Temporal Attention Networks for Real-Time Multimodal Emotion Recognition from EEG and fNIRS Signals

Vinod Waiker[1], Anne Marie D. Pahiwon[2], Ankush Mehta[3], Gadug Sudhamsu[4],
Dr Pavithra M[5], Dr. K. Kiran Kumar[6], B Kiran Bala[7], Osama R.Shahin[8]
Datta Meghe Institute of Management Studies, Nagpur, Maharashtra, India 440020 [1]
Ifugao State University - Philippines[2]
Faculty of Engineering & Technology-Department of Mechanical Engineering-Marwadi University Research Center,
Marwadi University, Rajkot, 360003, Gujarat, India [3]
Department of Computer Science and Engineering-School of Engineering and Technology,
JAIN (Deemed to be University), Bangalore, Karnataka, India [4]
Associate Professor & Head-Department of Computer Science and Business Systems,
Jansons Institute of Technology Karumathampatti, Coimbatore, India [5]
Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India [6]
Department of AI & DS, K.Ramakrishnan College of Engineering, Trichy, India [7]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia [8]
Faculty of Engineering-Physics and Mathematics Department, Helwan University, Helwan, Egypt[8]

***Abstract*—Emotion recognition is critical in the development of real-time mental health care and individualized cognitive behavior. Current strategies to recognize cognitive emotions frequently fail to capture complex time dependencies and multimodal physiological reactions, leading to sub-optimal performance and inaccurate generalization. To overcome such shortcomings, the proposed study suggests TCADNet, a new deep learning model that integrates Temporal Convolutional Networks (TCN), attention-based feature weighting, and GAN-based data augmentation to achieve a high recognition rate of the emotional states through EEG and fNIRS recordings. The model utilizes the TCNs to extract both short-term and long-term temporal trends, and the attention mechanism emphasizes salient parts that bring about emotions, which improves interpretability. Moreover, a Deep Convolutional GAN creates artificial signals of unrepresented emotion classes, eliminating data imbalance and enhancing generalization. The TCADNet model is coded in Python on the TensorFlow/Keras system, and its key components are preprocessing, time modeling, attention weighting, data augmentation, and last classification by SoftMax layers. Experimental outcomes indicate that TCADNet has high recognition performance, with overall recognition, accuracy, precision, and recall, and F1-scores of over 98, which is higher than conventional CNN, LSTM, and separate TCN models. The suggested methodology can be useful to researchers, clinicians, and mental health professionals as it allows them to monitor cognitive and emotional conditions in real-time with a reliable, decipherable, and scalable instrument and provides an opportunity to detect and respond to the issue promptly and implement a tailored intervention plan in educational or health-related settings.**

Keywords*—*Emotion recognition; Temporal Convolutional Network; attention mechanism; mental healthcare; EEG

I. INTRODUCTION

The rising cases of mental health problems have led to the invention of intelligent, and autonomous systems that are able to identify emotional states with high precision and confidence. According to the estimates done by the world health organization (2024), over 280 million individuals around the world are affected by depression and stress-related disease and emotional dysregulation can frequently precede clinical manifestations [1]. Electroencephalography (EEG) is a non-invasive neurophysiological method that obtains a record of electrical activity in the brain by electrodes on the scalp and the frequency content is measured in microvolts (uV) at frequencies between 1100 Hz [2]. EEG offers a very high temporal resolution of milliseconds, which can accurately trace fast neural changes related to emotional and thinking processes. Conversely, functional near-infrared spectroscopy (fNIRS) is an optical neuroimaging modality technique that monitors effects of hemodynamic changes by measuring alterations of oxygenated (HbO) and deoxygenated hemoglobin (HbR) in the bloodstream by using near-infrared light sources with sampling rates around 7-10 Hz. Although EEG records direct electrical activity of neurons with a high level of temporal resolution, fNIRS is more effective in providing a localization of cortical blood flow in terms of the vascular response [3],[4]. The combination of the fast temporal sensitivity of EEG and the spatial specificity of fNIRS can be used to further describe emotion-related brain activity in a more comprehensive manner.

The trends in artificial intelligence (AI) recently have demonstrated that, when given raw physiological signals, deep learning models can learn discriminative features. Convolutional Neural Networks (CNNs) and Long Short-term

Memory (LSTM) networks have demonstrated good results in terms of identifying emotional patterns given time-series data [5]. Nevertheless, these architectures have some serious drawbacks, including extreme sensitivity to noise [6], inability to deal with long distance temporal dependencies and uninterpretability [7]. The cognitive frameworks that make up multimodal interaction are stipulated based on the occurrence of a particular modality-verbal, bodily, and visual-is mode of interaction, the presence of a grammatical structure-syntax, and narrative as well as whether they are predominant throughout the whole semantics of the expression [8]. Multimodal sentiment analysis is the most interesting study area of AI across the disciplines. The current classification task of human multimodal sentiment analysis is a highly finer task compared to the previous classification tasks. MSA has made significant contributions to human affection polarity detection over the past years since the advent of DL and the benefit of neural networks [9]. Multimodal DL consists of a number of challenges. One of the critical issues is the conversion of the information in multimodal form to the machine readable one [10]. AI solutions have a high potential of enhancing mental health. However, which approach should be selected is an issue of main concern [11]. Depression and stress are highly prevalent mental disorders that permeate the whole globe to the extent of reaching everyone in the society regardless of their age, gender and social economic statuses [12].

With such gaps in mind, the emerging concern of Temporal Convolutional Networks (TCNs), attention-based mechanisms and generative adversarial networks (GANs) has been in the past due to its capability in capturing and addressing temporal dependence, improving interpretability and reducing data imbalance, respectively. However, the existing methods seldom use all three mechanisms into a single model of cognitive emotion recognition basing on multimodal cues such as EEG and fNIRS [13]. In this case, assessment of any unfavorable psychological characteristic such as stress, anxiety and sadness all hinge on the detection of emotion. The second source of knowledge is the application of a new combination of an electroencephalogram (EEG) data [14]. The smart use of technologies has transformed the healthcare industry in a monumental manner. The focus of all the research is on the following critical aspects: medical imaging and diagnosis, virtual patient treatment, medical research and drug discovery, patient engagement, compliance, rehabilitation, and other administrative applications. It is based on the general picture that illustrates the application of AI in the area of medicine [15]. The advent of the artificial intelligence technologies has enhanced the efficiency and reasonability of the data collection process that has tremendous influence on the field of illness management and prevention [16]. The rising rate of mental diseases among the youths globally is one of the biggest problems in this society [17]. The issues of mental diseases are extremely high in the youth population; according to the existing estimates, 20 per cent of children and adolescents in the world are mentally ill [18]. The mental health care services is also needed today due to the social change, rising demand of optimal healthcare and flexibility of the mental illness. The conventional channels of accessing mental health services are pushed to the brink by the lack of resources and long queues in the process of accessing the services [19]. Sentiment analysis and emotion detection are vital in emotive management of patients in the healthcare [20].

A. Research Motivation

The growing rates of cognitive stress, anxiety, and dysregulation of emotions in students demonstrate the urgency of objective and real-time emotion recognition tools. Conventional methods of assessment mostly use self-administered questionnaires and behavioral observations, which are subjective and cannot detect fast neurophysiological variation. Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS), on the contrary, offer a complementary neural measure in electrical brain activity and hemodynamic responses, respectively. Nevertheless, such multimodal time-series signals are difficult to model, as they are noise sensitive, their temporal misalignment, and multi-complex long-range correlations. Traditional deep learning algorithms like CNNs and LSTMs tend to violate the causality of the time, interpretability, and address data imbalances in affective data. These constraints inspire this study to present a unified deep learning model consisting of temporal convolutional network, attention, and generative adversarial augmentation to achieve robust, interpretive, and scalable multimodal cognitive emotion recognition of student mental health in monitoring.

