Time-Series Transformer for EEG-Based Human Emotion Recognition

Shrushti Samant
STME
SVKM's NMIMS
Navi Mumbai, India
shrushti.samant@gmail.com

Shashank Jain
STME
SVKM's NMIMS
Navi Mumbai, India
shashankjain019@gmail.com

Shravani Devke
STME
SVKM's NMIMS
Navi Mumbai, India
devke22shravani@gmail.com

Sakshi Indolia
STME
SVKM's NMIMS
Navi Mumbai, India
sakshi.indolia@nmims.edu

Abstract—**Electroencephalogram (EEG) signal classification portrays a significant role in applications like Emotion Recognition, Cognitive Neuroscience and Brain Computer Interfaces (BCIs). Existing EEG-based emotion recognition models fails to completely capture spatial-temporal dependencies and also struggle to generalize across diverse datasets, limiting their real-world effectiveness. To overcome this, we propose deep learning frameworks like CNN, BILSTM and Transformers. This study has the objective of identifying the most efficient approach capable of effectively capturing spatial and temporal dependencies in EEG signals for accurately classifying human emotions as negative, neutral or positive. Performance of these are evaluated on a standard benchmark emotion recognition SEED dataset. The result shows that the time series transformer model, used along with a multi-head attention mechanism for capturing long-range dependencies outperforms CNN and BILSTM models, achieving an accuracy of 99.07%. These outcomes emphasize the potential of Transformer-based architecture in EEG based emotion classification.**

Keywords- ***Human Emotion Recognition, Facial Emotions, Speech, Physiological Signals, Electroencephalogram Signals (EEG), Audio, Video.***

I. Introduction

Human Emotion Recognition is an important area of research in the modern world due to its extensive range of applications in the field of mental health monitoring, adaptive learning systems and human-computer interactions. Multiple approaches have been explored till date to identify emotions, from facial expressions, voice and speech patterns, body gestures, and physiological signals. Physiological signals have gained the most attention among these because they can directly provide information about brain function, particularly the brain signals obtained by Electroencephalography (EEG).

Electroencephalography (EEG) is a neuroimaging technique that captures electrical activity from the brain using multiple electrodes placed on the scalp. Its high temporal resolution allows researchers to monitor and analyze brain activity in real time. Beyond emotion classification, EEG has been widely applied in, such as brain-computer interface (BCI) development, to provide direct interaction between the brain and external devices, like controlling of prosthetics and computer cursors through neural activity[1]. Additionally, EEG is employed in cognitive neuroscience to study attention, memory, and perception, providing insights into neural mechanisms underlying these processes [2]. Accurate emotion recognition using EEG signals can enhance systems aimed at improving adaptive learning, mental health assessment, and user-centered technologies.

Multiple machine learning methodologies have been explored in the past for extracting human emotions from EEG signals. Traditional approaches often involved combining classifiers such as Support Vector Machines (SVM) [3], Random Forests [4], and K-Nearest Neighbour (KNN) [5] with the manually created features. Several methods included both traditional machine learning and deep learning models for recognizing human emotions successfully [6]. CNN's have been consistent in successfully extracting the spatial features from EEG data, thereby helping to improve the classification performance [7]. To enhance the cross-subject and cross-session generalization multi-branch frameworks have been used in emotion recognition tasks[8].The latest researches have introduced a transformer-based model to capture global temporal relationship in EEG signals to enable better modelling of the long range dependencies [9]. Moreover, convolutional recurrent neural networks combined with multi-head self attention mechanisms have also been utilized to strengthen spatio-temporal feature extraction [10] . Ultimately, BILSTM models have also been employed successfully to capture the sequential dependencies within the EEG data [11].

To address the shortcomings of the traditional methods, deep learning models are developed with a powerful approach. These models are highly capable of automatically learning from complex spatial-temporal patterns from EEG data without the requirement for any extensive manual feature extraction. At the same time deep learning also has its own challenges like the need for large labeled datasets, higher computational requirements , and difficulty in obtaining strong generalization across the subjects due to individual differences in EEG Signals.

