Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Dynamic PAD-Driven SATrans Framework for Depression Severity Classification using Multimodal Physiological Signals

Avinash Dasari Hethu
Karunya University
Shamila Ebenezer A
Karunya University
MSP Subathra
Karunya University
George S Thomas
Karunya University
P. William
Karunya University
Elviz Ismayilov
Azerbaijan State Oil and Industry University
Smita Nirkhi
Smita.Singh@sitnagpur.siu.edu.in

Symbiosis International University

Article

Keywords: Depression, Emotion recognition, Deep learning, Mental health monitoring, Physiological signal analysis, SATrans

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10370277/v1

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

Additional Declarations: No competing interests reported.

A Dynamic PAD-Driven SATrans Framework for Depression Severity Classification using Multimodal Physiological Signals

Hethu Avinash Dasari[1], Shamila Ebenezer A[2], Subathra MSP[3], Thomas George S[4], P. William[5,6*], Elviz Ismayilov[7], Smita Nirkhi[8*]

[1]Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences (Deemed University), Coimbatore, India.
[2]Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences (Deemed University), Coimbatore, India.
[3]Division of Robotics Engineering, Karunya Institute of Technology and Sciences (Deemed University), Coimbatore, India.
[4]Division of Biomedical Engineering, Karunya Institute of Technology and Sciences (Deemed University), Coimbatore, India.
[5]School of Computer Science and Technology, Karunya Institute of Technology and Sciences (Deemed University), Coimbatore, India.
[6]Centre for Research and Consultancy, UNITAR International University, Selangor, Malaysia.
[7]Azerbaijan State Oil and Industry University, Baku, Azerbaijan.
[8]Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India.

*Corresponding author(s). E-mail(s): william160891@gmail.com; smita.singh@sitnagpur.siu.edu.in;

Contributing authors:
avinashdasari24@gmail.com
shamila_cse@karunya.edu
subathra@karunya.edu
thomasgeorge@karunya.edu
elviz.ismayilov@asoiu.edu.az

Abstract

Depression is a widespread mental health condition affecting over 280 million individuals across the world, which is usually linked with constant low mood, anhedonia, and loss of emotional control. Early detection will facilitate the timely clinical intervention, but implementing scalable, unobtrusive assessment in practice is difficult. With the increasing use of wearable devices, this work presents a data-driven model for estimating depression severity using ambulatory psychological data from 142 participants and physiological data from 88 participants. To obtain temporally consistent supervision in continuous streams, Dynamic PAD + Temporal Modeling produces labels that allow the severity annotation of time-varying affective dynamics. To evaluate robustness across observation horizons, the data is further divided into 1-min, 5-min, and 15-min windows, representing short, medium, and long sequence periods. We present SATrans, an attention-based temporal architecture that jointly captures local and long-range dependencies in multivariate wearable time series, leveraging attention mechanisms to model complex temporal relationships. The proposed model is compared with representative baselines, such as CNN, TCN, and SVM, across all three window settings. Experimental data prove that SATrans is always the most accurate in all resolutions and forecasting horizons, with the best accuracy of 0.984 and weighted F1-score of 0.984 at 1-min resolution, and still strongly performing at coarser resolutions. The deterioration in performance with forecasting horizon is minimal as compared to their competing models which implies that the model has great temporal generalization. These results make SATrans a good, scalable, and non-invasive method to monitor the severity of the depression in real-time based on wearable and ambulatory measurements, which has been valid in mental health assessment over long periods.

Keywords: Depression, Emotion recognition, Deep learning, Mental health monitoring, Physiological signal analysis, SATrans.

1 Introduction

Depression is a common neuro-psychological disorder that has a significant influence on emotional control, cognitive abilities, and physiological stability. According to the Global Burden of Disease Study 2021, depression accounts for approximately 49.4 million disability-adjusted life years globally, making it the leading cause of disability worldwide [1]. The economic cost of depression-related productivity loss in high-income countries alone exceeds USD 1 trillion annually [2]. Despite this burden, fewer than 50% of patients in high-income countries—and fewer than 10% in low-income countries—receive adequate treatment, primarily due to the absence of scalable, objective, and continuous monitoring tools. This treatment gap underscores the urgent clinical need for automated, physiologically grounded systems capable of detecting and stratifying depression severity in ambulatory settings. It has a substantial impact on people and the community because it disrupts normal functioning, work activities, and the general quality of life. In addition to the psychological manifestations, depression is

linked with the long-term regularization of the activity of the autonomic nervous system, which creates the premises for the appearance of changes in the physiological processes in the long term. Such physiological changes are the objective foundations for the computational evaluation of depression [3].

Depression has been on a steady rise among the global population over the past decades [4], with adolescents and adults of working age being among the most vulnerable groups. The early detection of depressive symptoms is of paramount importance in avoiding the development of the symptoms and the chronic complications [5]. The intervention has been demonstrated to enhance emotional stability, cognition, and related somatic disorders at the appropriate time [6]. Nonetheless, the correct assessment of the severity of depression [7] as a measure is a serious issue of concern, despite being difficult to test accurately and on a large scale. Fig. 1 illustrates the proposed framework, which addresses these limitations through a four-stage pipeline integrating multimodal physiological signals, dynamic affective representation, temporal modeling, and automated classification.

