

Multi-Modal Bio signal Transformer Based Deep Learning Approach for Mental Stress Detection

Maria Saleem, Bushra Naz Soomro, Almas Ali, Zartasha Baloch

Mehran University of Engineering and Technology, Jamshoro

Received: 13 February 2025, Revised: 15 December 2025, Accepted: 24 December 2025, Published: 01 January 2026

KEY WORDS	ABSTRACT
Transformer Physiological Signals Stress Detection Deep Learning Convolutional Neural Network	Mental stress significantly affects both the emotional and physical health of the person making its accurate detection crucial for preventive mental healthcare. Our study presents a hybrid Convolutional Neural Network–attention-based Transformer architecture for stress detection using physiological signals and sensor data, Electroencephalogram, Electrocardiogram, Electromyogram, Electrodermal Activity, Respiration, and Temperature from the publicly available Wearable Stress and Affect Detection dataset. The CNN technique efficiently extracts spatial and temporal features from bio signal data, while the attention-based Transformer, equipped with multi-head attention and layer normalization, captures long-range dependencies and inter-signal relationships. Through optimized preprocessing, feature standardization, and balanced class training, the model effectively classifies three classes: baseline, stress, and amusement. Our model achieves a 96% accuracy, outperforming existing deep-learning and machine-learning approaches in robustness and generalization. These findings demonstrate our model’s capability for real-time mental stress recognition, with promising applications in wearable health monitoring and intelligent human-computer interaction systems.

1. Introduction

Stress can be the most threatening factor in the daily life of many people; not only it influences emotional stability but also lack motivation and cognitive focus [1]. It has been regarded as a state of anxiety, and this is brought about by mental and emotional issues [2]. Stress leads to observable thermal, electrical, impedance, acoustic, and optical transformations in an individual. It is considered a common reaction to external economic forces, usually caused by inflation which adds to the unemployment rates and domestic abuse, among numerous other causes [3]. There are positive and negative ways of stress. Positive stress is usually based on feelings, such as happiness or passion, which energize the brain and make it more alert. In contrast, negative stress typically harms an individual’s well-being, performance, and attention [4]. Many studies have demonstrated that high levels of negative stress impair decision-making, diminish motivation, induce guilt, lower self-esteem, reduce situational awareness, and hinder overall performance. As a result, stress is often

perceived primarily in its negative form [5]. Wearable sensors and artificial intelligence are developing at an accelerating rate in our environment for mental stress, and physiological metrics can identify stress. Several researchers have discovered that the human nervous system’s autonomous function is linked to stress [6]. Experiencing stressful situations can initiate physiological reactions, including the activation of the sympathetic-adrenal-medullary (SAM) axis and an elevated secretion of metanephrine [7]. These physiological responses result in bronchodilation, higher vascular tone, improved cardiac contractility, elevated blood pressure and heart rate, and other associated consequences of health [8].

ECG-based techniques mostly depend on the RR interval and the P- and T-wave forms of the human body [9]. Similarly, EEG signals are crucial in the neuroscience field, cognitive research, and stress detection, as they reflect neural activity associated with stress, relaxation, and emotional states [10]. EMG detects muscle activation of the human body, making it a valuable tool for mental stress detection

[11]. EDA signals are commonly used for mental stress detection, as they measure sweat gland activity, which is controlled by the autonomic nervous system [12, 13].

Artificial intelligence has recently advanced due to the deep learning field [14]. Many researchers have used neural networks and attention-based techniques to infer psychological stress [15]. For stress analysis, a variety of neural network topologies have been used, including long short-term memory networks, gated recurrent units, and recurrent neural networks [16]. Multimodal fusion to detect stress using an attention-based convolutional neural network with long short-term memory [17]. Wearable devices are used to gather single-lead ECG data, and other sensor data utilize a 34-layer residual CNN to make a diagnosis [18]. LSTM and DNN on signals from ECG, EDA, and EMG, Using ECG and RESP as well as TEMP data, binarized CNN and LSTM models are employed for stress detection [19]. Although these techniques achieve strong performance, they often overlook the cyclical patterns inherent in physiological signals. Furthermore, deep learning models might stick in local optima due to their complexity. Because deep networks have many parameters, training is slow [20]. We suggest a novel paradigm with a transformer network to get around these issues. The transformer network, which does not have too complex a structure and solely depends on the attention mechanism, is first used in translation and natural language processing. To contextualize the words, the transformer network can encode the position information. It is thought that physiological signals can be managed similarly [21].

Transformer networks perform better than individual CNNs when trained on small data sets, and they are less effective at extracting local characteristics, even when they perform well on large-scale data [22]. There is limited research on applying transformer networks to stress recognition using physiological data, particularly regarding the effective integration of CNNs with transformers [23]. Our model introduces a novel approach for stress detection by utilizing multimodal bio signals (EEG, ECG, EMG, and EDA) in combination with deep learning techniques, specifically a CNN integrated with a Transformer architecture that employs multi-head attention and layer normalization. This robotized approach can be involved in stress detection and has great potential for improving mental health care [24]. Model, multi-head attention, layer normalization transformer and convolutional neural network. A convolutional neural network gathers spatial features of the model to extract features, and the transformer captures temporal features. This action is critical to the training setup of the model, where we apply the

technique of Adam optimizer and monitor learning by the categorical cross-entropy loss function. It also contains such measures of evaluation as accuracy, precision, recall, and F1-score to test the performance of the model. Finally, this evaluation approach categorizes the evaluation into three multi-emotional categories, including stress, amusement, and baseline.

2. Methodology And Methods

Our studies main goal was to create an excellent hybrid CNN transformer deep learning approach that could use multimodal bio signals to identify mental stress. The methodology follows four important phases, as shown in the methodology framework. Figure 1. 1: Data Acquisition, 2: Preprocessing, 3: Model Building and Training, and the last is 4: Evaluation. Input Data is collected from the WESAD dataset, including multimodal physiological signals and sensor data such as electromyograms, electrodermal activity, electroencephalograms, temperature, and respiration. We used standardization and down sampling for the preprocessing to reduce computational overhead and clean, normalize, and segment the signals for a smooth input. Also, we used a deep learning

2.1 Data Acquisition

Our stress and affect detector package, that is published online as the Wearable Stress and Affect Detection, has been used as the benchmark in our study. It contains sensor and physiological measurements of 15 respondents (3 females, 12 males) in various conditions in a controlled laboratory facility. Electroencephalogram is used to measure brain activity, electrocardiogram is used to check heart rates and heart rhythm, electromyogram is used to measure the activity of muscles and electrodermal activity to check skin conductance which varies according to stress conditions. Body temperature changes are recorded by temperature checks, and breathing patterns are recorded by respiration. Data on each subject has been annotated with respective class states allowing a guided learning strategy to be used in identification of mental stress.