To resolve few of the gaps our experiment completely focuses on enhancing the accurate recognition of emotions via performance evaluation of three deep learning frameworks applied directly to raw EEG signals:

- A deep multilayered Convolutional Neural Network (CNN) to extract spatial features specific to individual brain regions from multichannel EEG data.

- A three-layer Bidirectional Long Short-Term Memory (BiLSTM) network to capture temporal dependencies in both forward and backward directions across EEG sequences.
- A Time-series Transformer model with multi-head attention can effectively describe long-range temporal dependencies and improve global feature extraction.

This study compares all three of these deep learning approaches thoroughly and investigates their capability to learn meaningful spatial and temporal patterns from EEG signals, enabling the recognition of human emotions across negative, neutral, and positive states.

The structure of this paper is as follows: Section II presents a comprehensive review of related work in EEG signal analysis and deep learning approaches. The dataset, pre-processing steps, and the architecture of the three proposed models are all covered in Section III. Section IV presents experimental results and a comparative evaluation. Finally, Section V concludes the paper with key findings and potential future directions.

II. Related Works

In this section, we review the related work in terms of EEG-based emotion recognition and attention mechanisms.

EEG-based emotion recognition has become a critical area in affective computing due to its ability to directly capture brain activity. Kavitha et al. [6] compared various emotion recognition approaches: facial expressions, voice, video, and other physiological signals. Their findings showed that EEG signals are more effective for emotion classification, achieving a precision of 96.87% when using deep convolutional neural networks on the DEAP dataset. They concluded that EEG-based techniques provide more accurate results for emotion recognition compared to other approaches.

A deep CNN model for emotion recognition was proposed by Chen et al., [11] which integrates frequency and temporal characteristics from EEG data. They used the DEAP dataset to illustrate that their model outperformed conventional classifiers like SVM and BT, achieving near-perfect classification using FREQNORM features. Their findings presented the potential of deep CNNs in capturing relevant patterns from EEG data for emotion classification.

In order to overcome cross-subject and cross-session issues in EEG-based emotion recognition, Han et al. [8] introduced the MBEER framework. The model combines spatiotemporal and frequency features using key modules like BG-DGAT and CFFE for feature extraction and MBC for classification. In cross-subject assessments, they achieved accuracies of 92.10% and 83.57% on the SEED and SEED-IV datasets, showing its effectiveness in maintaining high classification accuracy despite differences across subjects and sessions.

Sharma et al. [9] presented two hybrid models—CNN-BiLSTM and CNN-Transformer—for EEG classification applied to visual brain decoding. The CNN-BiLSTM achieved 71% accuracy on the EEG-ImageNet dataset, outperforming GRU Gated Transformers. This study highlighted how important it is to separate temporal and spatial information in EEG signals and showed how hybrid architectures can be used to decode complex visual stimuli.

Hu et al. [10] developed a hybrid CNN-BiLSTM-MHSA model, combining convolutional layers, BiLSTM, and multi-head self-attention for emotion recognition using EEG data. Using CSP filtering and continuous wavelet transforms, the model achieved 98.10% accuracy for binary classification and 89.33% for four-class emotion recognition on the DEAP dataset. The multi-head self-attention component enhanced feature weighting, resulting in improved classification performance.

The application of BiLSTM models for emotion recognition in conversational AI systems has been studied by Amadi et al. [7] By capturing long-term dependencies in textual data, the BiLSTM model improved emotion recognition accuracy to over 80% on a labeled dataset. This study demonstrated how attention to temporal dynamics enhances emotion-aware chatbot interactions and overall user satisfaction.