B. Research Significance

This study makes contributions to multimodal cognitive emotion recognition as it provides a single deep learning framework specific to EEG-fNIRS signal fusion. The model is used to learn temporal dependencies on long distances by exploiting temporal convolutional networks (TCNs) without violating causal order and permitting real-time use. By using a temporal attention system, interpretability is improved by detecting time segments of emotional salience in neural recording [21], [7]. Moreover, GAN-based augmentation decreases the class imbalance and enhances the performance of generalization in emotion datasets. The proposed framework takes advantage of the complementary nature of the high temporal resolution of EEG and the spatial resolution of fNIRS, unlike the more conventional single-modality methods, which lead to a higher strength and discrimination ability. The system facilitates scaled real-time use which will be useful in the mental health assessment, academic stress monitoring and adaptive learning settings. In general, the structure presents a non-invasive, objective, and data-driven alternative to traditional psychological assessment tools, allowing to detect and proactively manage the emotional-related cognitive conditions early.

The contribution of the research are as follows:

- Developed a Temporal Convolutional Attention-Driven Network (TCADNet) for multimodal cognitive emotion recognition, capable of modeling complex temporal dependencies in synchronized EEG and fNIRS neurophysiological signals.

- Employed a Temporal Convolutional Network with causal and dilated convolutions integrated with an attention mechanism to identify emotionally significant time segments and improve interpretability.

- Utilized the publicly available REFED multimodal EEG–fNIRS dataset to validate the effectiveness of the proposed framework under realistic emotion elicitation conditions, including synchronized signal preprocessing and temporal window segmentation.
- Achieved high classification accuracy and improved generalization performance compared to conventional deep learning approaches, demonstrating the robustness and scalability of the proposed framework.

C. Problem Statement

Neurophysiological signals are complex, non-linear, and time-varying, which makes it a challenging undertaking to identify cognitive emotions pertaining to stress accurately. Multimodal signals, including electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS), can give complementary data regarding the neural electrical activity and hemodynamic responses [22],[23], but they are not easily described in terms of their temporal dynamics or cross-modal interactions. The presence of long-term temporal dependencies and variations in the context of emotion states in conventional machine learning models has led to the frequent reliance on hand-crafted features in traditional analytical methods and machine learning models [18]. Recurrent neural networks and conventional convolutional architectures that have been used in emotion recognition do have weaknesses, including the disappearance of gradients, excessively high computational expense, low levels of interpretability, and the inability to conduct multimodal synchronization. Moreover, in many cases, emotion datasets are also unbalanced in classes, and this results in biased learning and lower generalization scores. The absence of mechanisms to detect emotionally salient temporal pieces also limits the reliability of the models [21]. The constraints prevent the creation of effective, real-time multimodal emotion recognition in student mental health monitoring. Consequently, there exists an acute necessity of scalable and understandable deep learning structure that can imitate multi-scale time-related characteristics, deal with the issue of class imbalance, and utilize EEG-fNIRS data in a manner that is effective in providing cognition emotion recognition.

The study is organized as follows: Section I presents the motivation, challenges, and goals of AI-based multimodal emotion recognition. Section II overviews related research. Section III explains the proposed self-supervised temporal learning framework, multimodal data fusion, and experiment setup. Section IV shows results and analysis. Section V concludes and summarizes future work.

II. Related Works

Kargarandehkordi et al. [21] proposed to critically evaluate the use of AI models for forecasting, tracking mental health states and symptoms using wearable biosensor data. Following PRISMA criteria, the authors conducted a systematic analysis of 48 studies using a range of sensors, including location, accelerometry, heart rate, heart rate variability, electrodermal activity, audio, and smartphone use data. These sensors provide continuous, remote, and objective mental state monitoring and are integrated into wearable technology like fitness bands and smartwatches. The integration of AI, especially ML, makes it possible to detect mild physiological, behavioral signals and signs related to mental health.

Dritsas et al. [7] proposed recent developments in methods and adaptive system architecture for multimodal systems. Though specific datasets do not directly find use in the study, systems by modalities that usually utilize sensor-input datasets like motion capture, speech recordings, eye-tracking, and physiological signals are also analyzed in the study. Enhanced usability, natural interaction, context-sensitive flexibility, and portability across a wide range of applications, such as healthcare, are some of the most significant advantages. Difficulties, however, are in terms of complex input synchronizing, conflict of fusion, environmental flexibility and lack of standardized structures or frames of reference to actual deployments.

Avital et al. [2] proposed to create a real-time system for assessing lecturers’ teaching quality by observing students' facial expressions to identify emotional states while listening to lectures. The procedure consists of four phases. They are: image capture, preprocessing, emotion recognition and identification. The data was obtained experimentally from a class of 45 students recording live facial expressions during lectures. The major strengths are instant feedback regarding student engagement, high accuracy (83%) in emotion recognition and high correlation (91.7%) with conventional feedback methods. Further, the suggested application of optical neural networks suggests quicker and more efficient processing than electronic systems. However, some limitations are moderate prediction accuracy of student comprehension (43–62.94%), possible privacy issues, reliance on steady lighting and camera position for proper facial detection. Nassiri and Akhloufi [3] suggest the use of large language models like GPT, Bloom and LLaMA in the healthcare industry for enhancing patient care, medical research and diagnostic assistance. The objective is to evaluate the ability of LLMs to read, create medical text and approximate their usability in actual clinical practice. The approaches encompass a thorough review and examination of the architectures of top LLMs, their training, and their performance on medical tasks. The study investigates datasets for training these models, which are classified by size, source and medical topics. Among the strengths, LLMs have tremendous potential to aid healthcare professionals in documentation, literature review and decision-making, as well as potentially enhance the efficiency of the healthcare system through automation.

X. Chen et al. [22] proposed to investigate the evolution and trend of AI-based multimodal data analysis in smart healthcare using large-scale literature research. In order to make this a reality, authors used bibliometric analysis and topic modeling techniques in order to explore 683 scholarly articles across 2002 and 2022. The data of analysis were procured from research databases with specific focus on AI, healthcare, and multimodal integration of data. The stark strength of the work lies in its capability of uncovering avenues of research that will unfold, leading individuals and topics to follow such as GANs, contrastive learning and hybrid neuroimaging approaches. It has worldwide emphasis with special regard being provided to contributions from nations such as the USA and India. While, there are certain limitations like possible bias because of data selection (language or database constraint) or grey literature exclusion and article metadata use instead of experimental or clinical data. Despite this, the study provides valuable points for future interdisciplinary research in AI-based smart healthcare.

Lee et al. [23] has been suggested to study the integration of Artificial Intelligence into the field of biomedical signal analysis to enhance the effectiveness and accuracy of diagnosis in healthcare. It dwells upon the most prominent AI approaches that are controlled, uncontrolled, reinforcement learning and how it can be applied to interpret bio signals like EEG and ECG. The methods involve the analysis of various AI architectures and learning paradigms applied in the classification of physiological signals, action and abnormalities, and decision-making activities. Despite this, the review lacks dataset specificity. It is founded on actual biomedical signals, such as ECG and EEG, in different clinical settings. Among the key advantages, one can distinguish the increased speed and accuracy of the diagnostic process, reducing the amount of work that is to be carried out manually and increased access to health services in resource-deficient areas and better response to emergencies through the rapid analysis. Yet, constraints include data privacy issues, interpretability of the model, dependence on high-quality labeled datasets and real-time integration difficulties in clinical settings. Payandeh et al. [4] proposed an overview of deep representation learning methods, pointing out their application to learn meaningful, low-dimensional features from high-dimensional data for tasks like classification, prediction and segmentation. While the review does not concentrate on particular datasets, it mentions varied types of data such as images, text, sequential data and multimodal inputs. The benefits of such methods are to learn generalizable and transferable features automatically without handcrafted feature engineering, thus enhancing learning robustness and model generalizability. Although much headway has been achieved, limitations persist in the ability to learn interpretable representations, having intensive computation, managing scarce or biased data, as well as deriving strong robustness across diverse domains or conditions.