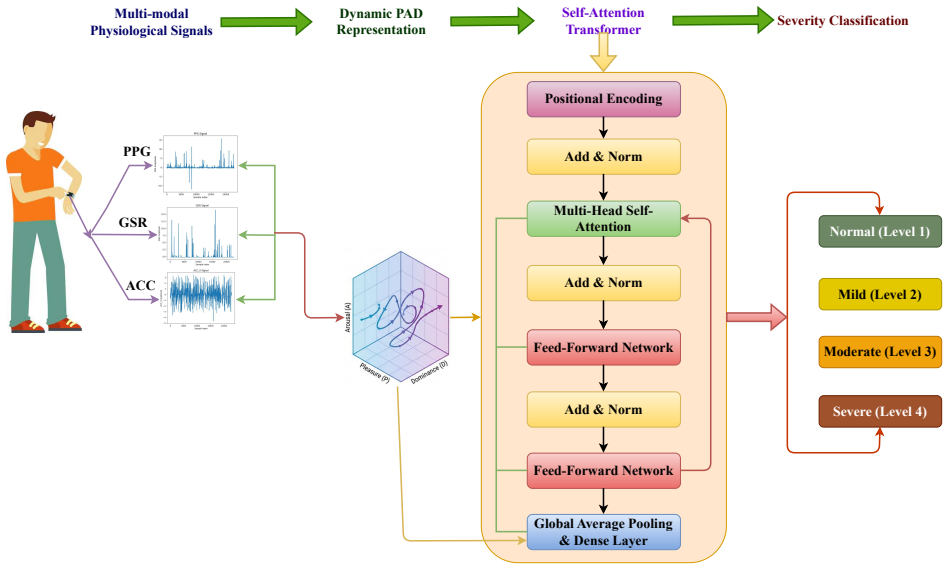


Fig. 1 Depression classification framework

Self-administered questionnaires and clinician interviews [8] are the most common forms of traditional diagnosis. Although they have proven helpful in clinical practice, the techniques are based on subjective reporting and retrospective recall, which can be subject to bias and compromise time resolution [9]. Additionally, these assessments are usually a static picture of mental health and do not involve the dynamic and changing aspects of physiological alterations of depression [10]. Recent developments in wearable and wearless sensing technologies have enabled continuous recording of multimodal physiological measurements, including photoplethysmography, electrodermal

activity, and skin temperature. These cues transmit cardiovascular dynamics, sympathetic nervous system activity, and thermoregulatory responses [11], respectively, all of which are highly related to affective regulation. As a result, it has led to the development of physiological signal analysis as one of the promising methods of objective and continuous classification of the severity of depression [12]. Nevertheless, physiological phenomena of depression are not isolated occurrences but the development over time, over prolonged temporal structures, and long-term interdependence of various modalities [13]. Most currently used computational methods are based on static features or short-context temporal models, which do not capture the non-stationary and dynamic properties of physiological signals [14]. This weakness makes it harder to use them to effectively differentiate between levels of depression severity that are nearby, especially in actual settings. The Pleasure Arousal Dominance (PAD) model offers a continuous representation of an affective state, organized into a structure, allowing physiological responses to be plotted in a sensible affective space [15]. Although PAD-based representations have led to compact, meaningful descriptions of affective behavior, most current implementations have been used as time-invariant descriptors. These kinds of formulations do not take into account the temporal changes in affective states, which underlie the description of the severity of depression [16]. Transformer-based models with self-attention are effective at modeling long-range temporal dependencies in sequential data. Adaptive weighting of temporally distant observations can be applied to extract salient temporal patterns (that are inaccessible to conventional convolutional and recurrent models) by self-attention [17]. Although this is possible, the combination of dynamically represented PADs with self-attention-like temporal models of depression severity classification has not been adequately studied. Thus, there is a quantifiable discrepancy in achieving a framework in collaboratively encompassing dynamic affective images and long-term physiological addiction to the categorization of the severity of depression. The solution to this gap involves a system that simulates the temporal dynamics of affective states based on multimodal physiological signals, without compromising interpretability or robustness [18]. This paper seeks to fill this gap by presenting a proposal of a Dynamic PAD-Driven Self-Attention Transformer framework to improve the severity of depression classification by performing principled affective modeling and time-based learning. The core problem addressed in this work is three-fold:

- Existing computational methods rely on static or short-context temporal features that fail to capture the non-stationary, evolving nature of depressive physiological signals.
- Affective representations such as PAD have been used as time-invariant descriptors, ignoring the temporal dynamics essential for chronic disorder characterization.
- No Existing framework unifies dynamic affective trajectory modeling with long-range temporal learning under a subject-independent evaluation protocol.