III. Methodology

A. Dataset

The pre-processed SEED (SJTU Emotion EEG Dataset) [12] contains 62-channel EEG signals from 15 subjects, each undergoing 3 experiments. In each experiment, subjects watched 15 emotion-evoking clips (positive, negative, or neutral). This setup results in 45 trials per subject, with 15 labeled segments per trial, totaling 675 labeled samples. This data was transformed to a different format to enable ease of use and reduce the time required to reload the data every time for execution. Each trial in the initial dataset consisted of multichannel EEG signals registered from 62 electrodes (channels) positioned on the head at a sampling frequency of about 200 Hz. Fast fourier transform is utilized to extract the significant frequency features and the band power was calculated across the five frequency brands present in the SEED dataset namely: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–100 Hz). This technique yields a complete feature representation capturing the spectral feature of brain activity, producing the shape of the data as (675,62,250,5) where 675 is the (15 x 3 X15) total number of trials, 62 are the number of EEG channels that is the number of electrodes connected to the scalp of the human during the trial, 250 is the average number of time steps per trial, and 5 are the frequency brands of SEED dataset. The correlative labels for each of the 15 trials of one particular label were originally in the form [-1,0,1] depicting negative, neutral and positive emotions were first altered using label encoding to [0,1,2] and then further reconstructed using categorical encoding resulting in output labels as [1,0,0], [0,1,0] and [0,0,1] for representing negative,positive and neutral respectively.Lastly both the features extracted as well as the transformed labels are stored in an HDF5 file format for easy loading and use in implementing deep learning models. The frequency band power data is normalized using standard scalers before splitting the dataset into training and testing sets.

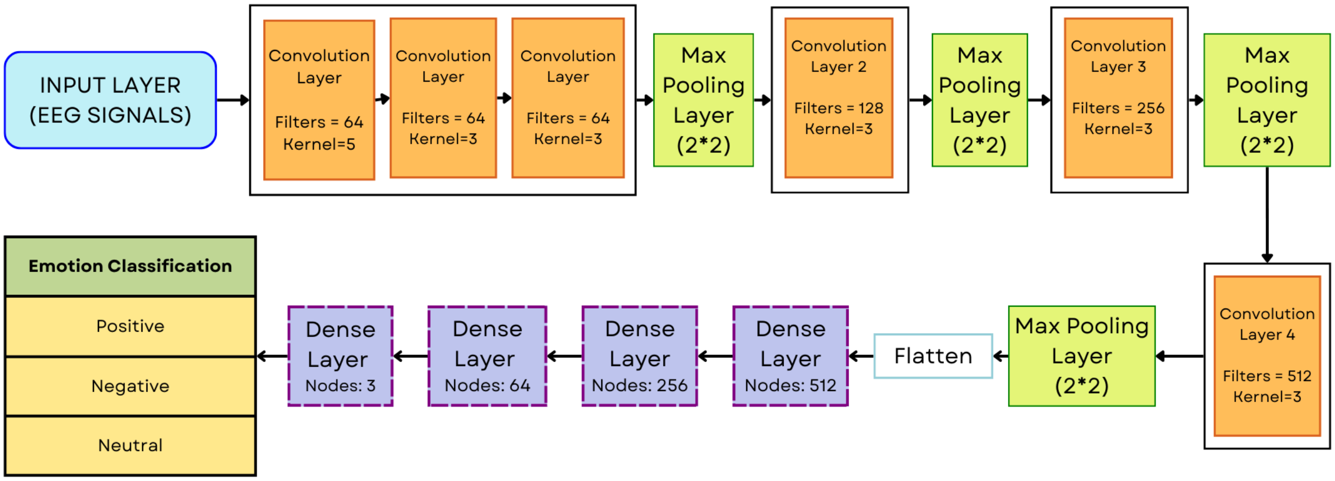
INPUT LAYER (EEG SIGNALS)
Convolution Layer Filters = 64 Kernel=5
Convolution Layer Filters = 64 Kernel=3
Convolution Layer Filters = 64 Kernel=3
Max Pooling Layer (2*2)
Convolution Layer 2 Filters = 128 Kernel=3
Max Pooling Layer (2*2)
Convolution Layer 3 Filters = 256 Kernel=3
Max Pooling Layer (2*2)
Convolution Layer 4 Filters = 512 Kernel=3
Max Pooling Layer (2*2)
Flatten
Dense Layer Nodes: 512
Dense Layer Nodes: 256
Dense Layer Nodes: 64
Dense Layer Nodes: 3
Emotion Classification
Positive
Negative
Neutral

