% with a minimal standard deviation of 0.000002, thereby indicating a consistently high performance across trials. This model's preponderance is attributed to its architectural sophistication, which facilitated superior performance relative to its counterparts.

Table 4: Precision and Recall

Model	Precision (%)	Recall (%)
GRU*	$66.0 \pm 2 \times 10^{-7}$	$66.3 \pm 6 \times 10^{-6}$
LSTM*	$67.7 \pm 1 \times 10^{-7}$	$68.8 \pm 4 \times 10^{-6}$
CNN*	$28.0 \pm 3 \times 10^{-6}$	$28.0 \pm 1 \times 10^{-6}$
FBCNet*	$39.6 \pm 2 \times 10^{-7}$	$40.6 \pm 4 \times 10^{-6}$
FBCCNN*	$68.0 \pm 3 \times 10^{-6}$	$26.0 \pm 1 \times 10^{-6}$
ViT*	$98.5 \pm 3 \times 10^{-6}$	$99.2 \pm 2 \times 10^{-6}$

Table 5: F1 Score and Time per Epoch

Model	F1 Score (%)	Time per Epoch (seconds)
GRU*	$65.9 \pm 2 \times 10^{-6}$	$13.303 \pm 1 \times 10^{-6}$
LSTM*	$68.0 \pm 1 \times 10^{-6}$	$13.370 \pm 9 \times 10^{-7}$
CNN*	$28.1 \pm 2 \times 10^{-6}$	$18.93 \pm 1 \times 10^{-6}$
FBCNet*	$38.6 \pm 3 \times 10^{-6}$	$13.8146 \pm 1 \times 10^{-6}$
FBCCNN*	$26.8 \pm 1 \times 10^{-6}$	$18.8471 \pm 7 \times 10^{-7}$
ViT*	$98.9 \pm 2 \times 10^{-6}$	$17.6572 \pm 1 \times 10^{-6}$

The integration of specific features, namely ’de_movingAve’, ’de_LDS’, ’psd_movingAve’, and ’psd_LDS’, across all rhythmic wave categories, was instrumental in enhancing model performance. These features, indicative of the underlying dynamics within the EEG signals, were pivotal in the models’ ability to discern and classify the data effectively.

It is imperative to acknowledge that the ViT model’s architectural superiority was a significant factor in its outperformance of other models we trained, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Gated Recurrent Units (GRUs). The latter models exhibited suboptimal accuracy, which can be ascribed to the absence of a grid search approach, a deficiency in hyperparameter optimization, and a limited training duration of merely 100 epochs.

When contrasted with referenced models, our models demonstrated enhanced accuracy. The CAN model recorded an accuracy of 87.71% with a standard deviation of 9.74, the SVM model 75.88% with a standard deviation of 16.14, the PR-PL model 85.56% with a standard deviation of 4.78, and the DCCA model 87.45% with a standard deviation of 9.23. Notwithstanding, the ViT model, as trained by our team, surpassed these referenced benchmarks, showcasing significantly elevated accuracy levels. This comparative analysis underscores the ViT model’s robustness and its potential as a leading framework for EEG signal classification.

The Vision Transformer (ViT) model, as depicted in Figure 4.16, underwent a rigorous training regimen spanning 100 epochs, yielding the following outcomes:

In the nascent stages of training, the model exhibited a gradual enhancement in performance metrics. Commencing with a training accuracy of 29.37% and a validation accuracy of 28.45% at epoch one, the model’s initial test accuracy was recorded at 29.06%. This incremental improvement was indicative of the model’s capacity to learn and adapt to the training data.

As the epochs advanced, a notable ascent in accuracy was observed. By the midpoint of the training epoch spectrum, specifically epoch 50, the model’s training accuracy had ascended to 69.12%, with the validation and test accuracies closely trailing at 67.21% and 66.76%, respectively.

The trajectory of accuracy continued its upward trend in the subsequent epochs. A marked milestone was achieved by the 86th epoch, where the training accuracy soared to 99.97%. Concurrently, the validation and test accuracies mirrored this upward trajectory, reaching an impressive 98.88% each. This consistent augmentation in accuracy is demonstrative of the model’s robust learning capabilities and its adeptness at generalizing from the training data to the validation and test datasets.