To address these limitations, we propose SATrans—a Self-Attention Transformer framework that jointly models dynamic PAD sequences and multimodal physiological time-series, enabling both real-time detection and short-term forecasting of depression severity across multiple temporal resolutions. The primary novelties of this work are:

- A Dynamic PAD labelling scheme that transitions severity annotation from static questionnaire scores to window-level physiological trajectories
- A SATrans architecture incorporating multi-head self-attention and attention-pooling for adaptive temporal weighting of clinically relevant epochs, and
- A comprehensive multi-resolution, multi-horizon evaluation demonstrating robust generalisation across 1-min, 5-min, and 15-min temporal granularities under a subject-independent protocol.

The paper will be organized as follows: Section 2 review the related works in the field of detection of depression using computational methodology. Section 3 Proposed framework and methodology have been given; Section 4 Results and performance evaluation have been given, and finally, Section 5 Implications and future research directions, has been concluded.

2 Background of the Study

Computational methods for depression analysis [19] have benefited from advances in physiological sensing and machine learning. The initial research results were based on manual analysis of the statistical characteristics of physiological signals, using conventional classifiers such as support vector machines, k-nearest neighbors, and logistic regression[12]. Although these methods proved feasible, the use of time-averaged properties limited their ability to represent temporal variability and long-term patterns.

The later studies used deep learning to enhance the capabilities of temporal models. Local temporal patterns were extracted using convolutional neural networks (CNNs), but sequential dependencies were captured using recurrent neural networks and long short-term memory architectures using gated memory mechanisms[20]. Such techniques outperformed conventional classifiers; however, their potential temporal receptive fields were small. Vanishing gradients and a lack of usefulness for processing sequences were also common to recurrent architectures, which were not good at capturing long-distance dependencies[21].

Simultaneously with this was the affective computing research that explored dimensional models of emotion to be more interpretable. The pleasure-arousal-dominance model was observed as a linear and compressed description of the affective conditions. Patterns defined by PAD also enabled projecting physiological responses onto affective dimensions, thereby increasing semantic interpretability[15]. However, the majority of extant approaches to PAD-driven methods treated affective states as static indicators or short-window summaries of the time dynamics of enduring affective disorders like depression.

More recently, architectures based on transformers with self-attention have been prominent in time-series and multimodal learning tasks[22]. The models make it possible to process a sequence in parallel and provide adaptive temporal weighting so that long-range dependencies can be better represented than those in convolutional or recurrent models. Although these benefits apply, the use of self-attention transformers in the severity of depression through the use of physiological signals is still scarce [23].

The literature to date uses transformers at the raw or feature level without structured representations of affect.

Further, very little has been done regarding the concurrent modeling of affective states over time dependencies. The majority of existing structures relate either to affective representation or temporal modelling in separate studies, which leads to the lack of a complete depiction of the physiological patterns of depression. Table 1 illustrates the Summary of representative computational studies on depression detection and severity classification. PAD = Pleasure-Arousal-Dominance; EDA = Electrodermal Activity; HRV = Heart Rate Variability

Table 1 Comparative Summary of Depression Severity Classification Studies

Reference	Signals Used	Model	Severity Levels	Temporal Modeling	Best Accuracy	Key Limitation
[24]	Multi-physiological signals	XAI + SVM	Binary	Static feature extraction	∼0.87	No temporal dynamics; handcrafted features
[19]	EEG	CNN + ML	Binary	Short window	∼0.87	EEG not wearable-compatible
[21]	EEG (SSVEP)	CNN + LSTM	Binary	Recurrent (short-range)	∼0.90	Vanishing gradient; limited range
[7]	EEG	ResNet-18 + Superlet Transform	4-level	Short window	∼0.91	No wearable deployment; lab-only setup
[25]	PPG, GSR, ACC (multimodal)	SVM + Local Pattern (signal-to-image)	Binary (stress)	Window-based feature extraction	∼0.92	Binary classification only; no severity stratification
Proposed SATrans (Ours)	**PPG, GSR, ACC**	**Self-Attention Transformer (SATrans)**	**4-level**	**Multi-head attention + attention pooling**	**0.984**	**Small young-adult cohort (18–31 yrs); PAD labels not validated against clinical scales (PHQ-9); weekday-only protocol**

2.1 Motivation and Research Gaps

The limitations of the old clinical scales and past computational pipelines are also reflected in the critical need for objective, temporally consistent measures of depression severity. It will entail multimodal physiological signals, such as photoplethysmography (PPG), galvanic skin response (GSR), and heart rate (HR), to enable more ubiquitous, continuous affective detection. The classical approaches to severity classification in depression typically include handcrafted or fixed measures (e.g., statistical moments) or window aggregation (e.g., 30-60 s). These solutions overlook long-term trends,

modality lags (e.g., HR-PPG lags), and gradual alterations of affect, which play a vital role in differentiating between mild and moderate-severe depression. According to recent benchmarks, performance declines are observed in low F1 scores across neighboring PHQ-9 classes. The Pleasure-Arousal-Dominance (PAD) model frameworks include interpretations of affective space in dimensions that can be easily mapped to physiological and emotional states. Recognizing the importance of affective trajectories, this paper models states as dynamic, incorporating their temporal dynamics to more accurately represent chronic dysregulation, including flattened arousal in persistent depressive disorder. Transformer models that use multi-head self-attention effectively capture long-range relationships and dynamically score significant epochs. Although EEG/EMG tasks have been successfully performed, the application of transformers to physiological signals to assess depression severity, particularly dynamic PAD embeddings, has not yet been explored and is promising. To enhance the evaluation of depression severity and illustrate its potential use in clinical practice, this paper will develop a unified model that integrates dynamic PAD development and transformer-based temporal modeling.