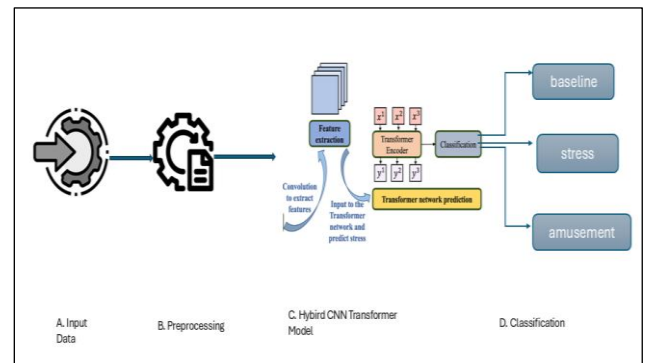


Fig. 1. Methodology Framework

2.2 Data Preprocessing

These data preprocessing methods are used in preprocessing to form a hybrid CNN-Transformer deep learning model, which is accurate and computationally efficient. Since preprocessing is key to converting unstructured, noisy, high-frequency physiological features into a simple format that is easy to work with models and in addition allows the data set to retain the relevant, inherent patterns in a data set and reduce noise. 1. Down sampling: All signals are down sampled at all frequencies by a factor of 10 to minimize memory requirements and computational requirements without losing any important information. 2. Standardization: Standard Scaler method was used to scale the data such that similar signals will have the same value of the feature. This enhances our model's convergence and accuracy and the 3. Data Structuring: The preprocessed signals are transformed into a 3D tensor format. (SEQ_LENGTH, FEATURES)(SEQ_LENGTH, FEATURES)(SEQ_LENGTH, FEATURES), which is the required input for our hybrid CNN-Transformer model. 4. Data Splitting: The data is divided into two parts with an 80:20 ratio for training and testing. This split ensures unbiased performance in the evaluation of the model, and last is 5. Batch Processing: in which efficient TensorFlow datasets are generated for streamlined model training.

2.3 Model Architecture

We provide a hybrid model that combines Transformer encoders to learn long-range dependencies with Convolutional Neural Networks (CNNs) for capturing local features to efficiently capture both local and global temporal patterns in multimodality. Figure 2 shows the entire framework of the suggested CNN-Transformer architecture.

The input to the model is a one-dimensional signal vector of shape (25×1) , which represents 25 features extracted from EEG and EMG and other signals following the preprocessing and feature extraction stages. This input is first processed by a 1D convolutional layer with a kernel size of 3 and 64 filters, enabling the network to extract short-range, localized temporal features from the signal. The output of this convolutional layer undergoes ReLU activation, resulting in a feature map of shape 23×64 .

The study of our model proceeded more smoothly as a result, and further adjustments were implemented consistently. The model's capacity to generalize across different physiological stress levels was improved by adding data from a diverse range of patients to the dataset. The target labels window for the three classes, Baseline, Stress, and Amusement, were removed to

prepare the feature matrix for supervised learning of the model.

Because they allow the learning algorithm to connect signal patterns to stress-related processes. Because of the multimodal character of the dataset, the feature values vary greatly in both scale and distribution level; for example, multimodal bio signals frequently show higher amplitude ranges than sensor temperature and respiration. All feature sets were standardized using scikit-learn's StandardScaler approach, which converts each feature to have a zero mean and unit variance for the model to reduce the possibility of such disproportionate influence during model training. Because it improves convergence, stabilizes gradient descent, and guarantees that the learning algorithm handles all input dimensions equally, this step is crucial for deep learning models. Additionally, the input was converted into a 3D dimension tensor format for the standardized scaler technique, which is essential for sequential learning deep learning architectures like CNNs and Transformer networks. Features, sequence length, and samples are the dimensions of the tensor in that order. In multivariate bio signals, which are time-series data structures, our model can identify both short-term spatial dependencies and long-term temporal patterns. By combining the advantages of CNNs and Transformers, two deep learning models, hybrid architecture captures fine-grained local information and models the global temporal environment, producing a reliable and efficient multimodality solution for mental stress detection.

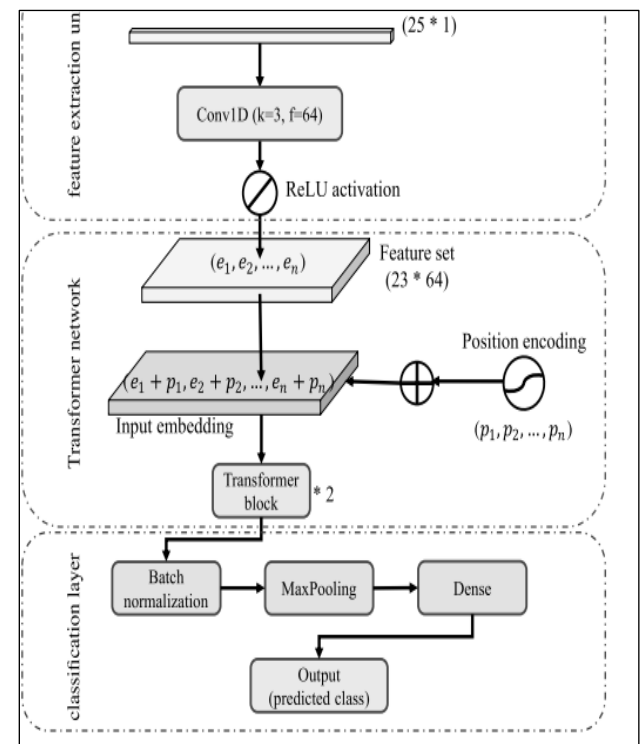


Fig. 2. The Architecture of the Hybrid Transformer-Based Model.

2.3.1 Feature Extraction using CNN

In the first part of the model, we used one-dimensional convolutional layers (Conv1D) to extract localized temporal and spatial information from the raw input data. These layers do a decent job of identifying the different model classes. ECG, EEG peaks, and EMG waveforms are a few instances of short-term patterns in time-series data. This is accomplished by using three different Conv1D layers, each with a fixed kernel size of three and rising filter sizes of 64, 128, and 256. For its advancement, the network can now recognize features at various scales and resolutions. For smooth learning, an activation function called a Rectified Linear Unit comes after each convolutional layer.