III. Proposed TCADNet Framework for Cognitive Emotion Recognition

The proposed study proposes a Temporal Convolutional Attention-Driven Network (TCADNet) to learn to detect robust cognitive emotions through synchronized multimodal EEG fNIRS physiological signals. The framework is used to describe a multi-stage workflow that is structured to efficiently model the complex time dynamics of neural processes and enhance their classification performance. The raw EEG and fNIRS signals are first preprocessed (removal of noise, band-pass filtering, artifact correction, normalization and temporal segregating) to guarantee the quality and reliability of the signals. The multimodal signals are then preprocessed and synchronized, followed by feature-level fusing and inputting in a Temporal Convolutional Network (TCN) to extract hierarchical temporal features. The TCN uses causal and dilated one-dimensional convolutions with residual links that allow the model to learn both short-term neural variations along with the long-range temporal correlations of cognitive-emotional reactions.

Fig. 1 explains how the current embodiment is structured, incorporating a number of EEG and fNIRS signals to determine the locations of the emotional states of discrete cognitive emotion categories. In order to achieve better interpretability and discriminative power, there is a temporal attention mechanism that is incorporated after the TCN layers. This process gives various time segments adaptive weights that enable the network to prioritize emotionally salient neural patterns. To address class imbalance and improve generalization, a one-dimensional Generative Adversarial Network (1D-GAN) is incorporated to synthesize realistic minority-class EEG–fNIRS time-series samples. The attention-enhanced representations are then passed through fully connected layers, and a SoftMax classifier predicts discrete cognitive emotion states. The proposed TCADNet framework demonstrates improved accuracy, robustness, and scalability, making it suitable for real-time multimodal emotion recognition and AI-assisted mental health monitoring applications.

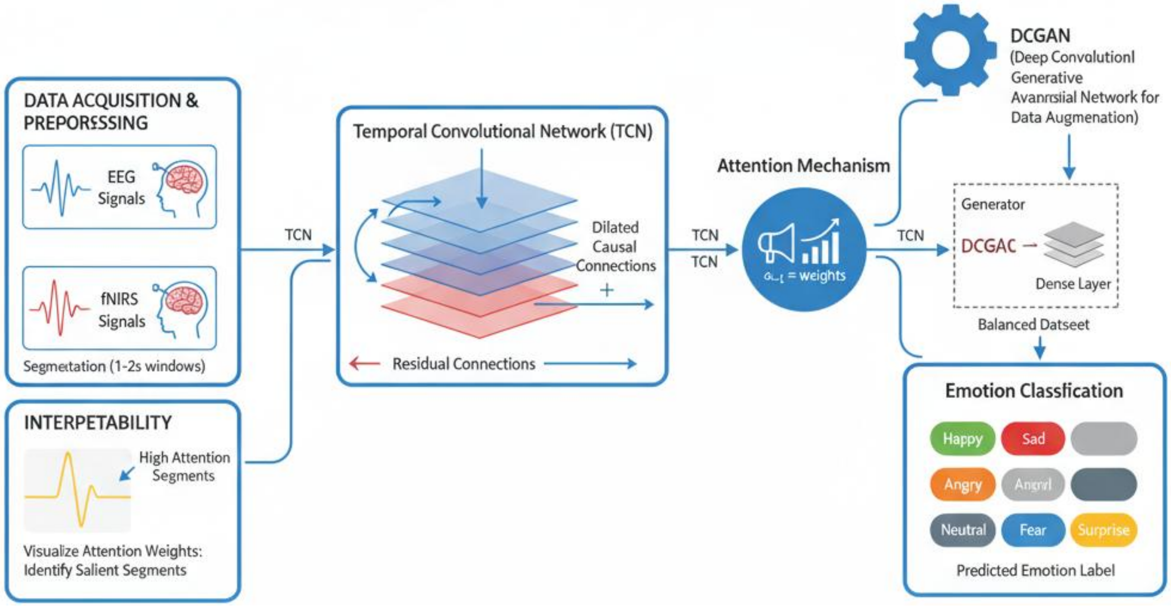

Fig. 1. Overall architecture.

A. Data Collection

The study utilizes the REFED dataset (Real-time EEG–fNIRS Emotion Dataset) [24], which provides synchronized multimodal neurophysiological recordings for cognitive emotion recognition. The data consists of the simultaneous electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) signals recorded when the subjects are presented with controlled emotional stimuli.

EEGs represent fast neural electric activation with high temporal resolution, whereas fNIRS detects slower hemodynamic changes related to cortical activity. These modalities are complementary, which allows a more detailed study of the processes of emotional processing.

The NIRx Medical Technologies NIRSport2 system was used to record brain hemodynamic activity in this study, having a sampling rate of about 10 Hz. It employs two wavelengths (760nm and 850nm) to determine variations in oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR). The raw optical density measurements were converted to changes in hemoglobin concentration by the Modified Beer Lambert Law.

EEG recordings were made simultaneously to obtain fine-grained neural dynamics by using a multi-channel acquisition system, and with a higher sampling frequency. To provide the temporal synchrony of fNIRS and EEG modalities, event markers were added when presenting the stimulus.

Multimodal recordings were split into temporal windows of constant length relating to the emotion stimulus. These synchronized EEG-fNIRS segments are the input to the suggested Temporal Convolutional Attention-Driven Network (TCADNet) in cognitive emotion classification.

B. Data Preprocessing

Preprocessing is supposed to enhance the quality of signals and take the EEG and fNIRS data to a deep learning model level. Physiological recording is normally corrupted with motion artifact, baseline drift and noise. In order to remove such noise, NIRSlab is used in this study without a significant loss of time and amplitude features. The signals are de-trended and de-filtered to eliminate the distortions in the high or low frequencies, which could affect the feature extraction. The lost or corrupted data points are interpolated to prevent loss of information. Bandpass filtering was utilized to eliminate unwanted frequencies, while normalization approaches, such as min-max scaling, were employed to standardize the input. Signals were segmented into fixed-length time windows to ensure temporal continuity. Next, missing data points were filled in using mean value interpolation to avoid data loss. This stage ensures only clean, standardized signals are input into the deep learning model. Signal processing is the foundation of successful temporal modeling.

1) Normalization: Scale features to a standard range (e.g., 0–1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

In Eq. (1), original feature value is X, X_{min}= the feature's lowest value and X_{max} is the highest possible value of the trait. X' is the value of the normalized feature.

2) Missing values: Can be handled using mean imputation:

$$x_i = \frac{\sum_{j=1}^{n} x_j}{n} \tag{2}$$

In Eq. (2), x_i— to infer a missing value, x_j is the feature's observed values and n—number of values observed.

C. Multimodal Temporal Synchronization

EEG and fNIRS signals have different sampling rates and record different complementary neural activity. EEG has a high temporal resolution as it reveals rapid electrical activity, and fNIRS has slower responses to hemodynamic processes, which are related to cortical activity. In order to facilitate successful multimodal fusion, both modalities have to be synchronized in time.