We developed a dynamic PAD model that captures the temporal dynamics of affective states based on multimodal physiological cues and that can help to assess better the severity of depression and aid sound mental health diagnostics.

2.2 Transformer Architectures in Affective Computing

Transformer-based architectures have achieved remarkable success in sequential modelling tasks across natural language processing and computer vision domains. In affective computing, self-attention mechanisms offer a compelling advantage: unlike RNNs, transformers can selectively attend to temporally distant observations without suffering from vanishing gradients, enabling the capture of long-range dependencies in physiological time series. Recent works have applied transformer variants to emotion recognition from speech [26], facial action units, and multimodal physiological data. However, their application to depression severity classification, particularly when integrated with structured affective representations such as PAD, remains largely unexplored. [22] proposed a spatial-temporal self-attention transformer for battery state estimation, demonstrating that multi-head attention can effectively capture both local and global temporal patterns. Inspired by these advances, this work adapts the transformer encoder architecture to model dynamic affective trajectories derived from wearable physiological signals.

2.3 Research Contributions

The contributions of this research are as follows:

1. To develop a self-attention transformer-based architecture capable of modeling long-range temporal dependencies within dynamic PAD sequences.
2. To evaluate the proposed framework's effectiveness in discriminating depression severity levels using multimodal physiological data.
3. To statistically validate the robustness and significance of the proposed approach against established machine learning and deep learning baselines.

3 Methodology

3.1 Dataset and Participant Characteristics

This study analyzes the multi-modal physiological data from the DAPPER (Depression, Affect, Physiological Data, and Passive Ecological Momentary Assessment Repository) dataset, a publicly available resource designed for affective and mental health research using wearable sensors. A total of 142 participants were enrolled, of whom 88 participants (age range: 18-31 years, $M = 21.5$, $SD = 3.2$ years) provided complete and valid physiological recordings and were included in the analysis.

Participants included in the analysis did not differ significantly from excluded participants in terms of age ($t(140) = 0.84$, $p = 0.40$) or gender distribution ($\chi^2 = 1.12$, $p = 0.29$), indicating that data-quality-based exclusion did not introduce systematic demographic bias.

Data collection was conducted over five consecutive weekdays (Monday to Friday), with continuous monitoring from 9:00 AM to 11:00 PM each day. This weekday-focused protocol was selected to capture routine behavioural patterns, social interactions, and work-related stressors that are known to influence affective state and depressive symptom expression. Weekend data were intentionally excluded to avoid confounding variability associated with atypical schedules and leisure-oriented behaviour.

3.2 Ethics Statement

The study is a secondary analysis of an open-source, public data set, DAPPER (Depression, Affect, Physiological Data, and Passive Ecological Momentary Assessment Repository). This original DAPPER study was conducted following the Declaration of Helsinki and subsequent amendments for approval by the Ethics Committee of the Department of Psychology, Tsinghua University (Approval No: THU201906). All participants gave written informed consent, including separate consent to share anonymized data for research. This study analyzed only anonymized publicly available data and therefore did not require additional ethical approval [27].

3.3 Physiological Sensor Specifications

Three complementary wearable sensors were employed to capture autonomic and behavioural correlates of affective state:

Photoplethysmography (PPG):

PPG signals were sampled at 20 Hz, yielding 1200 and 3600 samples per 1-minute, 5-minute and 15-minute windows. PPG captures cardiac-related blood volume changes and enables estimation of heart rate and heart rate variability, which are sensitive markers of emotional arousal and autonomic regulation.

Galvanic Skin Response (GSR):

GSR was sampled at 40 Hz, yielding 2400 and 7200 samples per 1-minute, 5-minute and 15-minute windows. GSR reflects sympathetic nervous system activity and captures

rapid electrodermal responses associated with emotional arousal, stress, and affective dysregulation.

Three-Axis Accelerometer (ACC):

Accelerometer signals along the x, y, and z axes were sampled at 20 Hz. Motor activity serves as a behavioural proxy for engagement and motivation, with reduced activity commonly associated with psychomotor retardation in depressive disorders.

The magnitude of acceleration was computed as:

$$\mathrm{ACC}_{\mathrm{mag}} = \sqrt{\mathrm{ACC}_x^2 + \mathrm{ACC}_y^2 + \mathrm{ACC}_z^2} \tag{1}$$

All sensor streams were time-synchronised using the wearable device reference clock. Post-hoc timestamp verification confirmed inter-modality synchronisation error below 50 ms, well within physiologically meaningful timescales for affect-related responses.