2.3.2 Transformer Encoder

After feature extraction, the output representations are fed into the Transformer Encoder, which is designed specifically to simulate temporal dependencies and contextual interactions throughout the series of bio signals and sensory data. Because attention-based transformer design can handle sequential data without depending on recurrent aspects of the model, it is ideal for processing physiological signals. Text and natural language processing were the initial applications of transformer architecture. To maintain the temporal structure of the input sequence, transformers apply a positional encoding to the feature vectors rather than automatically encoding order. Encoding allows the attention processes to learn from the relative and absolute order of the sequence by including information about each step's location in the model.

The Transformer encoder's multi-head self-attention mechanism allows the model to compare the relationships between many sequences sets and pay attention to multiple time series data at once. Each of the four attention heads in this model learns distinct representations and interactions within the signal timeline. The model's capacity to capture both short-term fluctuations and long-term dependence is improved by this parallelization strategy. After going through a Layer Normalization stage, the self-attention mechanism's output is combined with the input utilizing residual connections.

During training, these residual linkages enhance gradient flow and help maintain the original input signal, stabilizing the deep learning networks. The data is fed through a position-wise feed-forward network after the attention mechanism, which further processes each stage separately to improve the retrieved representations. To guarantee the transformer's uniformity and seamless convergence, a

second layer of normalization is used. The Transformer encoder is a potent tool for comprehending the dynamic nature of physiological states during three distinct classes of the model because it efficiently captures cross-temporal linkages and modality interactions that may span several seconds. Figure 3 depicts the Transformer encoder's whole arrangement.

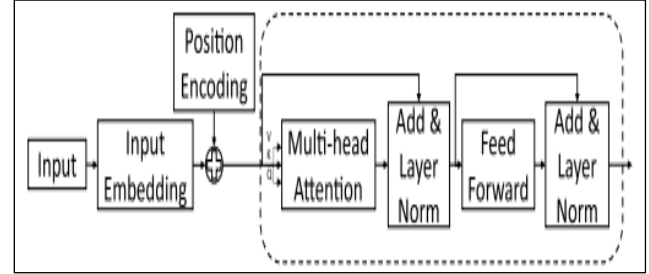


Fig. 3. Transformer Encoder

2.3.3 Classification Layer

We extracted features through dense layers to perform classification. After SoftMax activation which generates probabilities for each class. We selected Different hyperparameters and optimized for both the CNN and Transformer components of the model, as shown in Table 1 key parameters. These included the number of filters, kernel size, activation function, and pooling strategy for the CNN, as well as the number of Transformer blocks, attention heads, and positional encoding method for the Transformer network.

Table 1

Key Parameters Of The Model

Hyperparameter	Description
Conv1D Layer 1	64
Conv1D Layer 2	128
Conv1D Layer 3	256
Kernal Size	3
Polling Layer	2
Global Max Polling	1D
Multi-Head Attention Heads	4
Dropout Rate	0.5
Epochs	20
Transformer Key Dimensions	32
Dense Layer	128

2.4 Training and Evaluation

2.4.1 Training

We trained models by using the following configurations: Part A. Optimizer: Adam optimizer with a learning rate of 0.0001 for effective gradient-based optimization. Part B. Loss Function: Sparse Categorical Cross-Entropy was applied to minimize classification error. Part C. Batch Size: We trained the

model using a 32-batch size. Part D Epochs: Training conducted for 20 epochs.

2.4.2 Evaluation Metrics

Several important criteria were employed in the assessment metrics section to assess the model's performance and ascertain how effectively it is classified in three of the model's tasks. The main objective was to measure the total classification and correctness accuracy of the model, which was defined as the proportion of properly predicted labels across all subjects. Precision, which was computed as the ratio of True Positives to the sum of True Positives and False Positives, shows the percentage of pertinent examples among the recovered stages, providing more insight into the model's performance. Generally, the ratio of True Positives to the sum of TP and False Negatives and sum of FN was used to calculate overall recall of the model, which measures the model's capacity to accurately identify each pertinent aspect of the data. To balance the trade-off functions between precision and recall, we used a single score that considers all performance criteria also we used Confusion Matrix to graphically represent model's performance among each multiclass classification; it highlights instances of misclassification and allows for a more comprehensive understanding of the model's advantages as well as disadvantages. The F1-Score, which is the harmonic meaning of the two, was also taken into consideration for the calculation of the model.

2.4.3 Feature Analysis

In Feature Analysis for our model, we utilize Principal Component Analysis, which reduces data dimensions of the model and visualizes the feature space in 1 dimension (1D), which is basically used with the bio signals, and Correlation Analysis, which determines the relationships between unique features of the model as well as their impact on stress detection classes.

3. Results and Discussion

The result of our model clearly indicates how well the hybrid CNN-attention-based Transformer model differentiates between the three mental stress classes: baseline, stress, and amusement. The most important physiological rhythms that define each scenario were effectively captured and taught by our model's dual-stage architecture, in which CNN is employed for local feature extraction, whereas transformer encoders capture global temporal pattern recognition. The baseline condition, which suggests a tranquil and emotionally neutral state, was indicated by consistent heart rate patterns in the electrocardiogram and

reduced levels of activity in the electromyogram electrodermal signals.

The deep learning model retrieved over time from localized low variance features to teach the convolutional layers these muted and consistent signal behaviors. Whereas accuracy of our model indicates how sensitive it is to even the smallest physiological changes and how well it can recognize subjects' stable conditions.

Moreover, many bio signals, such as stiff muscles and elevated sweat gland activity linked to stress conditions, indicate strong activation in the stress state. The confusion matrix displays the model's high recall and accuracy in multi-class identification. This implies that by successfully capturing the extended oscillations and temporal surges that define stress events utilizing the transformer encoder's self-attention mechanism, the model was also able to distinguish stress-indicative patterns across modalities and time. Because the amused state once more showed physiological arousal that was like stress due to increased heart rate and facial muscle activity, it is difficult to categorize.

While our model continued to function well, it suggested that it could use small features in the multimodal data to distinguish between positive and negative arousal of the subjects. Our multimodal self-attention and layer normalization mechanism was essential in recognizing the minute differences between the respiratory patterns and signals in the Amusement, Baseline, and Stress classes, which frequently differed in rhythm and coherence of the signals.

It shows how the multi-head and layer normalization Transformers can identify long-range dependencies and subtle multi-modality interactions that are challenging for conventional models to discern. These findings were further supported by PCA heatmaps, which showed distinct feature clusters for each of the three different classes. The model acquired class-specific feature embeddings with little overlap, as evidenced by the separation of clusters for Baseline, Stress, and Amusement in the reduced feature space.