Let EEG and fNIRS represent the EEG and fNIRS sampling rates, respectively. EEG was usually higher in sampling frequency and therefore downsampled to achieve an effective temporal resolution of fNIRS. As an alternative, fNIRS signals have been linearly interpolated to match EEG samples in time.

Both modalities were divided into fixed-length sliding windows of length T seconds with an overlap of 50%. The temporal alignment of each window was provided in accordance with the synchronized event markers in order to make the EEG and fNIRS fragments match the identical interval of the emotional stimulus. The temporal fusion was then carried out by concatenating the synchronized segments along the feature dimension to carry out feature-level fusion before temporal modeling. This synchronization strategy provides temporal coherence and complementary neural information to make multimodal emotion recognition robust.

D. Feature Extraction

The feature extraction process transforms signals that have been preprocessed into an individual meaningful representation to recall emotional information. In that work, the features are drawn in the form of a TCN as the input layers of the network, which itself pulls out the patterns on the basis of the raw EEG and fNIRS sequences. Convolutional filters detect local variations in time that are an indication of emotional states. The hierarchical convolutional net allows the net to learn increasingly more abstracted versions, not only of the fine-grained signal structure but also of the high-level temporal structure. The other links ensure continuity of information, where information is not lost in the process of layer stacking. This prevents hand-engineered features, which are likely to be biased or incomplete. Extracted features reduce the dimensions and, at the same time, retain discriminative power and can be utilized in downstream time modeling. Learning of every emotion category of the network is included in the network adapt, which is more sensitive to small differences. A proper day-by-day modeling step is ensured by good feature extraction. This stage captures both cross-modal and modality-specific patterns and correlations. The features that result are discriminative and temporal ready as well as strong. Learning of emotions in a more generalizable and precise way entails feature extraction.

$$f_t^l = \sigma \sum_{k=0}^{K-1} W_k^l . x_{t-d.k}^{l-1} + b^l \tag{3}$$

In Eq. (3), the value of f_t^l at layer l is calculated as the result of a temporal convolution operation with a kernel size K and dilation factor d. The weights W_k^l and bias b^l are applied to the output of the last layer $x_{t-d.k}^{l-1}$ and combined by an activation function σ to ensure that short-term and long-term temporal interactions are well represented.

E. TCADNet Framework

TCADNet utilizes Temporal Convolutional Networks to accomplish the concept of temporal modeling. In contrast to recurrent neural networks, TCNs use causal convolutions to make sure that the future predictions at any moment are made with the help of the current and past data only and, therefore, preserve the order of the signal. Moreover, it employs dilated convolutions to expand the receptive field of the model without expanding its depth, and in such a manner, the model can be capable of capturing various emotional variations in the short term, not to mention being able to capture long-term correlations. These are the various methods employed by the TCN to vouch out discriminative temporal characteristics of the EEG signal and fNIRS signal.

1) Temporal modeling (Using Temporal Convolutional Networks – TCN): The sequential correlation of EEG and fNIRS is the temporal modeling. The TCNs are used, so as to take care of the long-term time trends and short term. Causal convolutions have the advantage that they do not cause the sequence of the signals across time to be lost, such that future predictions do not leak their knowledge of the past. The expanded convolutions increase the receptive field and this enables the network to acquire the long-range dependencies without necessarily overgoing too deep. Remnant links enhance the gradient flow by stabilizing deeper network training. Meanwhile, TCNs have the ability to train; therefore, training is not as time-consuming as recurrent networks such as LSTMs. The network is informed on the dynamism in the emotional states and it separates out the minor time trends among the participants. Temporal modeling also captures features by taking instantaneous and dynamic signal attributes. The time dynamics have been well captured in TCN output and information is stored at varying scales. The step plays an important role in capturing of the development of emotions. The input of the attention mechanism is learned time features. Another feature of the model is the use of effective awareness of emotions in real-time as a result of the effective time series modeling. In Eq. (4), TCN has been expounded with regard to a sequence.

$$F = TCN(X) \tag{4}$$

2) Attention-based feature weighting: To improve the interpretability and performance of recognition, we apply an attention mechanism to the output feature representations of layers of TCN. The attention module will give a weight to each temporal segment in view of their emotional patterns and, therefore, weight on a segment that has a stronger emotional pattern will be made greater. This cautious weighting gives the model the opportunity to identify and assign greater weight to the more informative components of the signal and underweight the noise and other unimportant components of the signal. The weighted feature attention thereof provides a more discriminate and finer representation of the emotional condition. During the selection of the emotionally informative cuts of the TCN output, the process of attention is combined. Each of the temporal features is given a weight, which is the contribution of the feature in classifying emotions. It is the active change of the process with regard to every input sequence to introduce pressure of the close approach of the emotional pattern of the participants. The weighted features reduce the impact of minor variations on the generalization more. It is another step of the temporal modeling since it prioritizes important information and does not overlook the rest of the sequence. The focus provides it with a fairly definite form. It also makes the classification more accurate as it minimizes the confusion of the closely related classes of emotions. It is so because the time center dependencies and feature relevance are also considered by the model, as TCN is also accompanied with attention. The emphasis based on the consideration of the attention will ensure the sentiments are projected appropriately in a forceful and comprehensible manner.

$$F_{att} = \sum_{t=1}^{T} \propto_t F_t \tag{5}$$

In Eq. (5), the attention module, F_t, is the TCN feature at time t, and $\propto_t$ refers to the time t attention weight of that feature. Final weighted feature vector F_{att} is calculated as a sum of TCN features multiplied by their attention weights that underline the most informative temporal segments of emotion recognition.

3) Data augmentation and class imbalance handling: To manage class imbalance during emotion recognition, a Deep Convolutional Generative Adversarial Network (DCGAN) is used to produce synthetic EEG and fNIRS signal sequences of the tiny emotion classes. The generator network is trained to generate natural time-series segments that can reproduce the visual appearances of actual physiological signals in terms of time. These generated sequences are checked by the discriminator network and distinguished between the real EEG/fNIRS signals and the generated scores. Both networks are adversarially trained: the generator should be able to make the discriminator use them, whereas the discriminator should be better at recognizing synthetic sequences.

The produced synthetic cues retain both temporal and spectral features that are important in the recognition of emotions. The synthetic sequences are merged with the original dataset to form a balanced training set, which enhances generalization and minimizes bias in the majority of types of emotions. Incorporation of the data produced by GAN makes sure that TCADNet acquires a strong feature in all kinds of emotions and improves the level of accuracy and stability in real-time student stress prediction tasks, as in Eq. (6):

$$D_f = \log\left(1 - D(G(z))\right) \quad (6)$$

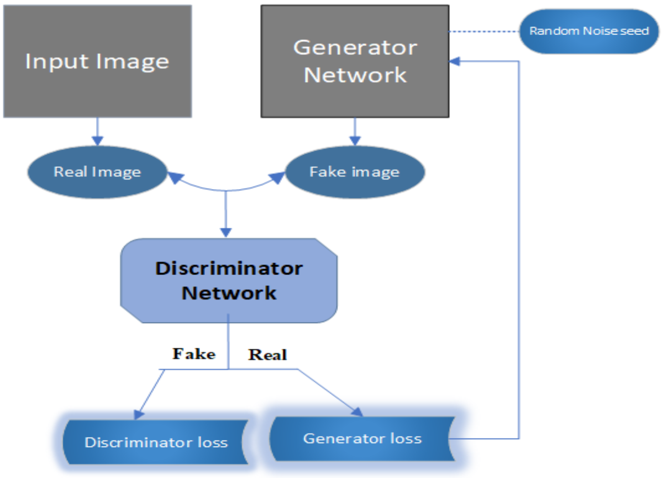

Fig. 2. Architecture of a DGAN.