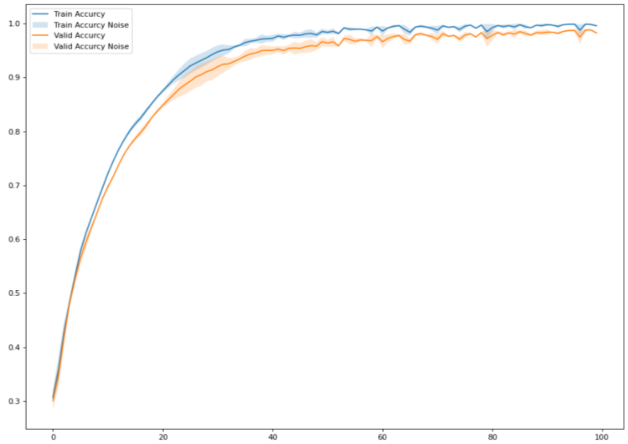

Figure 3: Train and Validation Accuracies of the ViT Model

Concomitant with the accuracy improvements, there was a steady decrement in loss values throughout the training duration. This decline is emblematic of the model’s increasing precision in aligning its predictions with the actual labels, thereby affirming its enhanced predictive fidelity.

In summation, the ViT model’s training journey was characterized by a steadfast amelioration in accuracy, culminating in exemplary performance on both validation and test datasets. The model’s adeptness at deciphering complex patterns from the image data and rendering accurate predictions is a testament to the efficacy of the Vision Transformer architecture, underscoring its potential as a robust tool for image-based data analysis.

The loss trajectory for the Vision Transformer (ViT) model delineates a consistent pattern of diminution, indicative of the model’s efficacious learning and its adeptness at adapting to the training corpus. Over the course of 100 epochs, the ViT model manifests a decrement in loss, culminating in a negligible final loss value in the vicinity of 0.0001. This metric corroborates the model’s proficiency in converging towards predictions of high veracity.

A noteworthy aspect of the ViT’s performance is the minimal variance in loss values observed across multiple training iterations. This uniformity is emblematic of the model’s reliable convergence to local optima, underscoring the stability and dependability of the solutions it renders.

In encapsulation, the loss curve analysis for the ViT model elucidates its exceptional capability to minimize the divergence between

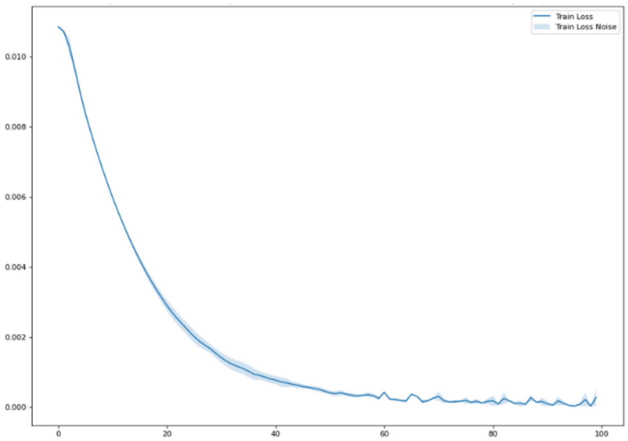

Figure 4: Train Loss of the ViT Model

its predictive outputs and the actual labels, thereby achieving a high degree of accuracy. The model's robust learning trajectory and its convergence to dependable solutions underscore its utility as an instrumental asset across a diverse array of applications.

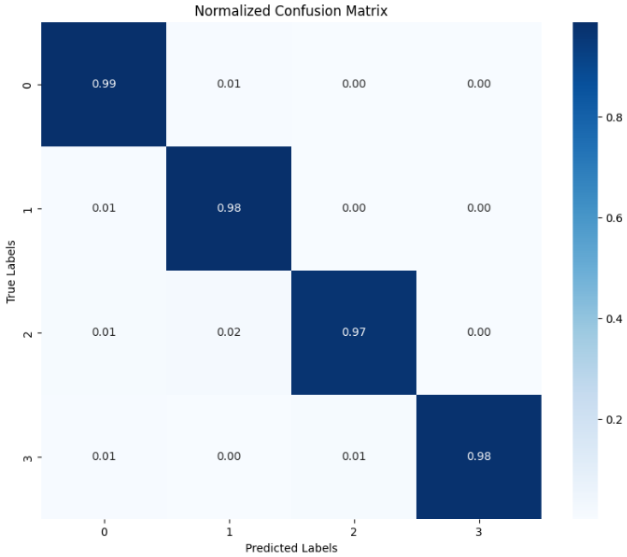

Figure 5: Confusion Matrix of the ViT Model

The confusion matrices provide valuable insights into the performance of classification models. Let's analyze the results for ViT model based on the given confusion matrice[Fig. 4].