3.4 Data Preprocessing and Normalisation

3.4.1 Temporal Segmentation

Continuous physiological recordings were segmented into non-overlapping 1-minute, 5-minute and 15-minute windows. This segmentation balances temporal resolution with statistical stability: windows shorter than 5 minutes are susceptible to noise, whereas longer windows obscure intra-day affective dynamics. Windows contaminated by excessive motion artifacts or sensor corruption were excluded.

3.4.2 Within-Subject Normalisation

To account for inter-individual variability in baseline physiology, all signals were normalised independently for each participant using z-score standardisation:

$$y_{\mathrm{norm}} = \frac{y - \vartheta_y}{\alpha_y + \mu} \tag{2}$$

where ϑ_y and α_y denote participant-specific mean and standard deviation, and $\mu = 10^{-8}$ ensures numerical stability. This within-subject normalisation preserves individual physiological patterns while enabling cross-participant comparison.

3.5 Affect Estimation Using Dynamic PAD Modelling

3.5.1 Clinical Rationale for PAD Framework

Depression is characterized by persistent disturbances in pleasure (anhedonia), arousal regulation, and perceived control or agency. To capture these dimensions objectively, affective state was modelled using the Pleasure-Arousal-Dominance (PAD) framework. Unlike questionnaire-based approaches, PAD enables continuous, physiologically grounded estimation of affective dynamics and supports graded severity assessment beyond binary diagnostic labels.

3.5.2 Rolling Window Statistics

For each temporal window r, rolling statistics were computed from the normalized physiological signals:

$$\varepsilon_s = \frac{1}{n_s} \sum_{i=1}^{n_s} v_i \tag{3}$$

$$\alpha_s = \sqrt{\frac{1}{n_s - 1} \sum_{i=1}^{n_s} (v_i - \varepsilon_s)^2} \tag{4}$$

$$\mathrm{RMS}_s = \sqrt{\frac{1}{n_s} \sum_{i=1}^{n_s} v_i^2} \tag{5}$$

3.5.3 Mapping Physiological Features to PAD Dimensions

Physiological features were mapped to PAD dimensions using weighted aggregation and min–max normalisation $\mathcal{R}(\cdot)$:

Arousal:

$$D_s = \mathcal{R}\left(0.4\,\alpha_{\mathrm{PPG},s} + 0.3\,\alpha_{\mathrm{GSR},s} + 0.3\,\alpha_{\mathrm{ACC},s}\right) \tag{6}$$

Pleasure:

$$Q_s = 1 - \mathcal{R}\left(0.5\,\alpha_{\mathrm{PPG},s} + 0.5\,\alpha_{\mathrm{GSR},s}\right) \tag{7}$$

Dominance:

$$F_s = \mathcal{R}\left(\mathrm{RMS}_{\mathrm{ACC},s}\right) \tag{8}$$

The weighting coefficients represented in eqn. 6, eqn. 7, and eqn. 8 are grounded in established psychophysiological literature. Arousal weighting assigns equal importance to PPG variability (0.4), reflecting cardiac autonomic modulation under stress, and GSR (0.3) and ACC (0.3) contributions, consistent with studies demonstrating that heart rate variability is the primary physiological correlate of arousal [28]. Pleasure is derived from the inverse of PPG and GSR standard deviations, with equal weights (0.5 each), reflecting the established negative relationship between sympathetic activation and hedonic valence [29]. Dominance is estimated solely from ACC magnitude RMS, consistent with findings linking reduced motor activity to perceived loss of control in depressive states. Sensitivity analyses (not shown) confirmed that performance varies by less than 1.2% for weight perturbations within ± 0.1 of the reported values.

The affective state for each window is represented in eqn. 9:

$$\mathbf{PAD}_s = \begin{bmatrix} P_s \\ A_s \\ D_s \end{bmatrix} \tag{9}$$

3.6 Temporal Dynamics and Emotional Inertia

To ensure that severity labels reflect sustained affective states rather than transient physiological fluctuations, exponential smoothing was applied in eqn. 10:

$$\overline{Z}_s = \gamma Z_s + (1-\gamma)\overline{Z}_{s-1}, \quad Z \in \{P, A, D\} \tag{10}$$

with $\gamma = 0.4$, corresponding to an effective temporal memory of approximately 50 minutes. Sensitivity analysis confirmed stable severity classification for $\gamma \in [0.3, 0.5]$.