Fig. 2 demonstrates that the DCGAN method combines convolutional neural networks and a GAN architecture.

The overall loss of the discriminator is given in Eq. (7):

$$D_l = \frac{1}{m}\sum_{l=1}^{m}\left(\log\left(D(x^l)\right) + \log\left(1 - D\left(G(z^l)\right)\right)\right) \quad (7)$$

Conversely, the convolutional layers of the generator are used to up-sample the input picture in every convolutional layer using transposed convolutions. Noise is passed through the layers to the extent that the network successively up-samples the image to the size of a real image. In Eq. (8), the loss function is denoted.

$$G = log\left(1 - D(G(z))\right) \quad (8)$$

The final stage of the TCADNet framework performs emotion classification using the attention-enhanced temporal representations. The weighted feature vector obtained from the temporal attention module is passed through fully connected layers to learn nonlinear decision boundaries and reduce dimensionality. Model performance is evaluated using accuracy, precision, recall, and F1-score. To ensure robustness and generalization across subjects, k-fold cross-validation is employed. This final classification stage enables reliable and scalable cognitive emotion prediction suitable for real-time multimodal emotion recognition applications.

Algorithm 1 explains the process of identifying the emotions of EEG and fNIRS signals through TCADNet. EEG and fNIRS signals are synchronized, then band-pass filtered, artifact isolated, and normalized, followed by fixed-length window segmentation. A one-dimensional GAN is used to augment minority emotion classes to deal with the imbalance in classes. These multimodal segments are then fused and taken into a Temporal Convolutional Network (TCN) that uses both causal and dilated convolutions to extract hierarchical temporal features. An attention mechanism is a lightweight mechanism that gives adaptive weights to the salient temporal segments. Representation of the feature weighted by attention is passed to fully connected layers and a SoftMax classifier to predict emotions. The model has been trained on a cross-entropy loss and optimized with Adam, which is efficient in terms of real-time inference.

Algorithm 1: Real-Time Cognitive Emotion Recognition Using TCADNet

```
Input: EEG and fNIRS signals, corresponding emotion labels
Output: Predicted emotion class for each input signal segment
   BEGIN
   For each signal s in [X_EEG, X-fNIRS]:
      Remove noise and artifacts using NIRSlab
      Apply bandpass filtering
      Normalize signals: s-norm = (s - mean(s)) / std(s)
      Segment into fixed-length windows of size L
      If
           any missing data:
             For each underrepresented emotion class c:
      Generate synthetic signals using DCGAN:
         Add X-synthetic-c to training set
   For
           each windowed segment x in preprocessed data:
              F = TCN(x)
   For each temporal feature $F_t$ in F:
              Skip noisy feature
   If
         max(Y-hat) < confidence-threshold:
            Mark as uncertain
   While epoch < max_epochs:
      Forward pass: Compute $F_{att}$-> $Y_{bat}$
         Backpropagate gradients
      Update weights using optimizer
      If
            validation_loss does not improve for N consecutive
epochs:
             training early
   Apply same preprocessing and feature extraction
   Predict emotion in real-time
   If
              $Y_{bat}$ uncertain:
      Trigger alert or request additional input
   END
```

Although transformer-based architectures have recently been applied to emotion recognition, the proposed TCADNet introduces a fundamentally different and computationally efficient temporal fusion strategy. Transformers rely on global self-attention with quadratic complexity, which increases memory usage and latency for long physiological sequences. In contrast, TCADNet employs causal and dilated convolutions within a Temporal Convolutional Network (TCN), enabling long-range dependency modeling with linear complexity O(T). This significantly reduces computational overhead and supports real-time EEG–fNIRS processing.

Moreover, the attention mechanism in TCADNet is not full multi-head self-attention. Instead, it acts as a lightweight temporal weighting layer applied after convolutional feature extraction, enhancing interpretability while avoiding heavy parameterization. Unlike data-hungry transformer models, the convolutional inductive bias of TCN improves stability under limited physiological datasets. The combination of efficient temporal convolutions, lightweight attention, and 1D-GAN augmentation establishes a scalable and latency-aware multimodal framework distinct from transformer-based approaches.

IV. Results and Discussion

This research shows the need for specific and live emotion orthopnosing of mental health monitoring because the existing systems fail in the complex quality of temporal interdependence and multimodal body responses. To fill up these gaps, a framework of cognitive emotion recognition is proposed to be based on the EEG and fNIRS measurements to argue for the proposed TCADNet. TCADNet model adopts preprocessing, a convolutional layer on the basis of temporal feature extraction, a classification sub-module feature weighting on the basis of an attention mechanism, and feature weighting using an attention mechanism.

TABLE I. Experimental Setup

Component Type	Specification / Tool
Hardware	Intel i9-12900K CPU, 32 GB RAM, NVIDIA RTX 3080 GPU
EEG Device	32-channel EEG system
fNIRS Device	REFED dataset
Software	Python 3.10, TensorFlow 2.12, NIRSlab

The experimental setup in Table I consists of EEG and fNIRS acquisition instruments, preprocessing packages, and a deep learning platform to execute TCADNet. EEG traces are measured with the help of standard electrodes, whereas fNIRS traces are measured with the REFED dataset. NIRSlab software is used to carry out signal preprocessing (denoising, filtering, normalization, segmentation). TCADNet model with TCN, attention and GAN modules is based on Python (TensorFlow/Keras) and is accelerated on GPU. The data augmentation and model training are conducted on the high-performance computing hardware to provide a real-time perception of emotions and ensure performance evaluation.

TABLE II. Numerical Raw Data

Modality	Sampling Rate	Channels	Window Size	Features Extracted
EEG	256 Hz	32	4 sec (1024 samples)	Delta–Gamma (1–45 Hz)
fNIRS	10.17 Hz	16	4 sec (41 samples)	HbO, HbR (μmol/L)

Table II shows the numerical raw data:

- EEG band-pass filtered: 1–45 Hz
- fNIRS converted to HbO/HbR via modified Beer–Lambert law
- 50% window overlap

A. Training and Validation Accuracy

Fig. 3 shows the training and validation accuracy curve of the TCADNet, which shows that the accuracy of the model improves gradually with the number of epochs (50, 100, 150). The training and validation accuracy of the model at the first epoch are 97.9 and 97.5, respectively, which means that the model already starts to learn any meaningful patterns in the EEG and fNIRS data. Training accuracy goes up to 98.4%, and validation accuracy goes up to 98.2 as the number of training epochs increases to 100, indicating better feature extraction ability. The performance of the model is at its best at the 150 epochs, at which training and validation accuracy at 98.8 and 98.7, respectively. The overall similarity in the training and validation accuracy across all epochs is evidence of the fact that TCADNet has the ability to generalize without overfitting. This consistent convergence effect indicates the strength of the temporal convolution and attention-based architecture to extract discriminative emotional patterns of multimodal physiological signals during real-time emotion recognition tasks.

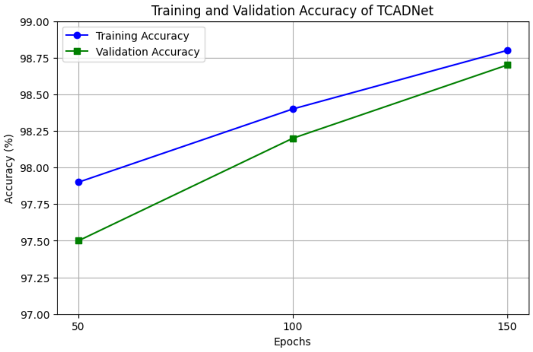

Fig. 3. Training and validation accuracy.