– Class 0 (Happiness): The model performs exceptionally well in classifying happiness, with a high accuracy of 0.99. Only a small percentage (0.01) of the samples in this class are misclassified as other emotions.

– Class 1 (Sadness): Similarly, the ViT model achieves a high accuracy of 0.98 in identifying sadness. It misclassifies only a negligible percentage (0.02) of the samples in this class.

– Class 2 (Fear): The ViT model demonstrates strong performance in detecting fear, with an accuracy of 0.97. However, it misclassifies a small portion (0.01) of the samples in this category as other emotions.

– Class 3 (Neutral): The ViT model shows excellent accuracy in recognizing neutral emotions, with a value of 0.98. Only a small percentage (0.01) of the samples in this class are misclassified.

Overall, the ViT model exhibits impressive classification results, with high accura- cies across all emotion classes. It displays a strong ability to differentiate between different emotions, particularly in identifying happiness and neutral emotions.

5 COMPARATIVE STUDY

The Vision Transformer (ViT) model has exhibited significant superiority over several referenced models, including both conventional and contemporary techniques, in the field of EEG-based emotion identification. In [2], the performance of ViT surpasses that of the Correlation-Aware Network (CAN) in terms of accuracy. CAN achieves an accuracy of 87.71% ± 9.74, whereas ViT achieves a much higher accuracy of 99.02% with a notably smaller standard deviation. The improved precision and decreased variability observed in the performance of Vision Transformer (ViT) highlight its advanced capacity to effectively process the complex patterns inherent in electroencephalogram (EEG) signals, which is a critical aspect for achieving accurate emotion recognition.

The gains gained by ViT are further illustrated by comparing it with the classic Support Vector Machine (SVM) technique, as indicated in [18] . The support vector machine (SVM) achieves an accuracy of 75.88% ± 16.14. However, the performance of the Vision Transformer (ViT) surpasses that of SVM, as seen by its better accuracy and significantly reduced variance. This observation demonstrates the improved capacity of ViT to generalise and maintain performance across various data circumstances, which is a crucial characteristic for applications utilising EEG data.

The ViT model consistently outperforms the Probabilistic Representation and Pairwise Learning (PR-PL) model, as mentioned in [30]. The PR-PL model attains an accuracy of 85.56% ± 4.78. However, the ViT model exhibits near-perfect accuracy and minimum variance, indicating its advanced methodology in capturing both local and global dependencies within the EEG data. This aspect holds special significance in tasks related to emotion perception, as comprehending the intricate interaction among different signal components is crucial.

Moreover, when compared to Deep Canonical Correlation Analysis (DCCA) as stated in [9], ViT once again demonstrates its effectiveness. The performance of DCCA, as measured by its accuracy of 87.45% ± 9.23, does not meet the rigorous standard established by ViT. While DCCA demonstrates efficacy in examining intricate correlations within multimodal data, it seems to possess inferior capabilities compared to ViT in terms of both accuracy and consistency when applied to EEG-based emotion detection tasks.

In general, the aforementioned comparative observations position the Vision Transformer model as a leading contender in the field of emotion recognition utilising EEG data. The utilisation of sophisticated self-attention mechanisms enhances the ability to effectively acquire and analyse the intricate nature of EEG data, resulting in improved accuracy and reliability in the classification of emotions. The present comparative investigation not only underscores the potential of Vision Transformer (ViT) as a prominent approach in the discipline, but also symbolises a substantial progression over both conventional and contemporary approaches in the intricate domain of emotion recognition utilising electroencephalogram (EEG) inputs.

6 CONCLUSION

Our study represents a notable progression in the domain of emotion classification by the use of EEG signals. Through the utilisation of a varied array of deep learning models, with a specific focus on the Vision Transformer (ViT), we have successfully showcased the feasibility of attaining elevated levels of precision in the classification of emotional states. The ViT model has demonstrated remarkable performance, mostly because to its advanced self-attention processes, which have resulted in significant improvements over conventional models.

The efficacy of our methodology, which integrates Frequency-Domain characteristics and Entropy within deep learning architectures, has been demonstrated. The findings derived from the SEED IV dataset provide empirical support for the effectiveness of our models, particularly the ViT. This model not only demonstrated superior accuracy but also demonstrated consistent performance and dependability throughout numerous experimental iterations.

The achievement of the ViT model highlights the capacity of transformer-based systems in comprehending intricate, high-dimensional data, such as EEG signals. This study presents novel opportunities for further investigation and practical implementation, particularly in domains where comprehensive comprehension of human emotions has significant importance, such as the evaluation of mental well-being, neuro-marketing strategies, and the enhancement of human-computer interaction.