Our models' overall findings support the ability of the suggested hybrid deep learning model to differentiate among the mental detection states, which are baseline, stress, and amusement. The ability of CNNs to extract short-term spatial and temporal information and Transformers to represent the long-range dependencies necessary to comprehend the complete context of emotional changes across time are credited with this accomplishment.

Additionally, the results demonstrate that multimodal physiological data can be employed for fine-grained emotional state classification with the development of our mental health stress-monitoring systems.

3.1 Experimental Setup

To ensure efficient and accurate processing of physiological signals in our data set, we put all input data through a structured preprocessing and preparation pipeline. The raw signals, extracted from the WESAD dataset, included EEG, EMG, ECG, EDA, temperature, respiration, and accelerometer data, along with their corresponding stress-level labels. For consistency and efficiency of the model, all physiological signals were down sampled by a factor of 10; down sampling effectively reduced the computational overhead without compromising the essential temporal dynamics of the data.

Moreover, the down sampling of the bio signals from all subjects was concatenated into a single unified data frame, and the feature set was standardized using the StandardScaler technique. This standardization step normalized the distribution of each feature to zero mean and unit variance, promoting numerical stability and improving models' convergence during the model training.

After that, the standardized features were then reshaped into a 3D tensor format, conforming to the input requirements of the CNN-Transformer hybrid model as shown in the detailed model architecture in Figure 2 and Figure 3, the architecture of the model. The dataset was subsequently partitioned into two parts, training and testing subsets, maintaining an 80:20 ratio (80 for training and 20 for testing). While TensorFlow Datasets were created from these partitions to enable efficient batch processing and shuffling during training.

The hybrid deep learning convolutional neural network-Transformer model was designed to extract both local and long-range dependencies in the multimodality sequences. The CNN component, composed of multiple Conv1D layers with varying filter sizes and kernel parameters, served to extract spatial and temporal features from the input signals. A GlobalMaxPooling1D layer was used to reduce feature dimensionality while preserving the most salient and important information.

Moreover, the output of the convolutional neural network was passed to the Transformer encoder block, which used multi-head and self-attention mechanisms to capture global temporal dependencies and contextual interactions across the signal sequence. While positional encoding was applied to retain the

order of the input sequence, and layer normalization was used to stabilize training of the model.

Our hybrid CNN multimodal transformer-based deep learning model was finalized with a series of dense layers, culminating in a SoftMax-activated activation function output layer that predicted one of three mental stress classes, which are Baseline, Amusement, and Stress. After that, the model was optimized with the Adam optimizer technique, using a learning rate of 0.0001, and the loss was calculated with the Sparse Categorical Cross-Entropy loss function, which is appropriate for multiclass classification frameworks.

Moreover, our model's performance was evaluated on the test dataset using metrics accuracy, precision, recall, and F1-score, along with a confusion matrix to visualize classification outcomes. After that, principal component analysis was applied to project high-dimensional features into a 2D space for visualization, and correlation heatmaps were generated to explore relationships among different features of the dataset. This experimental setup was designed to reduce model complexity and enable the model to effectively learn discriminative physiological patterns of different physiological signals associated with varying mental stress detection.

3.2 Training

For training we divided the data set into two sections 80% for training and 20% for testing then we used for our model Adam optimizer technique with a learning rate of 0.0001 and the Sparse Categorical Cross-Entropy function, which is appropriate for multiclass classification deep learning models, using 80% training and 20% for testing to evaluate the model's capacity for generalization. To keep an eye on the models' performance and avoid overfitting, validation was carried out on the test set during the training phase.

The graphical analysis in our model illustrates the learning behavior, classification performance, feature distribution, and comparative evaluation of the hybrid CNN-Transformer deep learning model as shown in the Table 2 compliance report. Training and validation loss curves plotted over 20 epochs as illustrated in Figure 4 the X-axis denotes the number of epochs, while the Y-axis shows the loss value of the graph. Two lines are displayed: the blue line indicates the training loss, while the orange line represents the validation loss of the model. The downward trend in both curves suggests steady model improvement and minimal overfitting, indicating strong generalization performance of the model.

And in Figure 5 training and validation accuracy which illustrates the curves, in the Y-axis represents accuracy percentage, while the X-axis indicates the number of epochs. The blue curve, which is training graph and the orange curve validation rise steadily, indicating that the model effectively captures discriminative features relevant to stress across the three classes Baseline, Amusement, and Stress.

The error of validation reduced with the advancement and the validation percentage showed a steady increase throughout the training epoch. It can be considered a successful case of the model to acquire the discriminative qualities to classify stress. The lack of overfitting and the steady increase in the quality of the validation process are the qualities of the model that help to illustrate its capability to generalize well to unknown data, which is a critical feature of the real-world implementation of physiological stress detection systems.

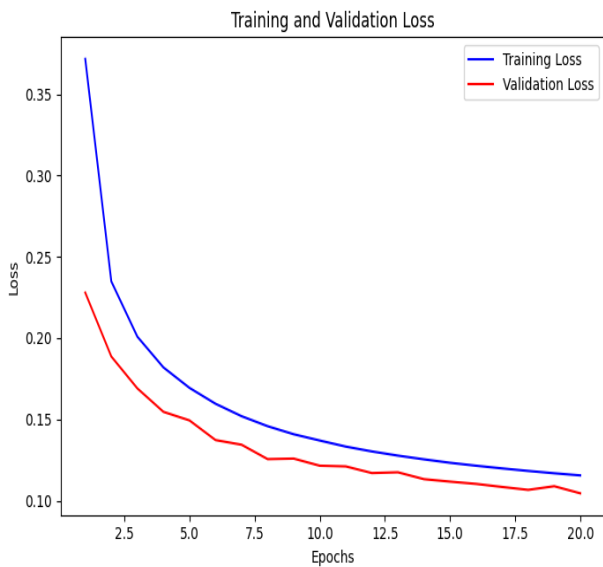


Fig. 4 Training and Validation Loss

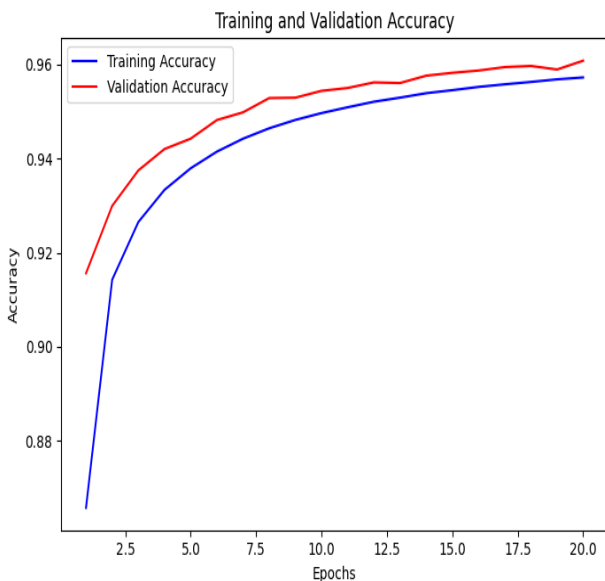


Fig. 5. Training and Validation Accuracy

3.3 Critical Analysis of Features

We employed an intensive study of the feature space to justify the fact that the predictions of the model are based on meaningful physiological representations or not. This was necessary to measure the quality of data preprocessing of the model, integrity of class distributions, and the discriminative capacity of the extracted features of the dataset as well to support our objective of constructing a robust and generalized framework of stress detection.