Fig. 1: CNN Architecture

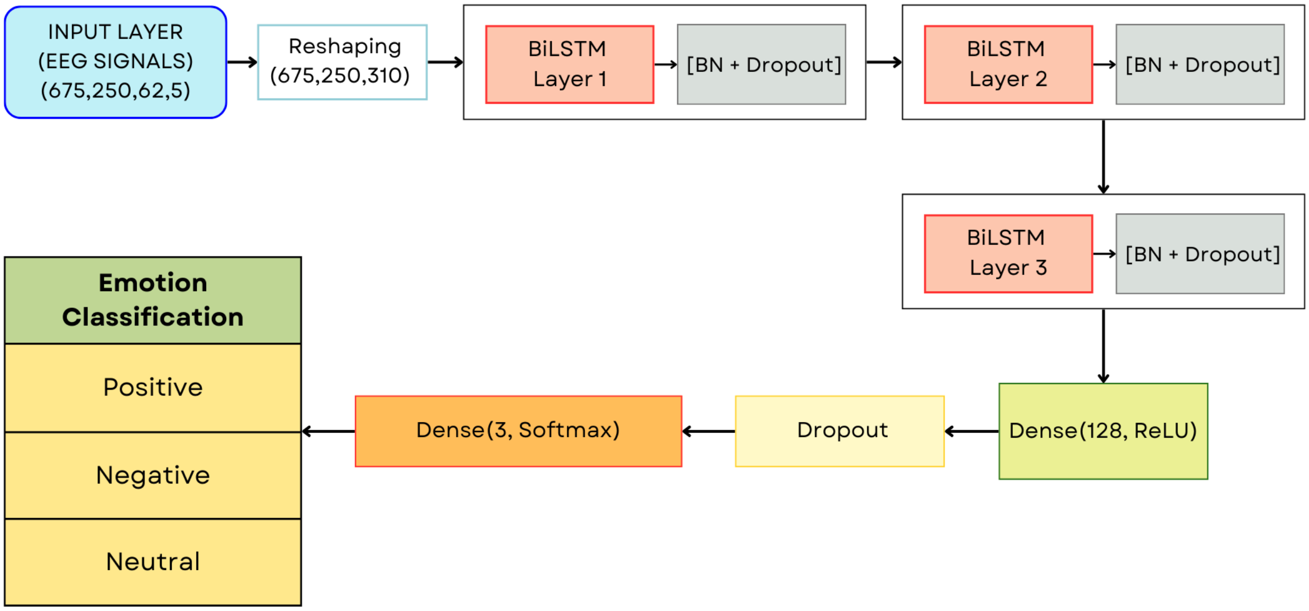

Fig. 2: Bidirectional Long Short-Term Memory

B. *Convolutional Neural Network (CNN)*

Convolutional neural network (CNN) is a framework of deep learning which extracts hierarchical spatial features from data using convolutional layers. In EEG analysis, CNNs are used to detect local spatial patterns over the EEG channels that have variations in voltage and frequency components. The pooling layers are responsible for reducing dimensionality while preserving critical information, thereby increasing the computational efficiency of the convolutional neural networks.

In the domain of emotion recognition from EEGs, CNNs automatically learn discriminative spatial features associated with emotional states, increasing the classifiers' performances.

A profound multilayered convolutional neural network (CNN) as shown in Fig. 1, is introduced for the multi-label classification of three output classes, namely negative, neutral, and positive. Our complex architecture comprises four convolutional blocks, each enhancing the complexity, with filters evolving from 64, 128, 256, 512 to gain information from both the high-level as well as low-level spatial features. Each block is equipped with multiple Conv2D layers, followed by Max Pooling 2D used to down sample the feature maps, preserving dominant features while reducing spatial dimensions and computational complexity and Dropout layers to prevent overfitting. After successful feature extraction, the complex network is then flattened to pass through the three fully connected layers (512, 256, 64 neurons) to allow the model learn from the deeper features of the data, followed