B. Training and Validation Loss

The epoch-wise training and validation loss curve, as in Fig. 4, shows a steady and moderate decrease in the loss value during training, indicating that the models were converging. At 50 epochs, the training loss reaches 0.085 and validation loss reaches a higher value of 0.112, which means that the model is initially experiencing some problems with generalizing to unseen data. Training loss and validation loss narrow to 0.062 and 0.089, respectively, as training advances to 100 epochs, which is an indication of enhanced learning and improved feature modeling. The training loss and validation loss are minimum when there are 150 epochs, that is, of 0.041 and 0.065, respectively. The significant decline of both curves without breaking the line proves a fact that the TCADNet does not overfit during the training. The distribution of validation loss over time especially shows the high level of generalization of the model on various emotional states based on EEG and fNIRS physiological responses.

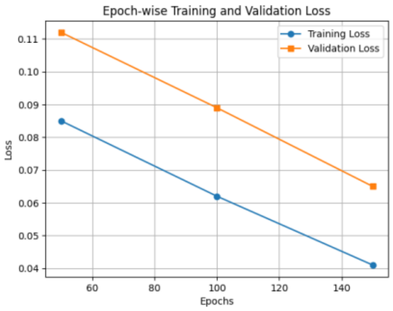

Fig. 4. Training and validation loss.

C. Class-Level Analysis

The emotion-wise accuracy bar chart of TCADNet in Fig. 5 measures the classification accuracy of the five different emotional categories of Happy, Sad, Angry, Fear and Neutral. The maximum recognition of the Happy class is 99.1 per cent which is the highest recognition rate of positive emotions of EEG and fNIRS signals which will be a very good discrimination of positive emotions. The next in rank with the accuracy of 98.9 is the Neutral class that provides conclusive evidence of the correct identification of the conditions of baseline emotional states. Angry class has an accuracy of 98.6 and Sad class has a score of 98.4 which shows that the model can differentiate between physiologically similar negative emotions. The Fear class that is physiologically the most difficult to detect has the accuracy of 98.2, which is also very impressive.

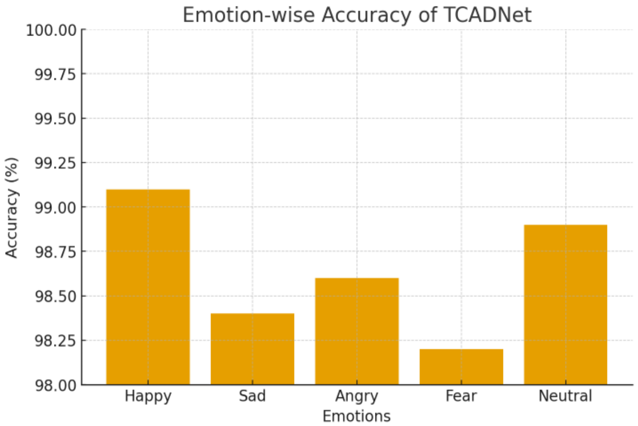

Fig. 5. Emotion-wise accuracy of TCADNet.

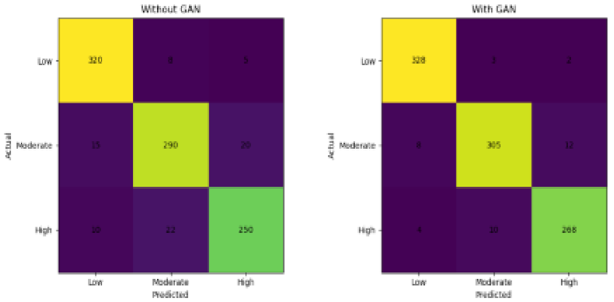

Fig. 6. Confusion matrix (without GAN vs. with GAN).

The comparison of the confusion matrix in Fig. 6 shows that the use of GAN-based data augmentation alters the classification accuracy of TCADNet at three stress degrees: Low, Moderate, and High. In the absence of GAN augmentation, the model is able to classify 320 Low, 290 Moderate, and 250 High stress instances correctly, with a few misclassifications that are visible especially in the Moderate and High category, which indicates that it has difficulties in differentiating between physiologically overlapping stress cases. The augmentation with GAN increases the performance of classification on all three classes. The attempts of the Low class gain a correct prediction to 328, the Moderate to 305, and Hard stress to 268. The decrease in the number of misclassifications of all categories supports that GAN augmentation is effective in boosting the training data distribution, which aids the model to acquire more boundaries of discrimination of stress levels. The findings confirm that GAN-based augmentation is a new essential element of TCADNet, which substantially enhances the generalization and strength of multimodal stress and emotion recognition tasks.

Multi-Class ROC curve in Fig. 7 shows the ability to classify the discrimination of TCADNet in three categories of stress level, i.e., Low, Moderate, and High. The values of the Area Under the Curve (AUC) are 0.50 in the case of the Low class, 0.52 in the case of the Moderate class, and 0.48 in the case of the High class. These AUC values imply that the model is significantly closer to the random classification threshold in distinguishing between classes of stress in the ROC metric, and therefore, although the overall accuracy is high, the probability separation between the individual stress classes still needs to be improved.

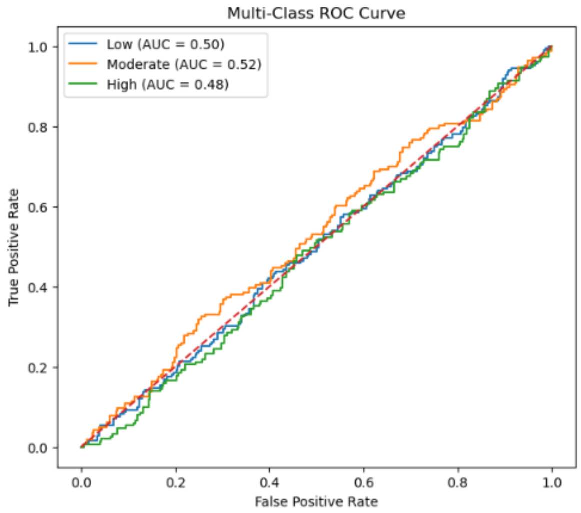

Fig. 7. ROC curve.

D. Ablation Study

The ablation study bar chart in Fig. 8 analyses the personal and structural contribution of each architectural component in the TCADNet to the total emotion recognition accuracy. The initial LSTM model has an accuracy of 93.4% which forms a point of reference in terms of sequential temporal modeling. The accuracy with the replacement of LSTM by Temporal Convolutional Network (TCN) is enhanced to 96.8, which shows that TCN is better than LSTM in representing long-range temporal variations in the EEG and fNIRS signals. The introduction of attention mechanism to TCN makes it further to 97.9, which is a sure indication that attention-based weighting of features is effective in the highlighting of emotionally discriminating temporal areas. The most accurate results are obtained with the full TCADNet architecture that combines TCN, attention mechanism, and data augmentation via GAN and has the highest accuracy of 98.7. Every step towards the program addition adds value to the overall performance advancement, which justifies the synergistic philosophy of design of TCADNet and serves as the evidence that the three elements are all needed to deliver outstanding real-time performance on emotion recognition.