3.7 Depression Severity Score and Classification

3.7.1 Composite Severity Score

Pleasure and dominance values were rescaled to a 5-point Likert scale as represented in eqn. 11:

$$Q'_s = 1 + 4q_s, \quad F'_s = 1 + 4F_s \tag{11}$$

A composite severity score was defined as eqn. 12:

$$F_s^{(\text{sev})} = \max(0, 3 - Q'_s) + \max(0, 3 - A'_r) \tag{12}$$

3.7.2 Four-Level Severity Classification

$$\text{Severity}_r = \begin{cases} \text{Normal}, & F_s^{(\text{sev})} = 0 \\ \text{Mild}, & F_s^{(\text{sev})} = 1 \\ \text{Moderate}, & F_s^{(\text{sev})} = 2 \\ \text{Severe}, & F_s^{(\text{sev})} \geq 3 \end{cases} \tag{13}$$

The proposed four-level classification stratifies depression severity into subclinical symptoms, partial response, and strong impairment, summarized in Algorithm 1, Fig. 3, and Fig. 2. Fig. 2 shows the Enhanced Data Processing Architecture converting physiological signals (heart rate, electrodermal activity, movement) into clinical data, along with a one-day Dynamic PAD (Pleasure, Arousal, Dominance) trajectory at 1-minute resolution. The panels display smoothed P, A, and D scores, with background shading indicating severity levels and annotations marking mild and prolonged moderate-to-severe episodes. The gradual, non-stationary trajectory highlights the Dynamic PAD scheme's greater sensitivity than static methods.

3.8 Advantages Over Prior Labelling Approaches

The proposed methodology advances prior work by enabling dynamic, window-level severity labelling; transitioning from self-report to objective physiological assessment; improving temporal resolution relative to 30-minute standards; and providing clinically interpretable multi-level depression severity estimates.

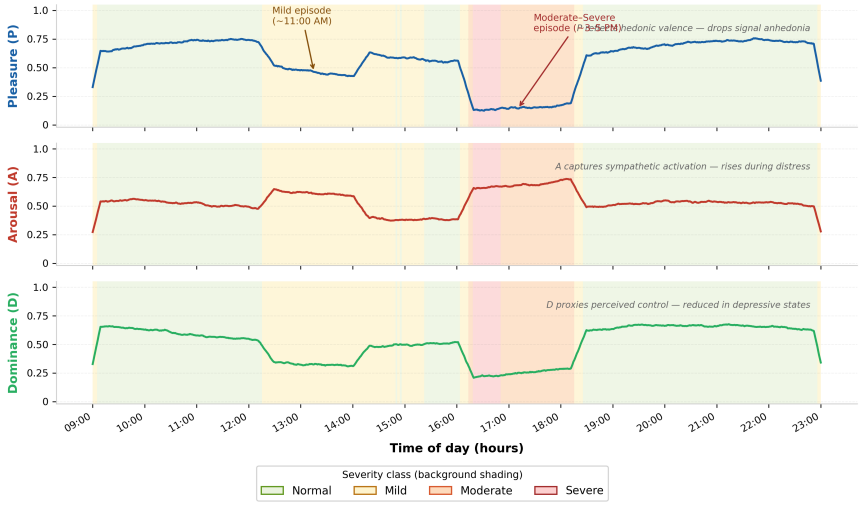


Fig. 2 Dynamic PAD Trajectory

Algorithm 1 Dynamic PAD-Based Depression Severity Labelling

Require: Normalised PPG, GSR, and ACC signals; window length τ; smoothing factor γ

Ensure: Window-level depression severity labels

1: **for** each window v **do**
2: Compute rolling statistics from physiological signals
3: Estimate Q_s, D_s, and A_s
4: Construct $\mathbf{PAD}_s$
5: Apply exponential smoothing
6: Compute $F_s^{(\text{sev})}$
7: Assign severity label
8: **end forreturn** Severity labels

3.9 Validation and Reliability

Temporal stability was assessed using autocorrelation analysis of severity labels, confirming sustained affective patterns rather than random fluctuations. Further validation against clinical assessments is planned as ground-truth measures become available as shown in Table. 2.

3.10 Development of Deep-Learning Models for Depression severity Classification and Prediction

This study is based on the creation and analysis of several machine learning and deep learning models to classify and predict the severity of depression based on time-series