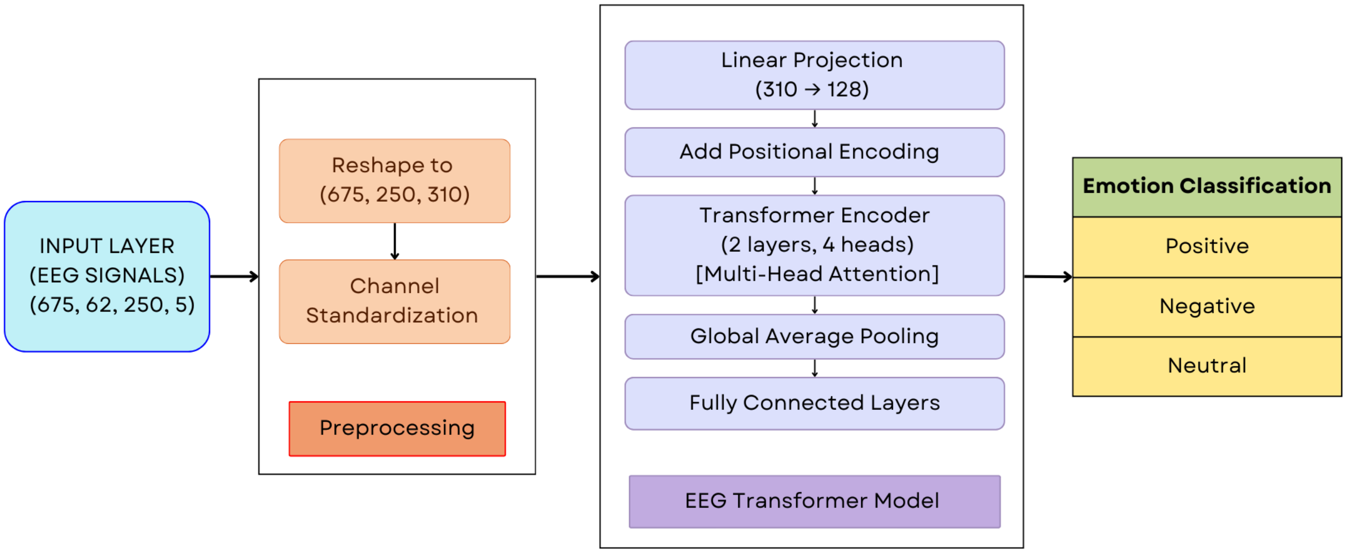

Fig. 3: Architecture Diagram for Time Series Transformer with multi-head attention for EEG

by a final softmax output layer with classification for three classes (negative, neutral and positive). Thus, our architecture is capable of balancing the feature extraction and classification, which ensures it is fit for use in complicated image-based tasks.

C. Bidirectional Long Short-Term Memory (BiLSTM)

A Bidirectional Long Short-Term Memory, is a particular kind of recurrent neural network for taking advantage of the forward and backward processing of data sequences allowing temporal dependencies to be extracted. BiLSTM differs from regular LSTM models because it can utilize both past and future contexts in order to capture richer temporal patterns.

Amadi et al. (2023) [7] emphasized the significance of a BiLSTM in sequential data handling and highlighted its use in emotion detection applications. Their findings demonstrated that BiLSTM considerably enhances emotion recognition performance in conversational systems by capturing bidirectional temporal information, as proven by their finding that BiLSTM achieved more than 80% accuracy in emotion classification.

As shown in Fig. 2, we have created a BILSTM model to capture the temporal dependencies intrinsic in our data. The implementation begins by reshaping each sample into 250 time steps with 310 features per step, permitting the model network to process sequential data efficaciously. The model network features three grouped BILSTM layers consisting of 256, 128, and 64 units respectively, where after every layer batch normalization and a dropout is used to improve upon the stability and reduce chances of overfitting. Before a final softmax layer which outputs the probabilities for three distinct labels (negative, positive and neutral), a fully connected dense layer is utilized to refine the extracted features. Thus, our model aims to provide a robust framework for EEG based multi-class classification.

D. Time Series Transformer with multi head attention for EEG

The self-attention mechanism allows deep learning models to dynamically allocate weights to various parts of an input sequence according to their connections. The model captures both local and global dependencies simultaneously through a self-attention mechanism, which is in contrast to recurrent networks that receive sequential inputs. This method is especially useful for EEG data since it can highlight important spatial-temporal patterns across multiple channels and time frames.