We also employ the bar chart that shows in the Figure 6 Label Distribution the count of each emotional state that is relevant to the stress that a subject is experiencing in each dataset and each label symbolizes a particular state of mind or emotion of a subject. The number of samples is in millions (1e6) as the vertical axis and the numbers of the labels (0-7) are the horizontal axis. All the bars have color codes that represent the different levels of emotions or levels of stress, and the descriptions are on the right-hand side of the graph.

Label 0 is the blue bar, which corresponds to the Calm state and includes the greatest number of samples of 2.8 million and above, which is to say that the data are also heavily biased towards the emotional state of calm. Label 1, orange, represents a stress label that has about 1.2 million samples. Label 2 is in green, which refers to a Relaxed state, and has approximately 0.9 million samples. The term High Stress has a label 3 that is represented in red and has a much lower number of samples of approximately 0.4 million, which means that the states are less noted or observed.

Label 4, which is "Moderate Stress", is reflected by the purple bar and has quite a significant number of samples of nearly 0.9 million. Label 5 of brown color is the category of the label No Stress, and the number of samples decreases noticeably to about 0.1 million. Likewise, label 6 is pink in color and it represents the "Very Low Stress" condition as well containing approximately 0.1 million samples. Lastly, label 7, which is gray colored is used to denote mixed or ambiguous states that are defined as Sometimes Low Stress and Sometimes No Stress and this category has a sample number of approximately 0.1 million as well.

In general, this sample distribution of our model indicates a vast imbalance of the classes, where most of our samples occupy the Calm class, and the number of samples is more in the high or mixed stress classes. This lack of balance may be manipulative to any machine learning model that is trained on this data and thereby more prone to predict the most common classes such as the "Calm" one and undermine less common classes such as the High Stress and Very Low

Stress classes. This imbalance would have to be addressed with some methods like resampling, data augmentation, or weight loss functions so that it could be accurately and fairly classified within all the emotional labels.

To enhance uniformity of signals among heterogeneous multimodality of the bio signal, we have used feature standardization as indicated in Figure 7. This was done by z-score normalization of the input features to a normal range. Through this, we managed to ensure that differences in amplitude within modalities like Electromyogram, electroencephalogram, electrodermal activity and electrocardiogram did not have a disproportionate effect on the learning process. This part played a major role in the successful combination of multimodal bio signals, enabling the hybrid deep learning CNN-Transformer model to be trained on all modalities on a similar scale.

PCA on the extracted features, which are represented as Features Analysis in Figure 8. As a dimensionality reduction algorithm, PCA was used to map the high-dimensional feature space to the two-dimensional space that is easily visualized. The ensuing 2D scatter plot depicts the distribution and segregation of the various levels of stresses in accordance with the learned features of the data set. The plot exhibits distinct and well-separated clusters, which means that the model has succeeded in capturing and distinguishing the patterns of relevancy that are associated with different levels of stress levels. The fact that these graphs visualize supports the learning of discriminative features of the model, thus playing a part in the correct classification of the stress detecting classes. These experiments confirm the usefulness of the preprocessing pipeline as well as the discriminative onfirm that the model can learn meaningful physiological patterns hence meet the prerequisite that a proper and broad generalized mental stress detection model needs.

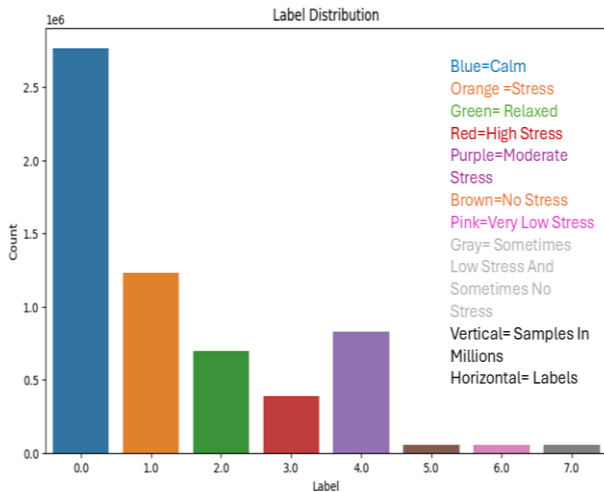


Fig. 6. Label Distribution of Bio signals

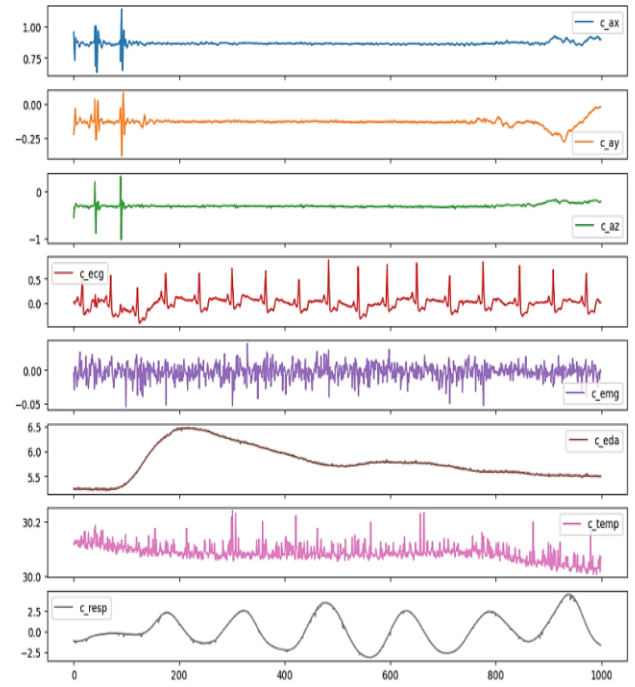


Fig. 7. Feature Preprocessability of the features produced by the model. They also cssing Standardization