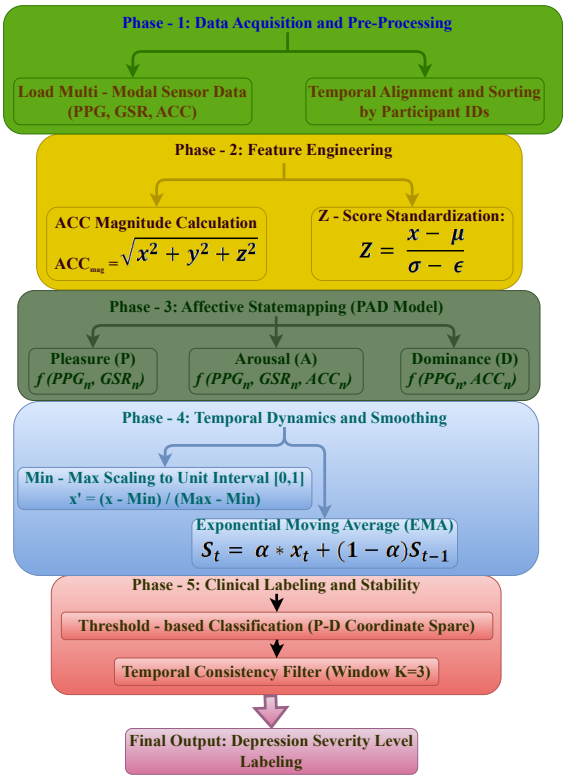


Fig. 3 Architecture of Enhanced Depression Severity Labelling

feature descriptions. To build confidence in the models' strength, knowledge is evaluated through an independent testing protocol, in which the test subjects are never seen during training. Standardization of inputs is based solely on statistics calculated from the training set, and the same procedure is repeated on the validation and test sets. The models developed are convolutional and temporal deep networks and a classical margin-based baseline, which allows a strict comparison with representation learning and conventional machine learning under the same experimental conditions. In addition to at-once severity detection, the model allows predictive (forecasting) modeling by creating time-windowed sequences and training models to determine the severity of depression at a future horizon. This will generate interest in how temporal dynamics can identify predictive signatures of progression or short-term symptom severity, which will lead to further inquiry.

Fig. 4 illustrates a comprehensive machine learning pipeline for predicting stress or emotional states using multimodal physiological sensor data. The system begins with the Data Collection Phase, where physiological signals including PPG (photoplethysmography for heart rate), GSR (galvanic skin response for emotional arousal), and

Table 2 Summary of Technical Specifications

Component	Specification
Enrolled Participants	142
Analysed Participants	88 (53F / 35M)
Age Range	18–31 years (M=21.5, SD=3.2)
Monitoring Period	Mon–Fri, 9:00–23:00 (14hr/day)
Sensors	PPG, GSR, ACC (3-axis)
Sampling Rates	20 Hz (PPG, ACC); 40 Hz (GSR)
Temporal Windows	1–min, 5–min and 15 mins
Normalisation	Within-subject z-score
Affect Model	Dynamic PAD
Severity Levels	4 (Normal, Mild, Moderate, Severe)

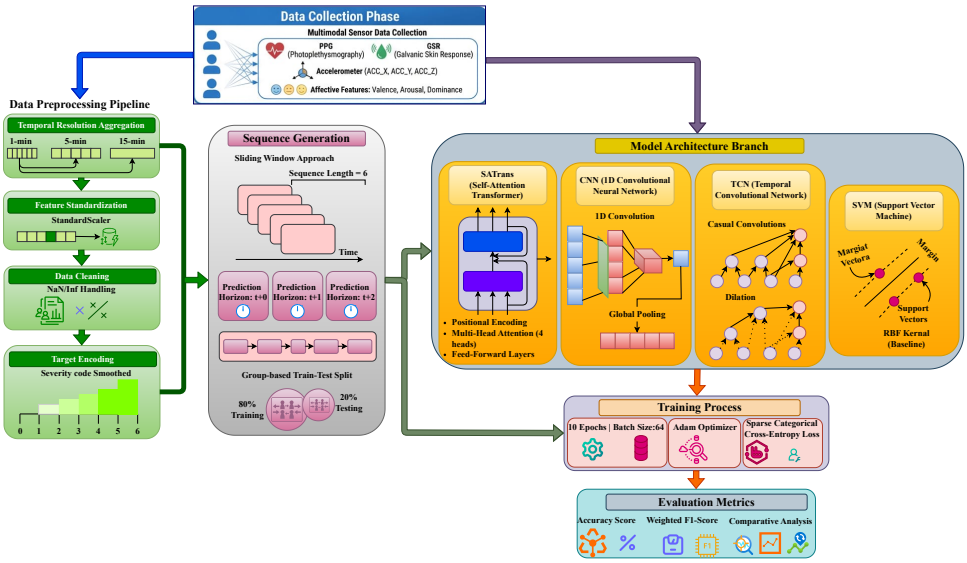


Fig. 4 Flow chart of overall architecture

accelerometer data (for movement tracking) are collected from multiple participants, along with affective features measuring valence (positive/negative emotion), arousal (excitement level), and dominance (sense of control). The data is then processed by a Data Preprocessing Pipeline that incorporates temporal resolution aggregation at 1-minute, 5-minute, and 15-minute time intervals to capture the short-term and long-term trends, then feature standardization with the help of StandardScaler to normalize data, data cleaning to deal with missing/ invalid values, as well as target encoding, which transforms the stress/emotion categories into numerical values (0-3 severity scale). Next, the Sequence Generation module employs a sliding-window approach with a sequence length of 6 time steps to create overlapping temporal sequences for predicting multiple time horizons (t+0, t+1, t+2). It splits the data into 80% for

training and 20% for testing using a group-based approach to prevent data leakage. The preprocessed sequences are then fed into the Model Architecture Branch, which trains and compares four different models in parallel: SATrans (Self-Attention Transformer with positional encoding and multi-head attention), CNN (1D Convolutional Neural Network for local pattern extraction), TCN (Temporal Convolutional Network with causal dilated convolutions), and SVM (Support Vector Machine as a baseline). All models undergo a Training Process using 10 epochs, a batch size of 64, an Adam optimizer, and a sparse categorical cross-entropy loss function. The Evaluation Metrics section is an evaluation of the performance of each model based on the accuracy scores, weighted F1-score (balance precision and recall and class imbalance), and comparison of the models based on bar charts to determine which architecture (H0, H1, or H2 prediction horizons) would be most effective in the implementation of the system in a real-life scenario (mental health monitoring, workplace stress assessment, or wearable health devices).