EEG data consists of signals that are registered synchronously from multiple electrodes (channels) placed on the scalp of a human being. These signals are recorded continuously, and thus inferred as time-series data, capturing brain activity at a very high resolution. Therefore, our efficient transformer model treats each time point as a sequence token and uses self-attention mechanisms to consider dependencies between time steps, enabling the model to prioritize important patterns and relationships.

The multi-head self-attention mechanism enables our transformer model to capture parallel information via multiple attention heads. Each head is capable of capturing a different characteristic of the temporal relationships, thus enhancing the learning of the model about temporal feature representation. This mechanism also helps our model learn how early and late the signals interact in a particular trial.

To enable the transformer-based sequence modeling, we reshaped the EEG data as shown in Fig. 3, in order to position time steps as sequence tokens. For all the time steps, EEG characteristics from all electrodes (channels) are concatenated, thus constructing a 310-dimensional feature vector per time step.

Secondly, a linear projection layer (embedding layer) is used to map the combined information obtained from all

TABLE I: The proposed methodology is compared alongside existing models for SEED dataset in terms of classification accuracy.

Reference	Stimulus	Classification Method	Accuracy (%)
Yun Su et al. [13]	SEED	TC-VIT	98.64
Weilong Tan et al. [14]	SEED	DAN	65.84
Zhijiang Wan et al. (2023) [15]	SEED	EEGformer	91.58
Peixiang Zhong et al. [16]	SEED	RGNN	85.30
Xingyi Wang et al. [17]	SEED	CNN	88.54
Cheng Cheng et al.[18]	SEED	MSDTT	97.52
Yiyuan Chen et al. [19]	SEED	Structural Deep Clustering	88.20
Zhe Wang et al. (2022) [20]	DEAP	CNN-LSTM	63.86
Utkarsh Dudeja et al. [21]	DEAP	GRU	98.90
Proposed Model	SEED	CNN	82.96
Proposed Model	SEED	BILSTM	85.00
Proposed Model	SEED	Time Series Transformer	92.59

channels and frequency bands at each time step of the EEG data. To establish compatibility with the desired input of the transformer model, all of the integrated features are projected into a 128-dimensional latent space. For capturing sequence awareness, which is not implicitly done by the transformer, positional encodings are appended to each time step's feature representation for the model to gain information about temporal ordering. To combine the information obtained about the temporal features from across sequences into a fixed-size vector, the global average pooling technique is used. The aggregated vector is relayed through a fully connected layer which outputs the class probabilities. Finally, the model is trained using Cross-Entropy Loss, and it is optimized using the Adam optimizer with a learning rate of 0.001. During training, the classification error is minimized over epochs.

IV. Result

This segment represents the evaluation of CNN, BILSTM and Time series based transformer models for EEG based emotion classification. The models are evaluated on the basis of the following performance metrics - Testing Accuracy (depicting how well the model predicts), Precision (representing the correctly identified positive cases from the predicted positive cases), Recall (depicting accurately classified positive cases out of all actual positive cases), F1-Score (represents the harmonic mean of precision and recall) and Computational Complexity (By measuring Giga Floating Point Operations to estimate the number of operations required for a single forward pass).

TABLE II: Performance Metrics for CNN, BILSTM, and Time Series Transformer

Performance Metrics	CNN	BILSTM	Time Series Transformer
Testing Accuracy (%)	82.96	85.00	92.59
Precision (%)	83.18	85.78	99.07
Recall (%)	82.96	84.44	99.07
F1-Score (%)	82.95	84.52	99.07
Computational Complexity (GFLOPs)	5.3382	4.4321	0.677

Additionally, the proposed methodology is compared alongside existing models for SEED dataset in terms of classification accuracy as illustrated in Table I.