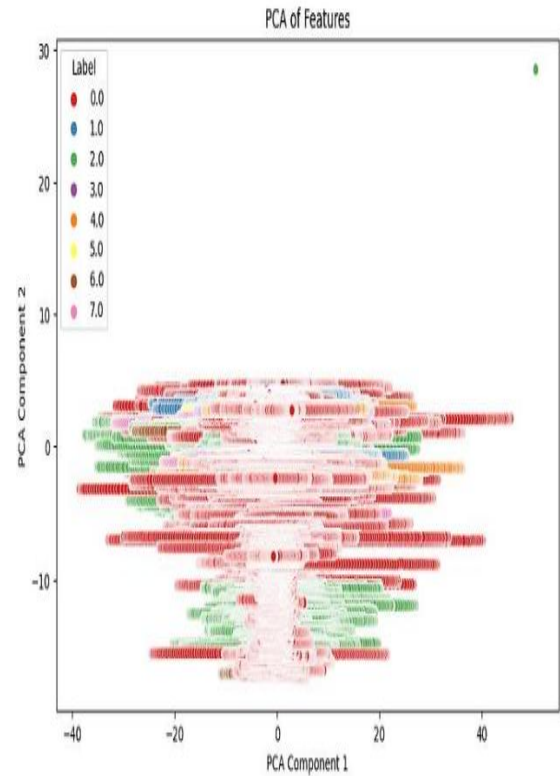


Fig. 8. PCA of Features of Multimodality

3.4 Benchmark Validation

Our model achieved notable accuracy, which is 96%, as shown in Figure 9, as compared to the other SOTA techniques. Our model's result shows the model's superior capability in capturing both local patterns convolutional neural network and long-range dependencies Transformer with multimodality of bio signals. The effective fusion of multimodal bio signals along with sensor data Temperature and respiration further contributed to the model's high performance.

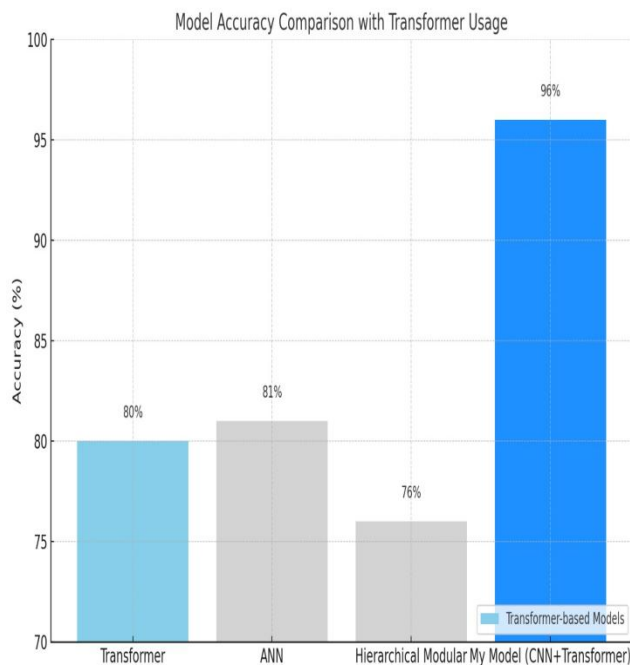


Fig. 9. Comparative Analysis with SOTA Approaches

3.5 Comparative Analysis

In comparative analysis we compare our model's accuracy across with many other deep learning and machine learning methods. The models evaluated include the single Transformer model, the Artificial Neural Network, the Hierarchical Modular model, and our hybrid CNN-Transformer based attention mechanism with multimodal bio signals.

To support the fact that the predictions of the model are made on meaningful physiological representations or not, we used an intensive study of the feature space. This was needed to gauge the data preprocessing of the model quality, integrity of class distributions and the discriminative capacity of the extracted features of data set as well to facilitate our goal of building a strong and genericized model of stress detection.

The bar chart we also use, and which is displayed in Figure 6 Label Distribution demonstrates the numbers of each emotional state which is applicable to the stress that a subject is undergoing in a specified dataset, and each label represents a specific state of mind or emotion of a subject. The size of the samples is millions (1e6) since the labels (0-7) are put as the horizontal axis and the numbers of the labels (0-7) are the vertical axis. The color codes of all the bars denote the various levels of feelings or level of stress, and the detail is on the right-hand side of the graph.

Label 0 is the blue bar, and it represents the Calm state and has the highest number of samples 2.8 million and above which means that the data also have intense bias to the emotional state of calm. Label 1, orange, is a stress label consisting of approximately 1.2 million samples. The green color on Label 2 is a Relaxed state with several about 0.9 million samples.

High Stress has a label 3 which is denoted in red and has far smaller number of samples of about 0.4 million meaning that the states are not so defined or observed.

The purple bar, representing "Moderate Stress" is the one with which almost half a million samples are almost equal. Label 5 brown color is the category of the label No Stress, the number of samples reduces significantly to approximately 0.1 million. Similarly, label 6 is pink color and it also corresponds to the very low stress condition and has a sample population of about 0.1 million. Finally, the gray-colored label 7 is to show the mixed or ambiguous states that are characterized as Sometimes Low Stress and Sometimes No Stress and this category has also the sample number of about 0.1 million.

This sample distribution of our model in general suggests a huge skew of the classes with most of our samples in the Calm classes and a larger sample in the high stress classes or mixed stress classes. Such a lack of balance can be manipulative to any machine learning model which is trained on this data and thus more subject to predict the most common classes like the Calm class and compromise less common classes like the High Stress and Very Low Stress classes. The imbalance would be required to be offset in some way such as resampling, data augmentation, weighted loss functions to be represented so that it could be faithfully and fairly classified between all the emotional terms.

To further standardize uniformity of signals amongst heterogeneous multimodality of the bio signal, we have adopted feature standardization as shown in Figure 7. This was achieved through normalization of the input features to normal range by z-scoring. It is through this that we were able to make sure that differences in amplitude within modalities such as Electromyogram, electroencephalogram, electrodermal activity and electrocardiogram did not disproportionately influence the learning process. This component significantly contributed to the successful integration of multimodal bio signals, where the hybrid deep learning CNN-Transformer model can be trained on a comparable scale on all modalities.