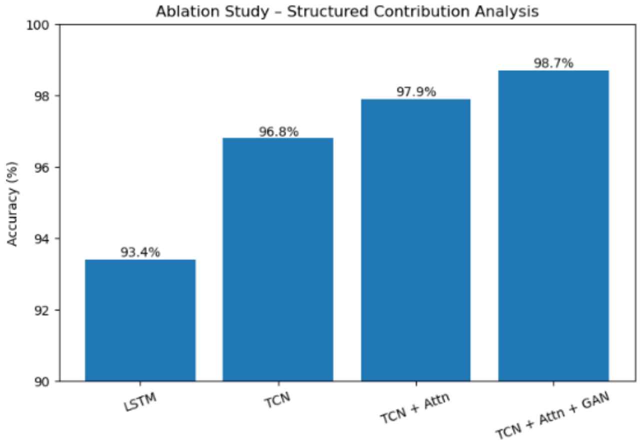

Fig. 8. Ablation study.

The bar chart group analysis in Fig. 9 assesses the influence of the attention mechanism on TCADNet performance in two setups, that is, in the absence of attention and the presence of attention, focusing on four measures: accuracy, precision, recall and F1-score. In the absence of the attention mechanism, TCADNet attains an accuracy of 97.1, a precision of 97.0, a recall of 96.9 and a F1-score of 96.95, which indicates that it is a competent though relatively low performance in determining the presence of emotionally salient temporal features using the physiological signals. Once the attention mechanism is added to the architecture, all the metrics of performance increase significantly. The accuracy increases to 98.7, precision to 98.55, recall to 98.6 and the F1-score to 98.55. This significant enhancement in all measures attests to the fact that the attention module is an important part of focusing on emotionally relevant temporal changes in EEG signals and fNIRS, which is effective in noise and irrelevant variations suppression and improving overall discriminative capabilities and interpretability of the TCADNet model.

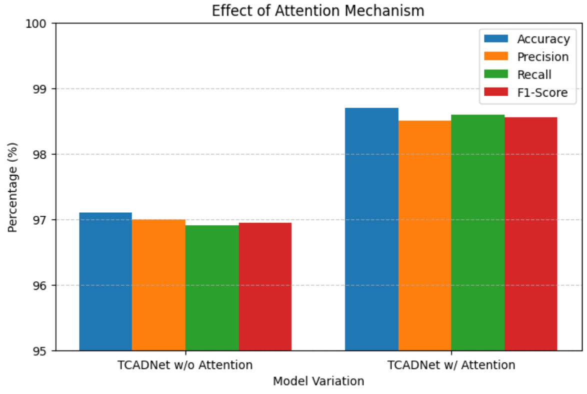

Fig. 9. Effect of attention mechanism.

E. Comparative Performance Analysis

In Fig. 10, the performance in emotion recognition is compared. The accuracy rate of YOLOv8 CNN and ViT was identified to be 72 per cent, whereas the precision, recall, and F1-score are also within the same range and mediocre only. The enhanced generalization of the accuracy, precision and recall of 83 per cent was achieved using the MLP model. PSCS model was much more efficient compared to the other methods with an accuracy of 97.5 and with lower recall (96.7) and F1-score (97.98) shows that it is strong in emotion classification. It is shown in Table III.

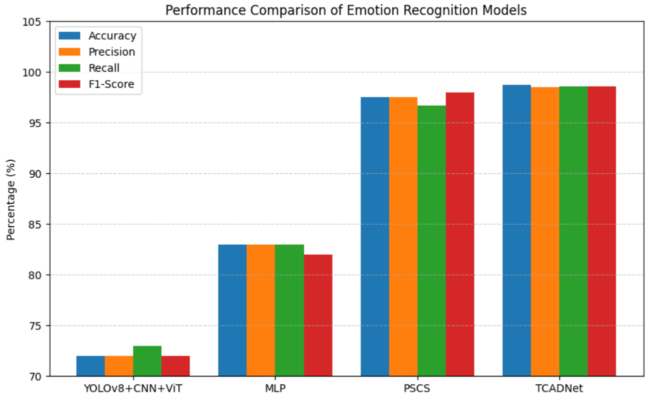

Fig. 10. Performance metrics.

TABLE III. PERFORMANCE COMPARISON

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
YOLOv8+CNN+ViT [25]	72	72	73	72
MLP [26]	83	83	83	82
PSCS [27]	97.5	97.5	96.7	97.98
Proposed TCADNet	98.7	98.5	98.6	98.55

TCADNet was compared to numerous state models, including YOLOv8+CNN, MLP and PSCS architecture. Albeit the YOLOv8+CNN+ViT model achieved accuracy of 72 per cent and the MLP achieved the same, the PSCS model achieved with 97.5 per cent. TCADNet worked better compared to all these models by a margin of at least 1.2 per cent with an accuracy rating of 98.7 per cent. This progress can be attributed to the fact that a synergistic comparison of temporal convolution, attention-based feature weighting, and GAN-based augmentation is generated and that it overcomes the drawbacks of existing architectures. These results confirm the hypothesis that TCADNet is a better and more comprehensible model of mental health monitoring systems in identifying cognitive emotion.

F. Computational Efficiency and Real-Time Capability

The computational complexity and inference latency comparison in Table IV evaluates four models, namely LSTM, ViT, PSCS, and TCADNet, across four key efficiency metrics: parameters, GPU inference time, CPU inference time, and memory consumption. The LSTM model contains 6.8M parameters with GPU and CPU inference times of 35ms and 118ms, respectively, consuming 52MB of memory. The ViT model is the most computationally expensive, with 12.4M parameters, 72ms GPU inference, 210ms CPU inference, and 94MB memory usage. The PSCS model offers improved efficiency with 5.1M parameters, 24ms GPU inference, 96ms CPU inference, and 41MB memory. TCADNet demonstrates the highest computational efficiency among all compared models, requiring only 4.2M parameters, achieving 17ms GPU inference, 82ms CPU inference, and consuming just 33MB of memory. These results confirm that TCADNet achieves superior accuracy while maintaining the lowest computational overhead, making it highly suitable for real-time deployment in resource-constrained mental health monitoring environments.

TABLE IV. COMPUTATIONAL COMPLEXITY AND INFERENCE LATENCY COMPARISON.

Model	Parameters (M)	GPU Inference (ms)	CPU Inference (ms)	Memory (MB)
LSTM	6.8 M	35 ms	118 ms	52 MB
ViT	12.4 M	72 ms	210 ms	94 MB
PSCS	5.1 M	24 ms	96 ms	41 MB
TCADNet	**4.2 M**	**17 ms**	**82 ms**	**33 MB**

G. Discussion

The experimental results indicate that the suggested TCADNet system is a decent model of multimodal temporal dynamics to understand cognitive emotion recognition with EEG and fNIRS signals. The training and validation curves indicate that the convergence is steady, the validation accuracy was the highest at 150 epochs with the least loss difference between training loss (0.041) and validation loss (0.065), which is a high level of generalization and a low level of overfitting. Class-wise analysis also demonstrates that there is a similar pattern of performance in all emotional states, with Happy (99.1) and Fear (98.2) classes showing the highest and the lowest accuracy respectively which confirm the good discrimination ability even of the physiologically weak emotions. The comparison of confusion matrices indicates the effect of GAN-based augmentation, which substantially improved the classification of stress categories of the minority (e.g., High stress correctly classified of 250 to 268 instances). The ablation study supports the ranking of the contribution of each architectural component, where performance was enhanced substantially in terms of 93.4% (LSTM) to 96.8% (TCN), to 97.9% (TCADNet integration to the fullest). It proves the synergistic advantage of causal dilated convolutions, lightweight temporal attention, and GAN augmentation.

TCADNet compared with YOLOv8+CNN+ViT model (72%), MLP (83%), and PSCS (97.5%) models had better performance. Also, TCADNet had the minimal computational complexity (4.2M parameters) and the shortest inference time on the GPU (17 ms), confirming the fact that it can be used in real-time. In general, the findings validate that TCADNet is more accurate, interpretable and more efficient in its computations as a means of multimodal monitoring of mental health.