As is clear from Table II, Time Series Transformer Model with multi-head attention has outperformed both CNN and BILSTM across all of the performance metrics. Even though both CNN and BILSTM managed to perform reasonably well, they were not able to achieve satisfactory testing accuracy. To address this issue and to enhance the capability of the model to capture spatial and temporal dependencies in EEG signals, we were motivated to introduce a Time-Series based Transformer model.

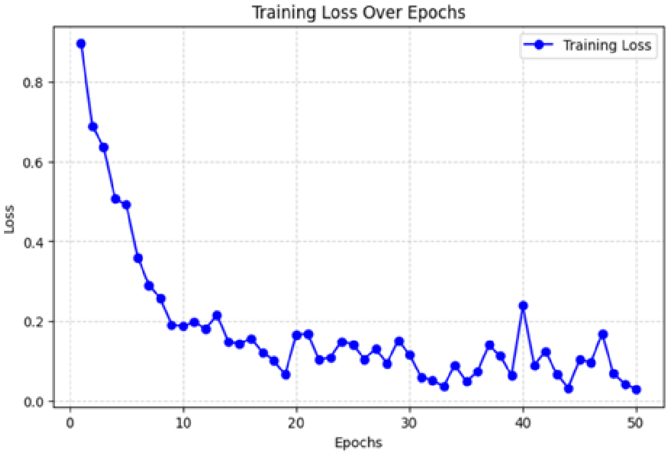

Fig. 4: Training Loss of Transformer Model

The proposed time series-based transformer model is hence capable of effective learning as shown in Fig. 4. Even after slight fluctuations, it manages to converge on completion of the epochs.

The confusion matrix shown in Fig. 5 shows the performance of our Transformer Model on a classification task with three classes (class 0 - negative, class 1 - neutral, class 2 - positive). Hence, our model is successfully capable of correct predictions as most predictions are aligning with their respective actual labels.

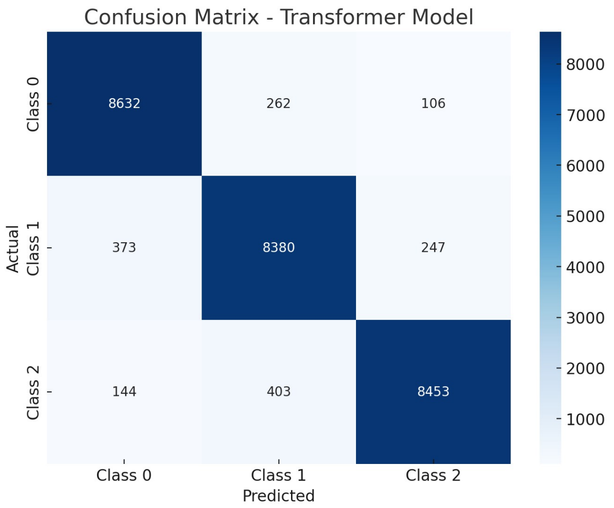

Fig. 5: Confusion Matrix for Transformer Model

V. Conclusion

The Time Series Transformer with multi-head attention model has excellent performance in EEG classification, as it is capable of achieving a Training Accuracy of 99.07%, Testing Accuracy of 92.59% and Precision, Recall, F1-Score of 99.07%, thereby outperforming the CNN and BILSTMs. Its multi-head attention mechanism effectively captures parallel information through multiple attention heads, making it a suitable fit for Brain-Computer Interfaces and Cognitive Monitoring. While the proposed approach effectively recognizes three emotional states using EEG signals, it is limited in scope and does not incorporate other modalities such as facial expressions or eye movements which could provide a more comprehensive understanding of human emotions. Future work will aim to extend emotion categories and explore multimodal integration to enhance the overall accuracy and robustness of the emotion recognition system.

References

[1] G. S., V. N., and S. Mahalakshmi, “Application of eeg signals – a case study,” *Advanced Aspects of Engineering Research Vol. 11*, p. 98–105, May 2021. [Online]. Available: https://stm.bookpi.org/AAER-V11/article/view/1255

[2] G. Winterer and R. W. McCarley, *Electrophysiology of Schizophrenia*. John Wiley Sons, Ltd, 2010, ch. 15, pp. 311–333. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444327298.ch15

[3] S. K. Jha, S. Suvvari, and M. Kumar, “Eeg-based emotion recognition: An in-depth analysis using deap and seed datasets,” in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2024, pp. 1816–1821.