PCA of extracted features, the results of which are shown as Features Analysis in Figure 8. PCA as a dimensionality reduction method was employed to project the high-dimensional space of features to the two-dimensional space that can easily be visualized. The following 2D histogram shows how the different levels of stress are distributed and segregated according to the acquired characteristics of the data set. The plot shows clearly and clearly distinct clusters and that is why the model has been able to capture and distinguish the patterns of relevancy that are linked to

various levels of stress levels. The very fact that these graphs represent encourages the acquisition of discriminative characteristics of the model, hence contributing to the accurate categorization of the stress detecting classes. This experiment validates both the usefulness of the preprocessing pipeline and the discriminative power of the features generated by the model. They also affirm that the model can learn significant patterns of physiological patterns therefore satisfies the requirement that an appropriate and general physiological stress detection model requires.

Table 2

Compliance Report

Serial No:	Paper	Datase t	Accuracy	Modality
1	Behnam et at.,2021.	WES AD	80%	ECG
2	Bobad et al.,2020.	WES AD	81%	ECG, EDA
3	F.Waloch et al., 2022.	Self-created Simulator Data set	76%	ECG
4	Our Model	WES AD	96%	EEG, ECG, EMG, EDA, TEMP, RESP

3.5.1 Model Insights and Implications

The significantly higher accuracy of our hybrid multimodal bio signal CNN-Transformer model emphasizes the importance of using specialized techniques for complex, multimodal datasets such as physiological signals. The model's ability to integrate spatial and temporal information enables it to outperform other architectures that are either less specialized in handling sequential data or unable to leverage spatial feature extraction effectively of the models. These results suggest that a hybrid approach, combining the strengths of two deep learning models CNNs and Transformers, is well-suited for tasks involving multimodal bio signals, where both spatial and temporal patterns play a crucial role in accurate classification of the multi classes. This finding opens

the door for further optimization of hybrid models as well as dual nature of models in future work.

3.6 Model Evaluation

To assess the performance of our CNN-Transformer hybrid model with multimodal bio signals for mental stress detection, we utilized a confusion matrix to evaluate the classification accuracy for each emotional state individually. As shown in Figure 10, the confusion matrix provides a detailed view of all features which show how well the model distinguishes between the different states: Baseline, Amusement, and Stress. The matrix compares the actual class labels to the true values on the vertical axis with the model's predicted outputs on the horizontal axis. In the confusion matrix diagonal cells represent correctly classified instances, while the off-diagonal cells indicate misclassifications. A high number of entries along the diagonal suggests that the model achieved excellent accuracy. Specifically, the model consistently predicted each emotional state with minimal confusion, demonstrating its robust discriminative capability and balanced performance. Confusion matrix reveals that the model distinctly classified the baseline state, minor misclassifications occurred between stress and amusement classes owing to overlapping physiological patterns of the signals, particularly in EDA and EEG signals during mild stress responses. This result confirms that the multimodal bio signal hybrid deep learning CNN-transformer method can capture both spatial and temporal patterns in multimodal physiological signals, leading to highly accurate and reliable mental stress detection.

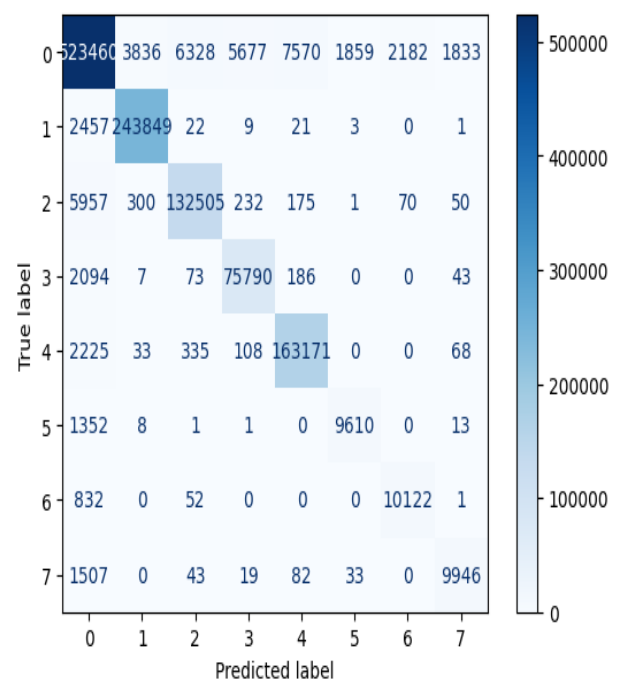


Fig. 10. Confusion Matrix illustrating Classification Performance

For the Evaluation we used metrics for our model, such as accuracy, precision, recall, and F1-score, to further assess the performance of our multimodal bio signal hybrid CNN-Transformer model. Figure 11's bar chart shows a comparison of the model's efficiency across the three classes Baseline, Amusement, and Stress. While accuracy evaluates overall correctness of the model, precision calculates the percentage of true positives among all predicted positives, recall checks the model's capacity to detect all true positive values, and the F1-score integrates precision and recall into a single performance of metric, each metric provides a unique role on the model's performance.

The uniformly high values observed across all four metrics confirm that the model not only achieves excellent performance but also maintains consistency across all three levels, avoiding bias toward any stress level. This consistent performance highlights and model's reliability.

	precision	recall	f1-score	support
0	0.97	0.95	0.96	552745
1	0.98	0.99	0.99	246362
2	0.95	0.95	0.95	139290
3	0.93	0.97	0.95	78193
4	0.95	0.98	0.97	165940
5	0.84	0.87	0.85	10985
6	0.82	0.92	0.87	11007
7	0.83	0.86	0.84	11630
accuracy			0.96	1216152
macro avg	0.91	0.94	0.92	1216152
weighted avg	0.96	0.96	0.96	1216152

Fig. 11. Classification Report

Here Table 3 shows an evaluation of the model.

Table 3

Performance Evaluation Metrics

Matric	Purpose	Results
Accuracy	Measures the overall correctness of predictions.	96%
Precision	Ensure stress predictions are reliable and minimize false alarms.	95%

Recall	Ensures all stress cases are detected, reducing missed cases.	97%
F 1 Score	Provides a balanced performance between precision and recall.	96%

4. Conclusion

We Are using multimodal bio signals and sensor data from the publicly accessible WESAD dataset, our work developed a hybrid CNN-Transformer-based deep learning model for the mental stress detection of three states: baseline, stress, and amusement. While the multi-head attention and layer normalization transformer was used to detect global temporal correlations, the CNN technique retrieved local spatial data. The model was trained using preprocessed biological signals and sensor data: EEG, EMG, ECG, EDA, respiration, temperature, and accelerometer data. Deep architecture feature extraction eliminated the need for human feature engineering by handling feature extraction automatically. Data normalization and reshaping were applied to optimize learning. After being trained over 20 epochs with the Adam optimizer and sparse categorical cross-entropy loss, the model's accuracy was 96%. It proved successful in distinguishing between stress states, particularly in correctly identifying the mental stress class, according to evaluation metrics such as precision, recall, F1-score, and confusion matrix. To sum up, our model's potential for identifying mental stress detection and contributing significantly to intelligent health monitoring systems and Our future work will focus on lightweight edge implementations for real-time stress monitoring systems through wearable sensors, larger multi-subject datasets, and inclusion of behavioral and contextual data to enhance model generalizability.