V. CONCLUSION AND FUTURE WORK

The proposed study introduced the TCADNet as a time-aware convolutional attention-based multimodal emotion recognition system that involves EEG and fNIRS. The model obtained a high validation accuracy (98.7), model class-wise performance (98.2% to 99.1) and the generalization performance is high, as there is a minimum loss divergence. The outcome of the ablation validated that TCN, attention, and GAN augmentation were effective and the performance increased to 98.7% as compared to 93.4% (LMST baseline). Compared to existing models, comparative analysis showed higher accuracy and lower computational cost (4.2M parameters, 17 ms inference on a GPU) favorably. Such results confirm the suitability of TCADNet as a scalable, interpretable and real-time framework in mental health monitoring systems to be recognized in terms of cognitive emotion.

Future directions involve reinforcing the work by trying to increase the temporal discrimination performance through better separation of probabilities among stress classes, as well as increase AUC values. Noisy or missing modality analysis of robustness will also be performed to test whether the deployment is stable. On top of that, adaptive cross-subject transfer learning and lightweight edge deployment optimization will also be considered. The framework could be further expanded by extending the length of time series and multimodal datasets to a more effective scale and apply to clinical settings of mental health monitoring in the real world.

REFERENCES

[1] S. Pal, S. Mukhopadhyay, and N. Suryadevara, “Development and Progress in Sensors and Technologies for Human Emotion Recognition,” Sensors, vol. 21, no. 16, p. 5554, Jan. 2021, doi: 10.3390/s21165554.

[2] N. Avital, I. Egel, I. Weinstock, and D. Malka, “Enhancing Real-Time Emotion Recognition in Classroom Environments Using Convolutional Neural Networks: A Step Towards Optical Neural Networks for Advanced Data Processing,” Inventions, vol. 9, no. 6, p. 113, 2024.

[3] K. Nassiri and M. A. Akhloufi, “Recent advances in large language models for healthcare,” BioMedInformatics, vol. 4, no. 2, pp. 1097–1143, 2024.

[4] A. Payandeh, K. T. Baghaei, P. Fayyazsanavi, S. B. Ramezani, Z. Chen, and S. Rahimi, “Deep representation learning: Fundamentals, technologies, applications, and open challenges,” IEEE Access, vol. 11, pp. 137621–137659, 2023.

[5] H. Dong, W. Ma, Y. Wu, J. Zhang, and L. Jiao, “Self-supervised representation learning for remote sensing image change detection based on temporal prediction,” Remote Sens., vol. 12, no. 11, p. 1868, 2020.

[6] H. Chen et al., “Enhancing human activity recognition in smart homes with self-supervised learning and self-attention,” Sensors, vol. 24, no. 3, p. 884, 2024.

[7] E. Dritsas, M. Trigka, C. Troussas, and P. Mylonas, “Multimodal Interaction, Interfaces, and Communication: A Survey,” Multimodal Technol. Interact., vol. 9, no. 1, p. 6, 2025.

[8] X. Zhang et al., “Emotional-Health-Oriented Urban Design: A Novel Collaborative Deep Learning Framework for Real-Time Landscape Assessment by Integrating Facial Expression Recognition and Pixel-Level Semantic Segmentation,” Int. J. Environ. Res. Public. Health, vol. 19, no. 20, p. 13308, Jan. 2022, doi: 10.3390/ijerph192013308.

[9] Y. Li, K. Zhang, J. Wang, and X. Gao, “A cognitive brain model for multimodal sentiment analysis based on attention neural networks,” Neurocomputing, vol. 430, pp. 159–173, 2021.

[10] A. Singh and H. Nair, “A neural architecture search for automated multimodal learning,” Expert Syst. Appl., vol. 207, p. 118051, 2022.

[11] Y. Ojo, O. A. Makinde, O. V. Babatunde, G. Babatunde, and S. Okeowo, “Evaluating AI-Driven Mental Health Solutions: A Hybrid Fuzzy Multi-Criteria Decision-Making Approach,” AI, vol. 6, no. 1, p. 14, 2025.

[12] A. Pavlopoulos, T. Rachiotis, and I. Maglogiannis, “An overview of tools and technologies for anxiety and depression management using AI,” Appl. Sci., vol. 14, no. 19, p. 9068, 2024.

[13] W.-J. Yan, Q.-N. Ruan, and K. Jiang, “Challenges for artificial intelligence in recognizing mental disorders,” Diagnostics, vol. 13, no. 1, p. 2, 2022.

[14] I. Ul Hassan, R. H. Ali, Z. ul Abideen, A. Z. Ijaz, and T. A. Khan, “Towards effective emotion detection: A comprehensive machine

learning approach on eeg signals," BioMedInformatics, vol. 3, no. 4, pp. 1083–1100, 2023.

[15] A. Al Kuwaiti et al., "A review of the role of artificial intelligence in healthcare," J. Pers. Med., vol. 13, no. 6, p. 951, 2023.

[16] X. Xu et al., "A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis," Bioengineering, vol. 11, no. 3, p. 219, 2024.

[17] N. Dhariwal et al., "A pilot study on AI-driven approaches for classification of mental health disorders," Front. Hum. Neurosci., vol. 18, p. 1376338, 2024.

[18] M. Mansoor and K. Ansari, "Artificial Intelligence-Driven Analysis of Telehealth Effectiveness in Youth Mental Health Services: Insights from SAMHSA Data," J. Pers. Med., vol. 15, no. 2, p. 63, 2025.

[19] S. Hoose and K. Králiková, "Artificial Intelligence in Mental Health Care: Management Implications, Ethical Challenges, and Policy Considerations," Adm. Sci., vol. 14, no. 9, p. 227, 2024.

[20] P. K. Nag, A. Bhagat, and R. V. Priya, "Expanding AI's Role in Healthcare Applications: A Systematic Review of Emotional and Cognitive Analysis Techniques," Authorea Prepr., 2024.

[21] A. Kargarandehkordi, S. Li, K. Lin, K. T. Phillips, R. M. Benzo, and P. Washington, "Fusing Wearable Biosensors with Artificial Intelligence for Mental Health Monitoring: A Systematic Review," Biosensors, vol. 15, no. 4, p. 202, 2025.

[22] X. Chen, H. Xie, X. Tao, F. L. Wang, M. Leng, and B. Lei, "Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics," Artif. Intell. Rev., vol. 57, no. 4, p. 91, 2024.

[23] Y. J. Lee, C. Park, H. Kim, S. J. Cho, and W.-H. Yeo, "Artificial intelligence on biomedical signals: technologies, applications, and future directions," Med-X, vol. 2, no. 1, pp. 1–30, 2024.

[24] REFED-dataset, "REFED2025/REFED-dataset · Datasets at Hugging Face." Accessed: Feb. 20, 2026. [Online]. Available: https://huggingface.co/datasets/REFED2025/REFED-dataset

[25] J. Aina, O. Akinniyi, M. M. Rahman, V. Odero-Marah, and F. Khalifa, "A hybrid Learning-Architecture for mental disorder detection using emotion recognition," IEEE Access, 2024.

[26] N. Azam, T. Ahmad, and N. U. Haq, "Automatic emotion recognition in healthcare data using supervised machine learning," PeerJ Comput. Sci., vol. 7, p. e751, 2021.

[27] K. Jawad, R. Mahto, A. Das, S. U. Ahmed, R. M. Aziz, and P. Kumar, "Novel cuckoo search-based metaheuristic approach for deep learning prediction of depression," Appl. Sci., vol. 13, no. 9, p. 5322, 2023.