[4] K. I, A. K, A. V. K, H. H, and B. Vidhya, “Eeg signal classification and analysis for upper limbs using svm and random forest algorithm,” in *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 2023, pp. 1508–1514.

[5] S. K, S. D, G. N M, L. S R, and S. K A K, “Emotion recognition using eeg signal classification of seed dataset,” in *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2023, pp. 1–6.

[6] K. K. V, S. L. R, and J. J. S, “A study on human emotion recognition techniques,” in *2023 International Conference on Innovations in Engineering and Technology (ICIET)*, 2023, pp. 1–6.

[7] C. Amadi, J. Odii, C. Ofoegbu, and C. Okpalla, “Emotion detection using a bidirectional long-short term memory (bilstm) neural network,” *International Journal of Current Pharmaceutical Review and Research*, vol. Vol 4, no 11, pp. 1718–1732, 11 2023.

[8] Q. Han, Y. Wei, Z. Pei, T. Weng, J. Qin, Y. Xu, S. Li, Y. Tian, Z. Li, and Y. Pei, “A multi-branch electroencephalogram emotion recognition framework based on cross-subject or cross-session multi-perspective representation fusion,” *SSRN Electronic Journal*, 2023.

[9] A. Sharma, J. Nigam, A. Rathore, and A. Bhavsar, “Eeg classification for visual brain decoding with spatio-temporal and transformer based paradigms.” in *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP 2024)*, ser. ICVGIP ’24. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3702250.3702286

[10] Z. Hu, L. Chen, Y. Luo, and J. Zhou, “Eeg-based emotion recognition using convolutional recurrent neural network with multi-head self-attention,” *Applied Sciences*, vol. 12, no. 21, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/21/11255

[11] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang, “Accurate eeg-based emotion recognition on combined features using deep convolutional neural networks,” *IEEE Access*, vol. 7, pp. 44 317–44 328, 2019.

[12] S. J. T. U. Brain-Computer Interface Laboratory, “Seed dataset,” https://bcmi.sjtu.edu.cn/home/seed/, 2015, accessed: 2025-04-11.

[13] Y. Su, Y. Zhou, X. Li, Q. Cai, and Y. Liu, “Eeg-based emotion recognition using temporal convolutional network and vision transformer,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[14] W. Tan, H. Zhang, Y. Wang, W. Wen, L. Chen, H. Li, X. Gao, and N. Zeng, “Seda-eeg: A semi-supervised emotion recognition network with domain adaptation for cross-subject eeg analysis,” *Neurocomputing*, vol. 622, p. 129315, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224020861

[15] Z. Wan, M. Li, S. Liu, J. Huang, H. Tan, and W. Duan, “Eegformer: A transformer–based brain activity classification method using eeg signal,” *Frontiers in Neuroscience*, vol. 17, 2023.

[16] P. Zhong, D. Wang, and C. Miao, “Eeg-based emotion recognition using regularized graph neural networks,” *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2022.

[17] X. Wang, Y. Ma, J. Cammon, F. Fang, Y. Gao, and Y. Zhang, “Self-supervised eeg emotion recognition models based on cnn,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1952–1962, 2023.

[18] C. Cheng, Y. Zhang, L. Liu, W. Liu, and L. Feng, “Multi-domain encoding of spatiotemporal dynamics in eeg for emotion recognition,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1342–1353, 2023.

[19] Y. Chen, X. Xu, and X. Qin, “Cross-subject and cross-session eeg emotion recognition based on multi-source structural deep clustering,” *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–15, 2025.

[20] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, “Transformers for eeg-based emotion recognition: A hierarchical spatial information learning model,” *IEEE Sensors Journal*, vol. 22, no. 5, pp. 4359–4368, 2022.

[21] U. Dudeja and S. K. Dubey, “Decoding emotions: Emotion classification from eeg brain signals using ai,” in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, vol. 10, 2023, pp. 1204–1208.