Acknowledgment

First, I thank Almighty Allah for His countless blessings and guidance. I would also like to express my heartfelt gratitude to the authors of the publicly available WESAD dataset, Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven, for their significant contributions to affective computing and human-computer interaction. Their meticulous efforts in collecting and making this invaluable data set available have supported my research.

Author's Contribution

- Author 1: Conceptualized the research idea, designed the experimental setup of the model,

and contributed to data analysis and manuscript writing.

- Author 2: Conceptualized the deep learning model and implemented the experiments, also analyzed the results.
- Author 3: Contributed to data collection, manuscript writing, and formatting.
- Author 4: Contributed to data collection and pre-processing.

Conflict of Interest

Regarding the publishing of this Research paper, we all authors have no conflicts of interest.

5. References

- [1] A. F. A. Lee and A. Roussos, "A privacy-preserved IoMT-based mental stress detection framework," 2024.
- [2] M. Abd Al-Alim, R. Mubarak, N. M. Salem, and I. Sadek, "A machine-learning approach for stress detection using wearable sensors in free-living environments," *Computers in Biology and Medicine*, vol. 179, p. 108918, 2024.
- [3] A. Abd-Alrazaq *et al.*, "The performance of wearable AI in detecting stress among students: A systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 26, p. e52622, 2024.
- [4] V. Adarsh and G. R. Gangadharan, "Mental stress detection from ultra-short heart rate variability using explainable graph convolutional network with network pruning and quantization," *Machine Learning*, vol. 113, no. 8, pp. 5467–5494, 2024.
- [5] Y. Badr *et al.*, "A review on evaluating mental stress by deep learning using EEG signals," *Neural Computing and Applications*, vol. 36, no. 21, pp. 12629–12654, 2024.
- [6] N. Bashir, S. Narejo, and B. Naz, "A machine learning framework for major depressive disorder (MDD) detection using non-invasive EEG signals," *Wireless Personal Communications*, 2023, doi: 10.1007/s11277-023-10445-w.
- [7] D. S. Benita, A. S. Ebenezer, L. Susmitha, M. S. P. Subathra, and S. J. Priya, "Stress detection using CNN on the WESAD dataset," in *Proc. 2024 Int. Conf. Emerging Systems and Intelligent Computing (ESIC)*, Feb. 2024, pp. 308–313.
- [8] M. Bolpagni, S. Pardini, M. Dianti, and S. Gabrielli, "Personalized stress detection using biosignals from wearables: A scoping review," *Sensors*, vol. 24, no. 10, p. 3221, 2024.
- [9] K. M. S. Boopathybalan, "Transformer-based deep learning model for mental health assessment," 2025.
- [10] A. Dasgupta and A. Basu, "Measurement and instrumentation for mental stress detection in humans," *IEEE Instrumentation & Measurement Magazine*, vol. 28, no. 2, pp. 10–17, 2025.
- [11] S. de Vries *et al.*, "Real-time stress detection based on artificial intelligence for people with an intellectual disability," 2024.
- [12] M. A. Hafeez and S. Shakil, "EEG-based stress identification and classification using deep learning," *Multimedia Tools and Applications*, vol. 83, no. 14, pp. 42703–42719, 2024.
- [13] M. Khandelwal and A. Sharma, "Machine learning and deep learning techniques to detect mental stress using various physiological signals: A critical insight," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 15, no. 3, p. e70035, 2025.
- [14] H. Lee, J. Kim, B. Han, S. M. Park, and J. Chang, "Developing an explainable deep neural network for stress detection using bio signals and human-engineered features," *Biomedical Signal Processing and Control*, vol. 109, p. 107960, 2025.
- [15] M. Naeem, S. A. Fawzi, H. Anwar, and A. S. Malek, "Wearable ECG systems for accurate mental stress detection: A scoping review," *Journal of Public Health*, vol. 33, no. 6, pp. 1181–1197, 2025.
- [16] N. Bashir, S. Narejo, and B. Naz, "EEG-based major depressive disorder (MDD) detection using machine learning," in *Proc. 5th Int. Conf. Pattern Recognition and Artificial Intelligence (MedPRAI 2021)*, Dec. 17–18, 2021.
- [17] A. Pandey and S. Kumar, "Mental health and stress prediction using NLP and transformer-based techniques," in *Proc. 2024 IEEE Symp. Wireless Technology & Applications (ISWTA)*, Jul. 2024, pp. 61–66.
- [18] M. Qorich and R. El Ouazzani, "Lightweight advanced deep-learning models for stress detection on social media," *Engineering Applications of Artificial Intelligence*, vol. 140, p. 109720, 2025.

- [19] R. K. Rai and D. K. Singh, "Stress detection through wearable EEG technology: A signal-based approach," *Computers and Electrical Engineering*, vol. 126, p. 110478, 2025.
- S. Hajano, B. Naz, S. Talpur, and S. Hajano, "Classification of amputee EMG signals using machine learning techniques," *International Journal of Innovations in Science & Technology*, Y category, 2023.
- [20] G. R. Reddy *et al.*, "Enhancing stress detection employing physiological signals from the WESAD dataset: A machine learning approach with SMOTE," *Milestone Transactions on Medical Technometrics*, vol. 2, no. 2, pp. 79–87, 2024.
- [21] R. Tanwar, O. C. Phukan, G. Singh, P. Pal, and S. Tiwari, "Attention-based hybrid deep learning model for wearable-based stress recognition," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107391, 2024.
- [22] G. Taskasaplidis, D. A. Fotiadis, and P. D. Bamidis, "Review of stress detection methods using wearable sensors," *IEEE Access*, vol. 12, pp. 38219–38246, 2024.
- [23] G. V. Tsakalidis *et al.*, "Advances, benefits, and challenges of wearable sensors for healthcare and stress management: A focus on hemodynamic parameters and cortisol measurement," *Engineering Proceedings*, vol. 82, no. 1, p. 42, 2024.
- [24] X. Yang and L. Yu, "Intelligent detection of cognitive stress in subway train operators using multimodal electrophysiological and behavioral signals," *Symmetry*, vol. 17, no. 8, p. 1298, 2025.