7 FUNDING

This paper has been partially supported by the Age-It: Ageing Well in an ageing society project, task 9.4.1 work package 4 spoke 9, within topic 8 extended partnership 8, under the National Recovery and Resilience Plan (PNRR), Mission 4 Component 2 Investment 1.3 - Call for tender No. 1557 of 11/10/2022 of Italian Ministry of University and Research funded by the European Union - NextGenerationEU, CUP B53C22004090006.

ACKNOWLEDGMENTS

This work has been developed at *is.Lab()* Intelligent Systems Laboratory at the Department of Computer, Control, and Management Engineering, Sapienza University of Rome (https://islab.diag.uniroma1.it).

REFERENCES

[1] Ali BOUSOUAR, Sallaheddine HARZALLAH, BACHIR Nail, and Imad Eddine TIBERMACINE. 2023. Comparative Study of Vibration Control using TMD on High Building response under Seismic Excitation. *Yantu Gongcheng Xuebao/Chinese Journal of Geotechnical Engineering* 45, 10 (2023), 9–19.

[2] Selma Büyükgöze. 2019. NON-INVASIVE BCI METHOD: EEG-ELECTROENCEPHALOGRAPHY. In *International Conference On Technics, Technologies And Education ICTTE.*

[3] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[5] TIBERMACINE IMAD EDDINE. [n. d.]. EEG Classification for Mind Controlling Applications using Multi-Method Approach. ([n. d.]).

[6] Maria Egger, Matthias Ley, and Sten Hanke. 2019. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science* 343 (2019), 35–55.

[7] Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution.* Springer, 162–190.

[8] Yuan-Pin Lin, Jyh-Horng Chen, Jeng-Ren Duann, Chin-Teng Lin, and Tzyy-Ping Jung. 2011. Generalizations of the subject-independent feature set for music-induced emotion recognition. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* IEEE, 6092–6095.

[9] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. 2019. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv preprint arXiv:1908.05349* (2019).

[10] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2016. Emotion recognition using multimodal deep learning. In *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23.* Springer, 521–529.

[11] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. 2015. Combining Eye Movements and EEG to Enhance Emotion Recognition.. In *IJCAI*, Vol. 15. Buenos Aires, 1170–1176.

[12] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[13] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11).* 689–696.

[14] Jorge Pérez, Javier Marinković, and Pablo Barceló. 2019. On the turing completeness of modern neural network architectures. *arXiv preprint arXiv:1901.03429* (2019).

[15] Rosalind W Picard. 2000. *Affective computing.* MIT press.

[16] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* 23, 10 (2001), 1175–1191.

[17] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion* 37 (2017), 98–125.

[18] Jie-Lin Qiu, Xiao-Yu Li, and Kai Hu. 2018. Correlated attention networks for multimodal emotion recognition. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM).* IEEE, 2656–2660.

[19] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2015. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* 7, 1 (2015), 17–28.

[20] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2011. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing* 3, 2 (2011), 211–223.

[21] Nazmi Sofian Suhaimi, James Mountstephens, Jason Teo, et al. 2020. EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational intelligence and neuroscience* 2020 (2020).

[22] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems* 28 (2015).

[23] Hao Tang, Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2017. Multimodal emotion recognition using deep neural networks. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24.* Springer, 811–819.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[25] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. 2014. Emotional state classification from EEG data using machine learning approach. *Neurocomputing* 129 (2014), 94–106.

[26] Yimin Yang, QM Jonathan Wu, Wei-Long Zheng, and Bao-Liang Lu. 2017. EEG-based emotion recognition using hierarchical network with subnetwork nodes. *IEEE Transactions on Cognitive and Developmental Systems* 10, 2 (2017), 408–419.

[27] Wei-Long Zheng, Bo-Nan Dong, and Bao-Liang Lu. 2014. Multimodal emotion recognition using EEG and eye tracking data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* IEEE, 5040–5043.

[28] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics* 49, 3 (2018), 1110–1122.

[29] Peixiang Zhong, Di Wang, and Chunyan Miao. 2020. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1290–1301.

[30] Rushuang Zhou, Zhiguo Zhang, Hong Fu, Li Zhang, Linling Li, Gan Huang, Yining Dong, Fali Li, Xin Yang, and Zhen Liang. 2022. PR-PL: A Novel Transfer Learning Framework with Prototypical Representation based Pairwise Learning for EEG-Based Emotion Recognition. *arXiv preprint arXiv:2202.06509* (2022).