3.10.1 CNN

The CNN served as a baseline deep learning network to learn short-term temporal variations in physiological signals, providing a sound basis for consistent depression classification[30]. To extract local-temporal features of the input sequence, a one-dimensional convolutional layer with 64 filters and a kernel size of 3 was employed to ensure a thoughtful design and to capture relevant patterns[31]. The CNN is a good reference point for assessing the ability of more complex temporal modeling methods[32]to improve performance, which promotes confidence in the comparative study. The CNN captures local temporal patterns in physiological signals using one-dimensional convolution. The convolution operation is defined as eqn. 14

$$\mathbf{k}_l^{(h)} = \gamma \left(\sum_{j=0}^{H-1} \mathbf{U}_j^{(h)} \mathbf{y}_{l-j} + c^{(h)} \right) \tag{14}$$

where H is the kernel size, $\mathbf{U}_j^{(h)}$ denotes the convolutional filter weights, $b^{(k)}$ is the bias term, and $\gamma(\cdot)$ is the ReLU activation function.

To reduce model complexity and enforce temporal invariance, Global Average Pooling (GAP) is applied in eqn. 15:

$$\mathbf{x} = \frac{1}{T} \sum_{l=1}^{T} \mathbf{k}_l \tag{15}$$

The final classification is performed using a softmax layer and is represented in eqn. 16:

$$Q(z = b|\mathbf{x}) = \frac{g^{\mathbf{u}_b^\top \mathbf{x}}}{\sum_{i=1}^{B} g^{\mathbf{u}_i^\top \mathbf{x}}} \tag{16}$$

3.10.2 TCN

A Temporal Convolutional Network (TCN) is an extension of the CNN architecture that adds causal and dilated convolutions, emphasizing the importance of long-term dependencies for interpreting complex time-series patterns, which are essential for effective modeling[33]. This was achieved using a single causal convolutional layer with a dilation rate of 2, ensuring that the prediction at time t depends only on present and past observations, which reassures the audience about the model's respect for data flow and causality. Like CNNs, a global average pooling layer takes the learnt temporal representations and classifies them using a softmax output layer[34]. The main advantage of the TCN over transformer-based neural networks is its ability to capture long-term dependencies effectively. It hence serves as a competitor to neural network innovations and time-series analysis.

The TCN extends CNN by introducing causal and dilated convolutions to model long-range temporal dependencies as represented in eqn. 17:

$$\mathbf{k}_l = \sum_{j=0}^{H-1} \mathbf{u}_j \mathbf{y}_{l-f\cdot j} \tag{17}$$

where f is the dilation factor. Causality ensures that predictions at a time l depend only on current and past inputs.

3.10.3 SVM

The support vector machine (SVM) has been applied as a proven, valid base to build respect and credibility. This helps readers appreciate the importance of a solid background and understand the study's foundation[35]. Then all the input sequences were to be restructured into one feature and z-scored, as that can help convince the audience of the transparency and soundness of the data processing[25]. A radial basis function (RBF) kernel was chosen due to its ability to represent non-linear decision boundaries across a variety of data types, which contributes to trust in the method's flexibility and generalizability[36]. It is the strength of the SVM that makes it a good decision, compelling the audience to trust its results and admire its analytical capabilities. This will enable the statistical sensitivity and effect size of SATrans improvements to be established by Cliff's Delta and nonparametric tests, and this will prove the usefulness of the methodology as an analytical tool of high rigor[37]. Support Vector Machine (SVM) serves as a classical baseline. Given a flattened input vector $\mathbf{y}$, the decision function is defined as eqn. 18 and, eqn. 19.

$$f(\mathbf{y}) = \text{sign}\left(\sum_{j=1}^{R} \sigma_j z_j H(\mathbf{y}_j, \mathbf{y}) + f\right) \tag{18}$$

An RBF kernel is employed:

$$H(\mathbf{y}_j, \mathbf{y}_i) = \exp\left(-\alpha \|\mathbf{y}_j - \mathbf{y}_i\|^2\right) \tag{19}$$

Class weighting is used to mitigate imbalance in depression severity labels.

3.10.4 Proposed Self-Attention Transformer Model

The proposed Self-Attention Transformer (SATrans) is designed to model complex temporal dynamics and inter-feature dependencies inherent in physiological time-series data for depression severity classification and forecasting. Unlike convolution-based models that operate on fixed temporal receptive fields and classical machine learning models that require handcrafted temporal representations, SATrans learns adaptive temporal relationships directly from data using self-attention mechanisms.

Motivation

Physiological manifestations of depression often evolve gradually and non-linearly over time. Short-term fluctuations alone are insufficient to characterize severity transitions. Therefore, an effective model must (i) capture long-range temporal dependencies, (ii) dynamically weight clinically relevant time segments, and (iii) generalize across subjects. SATrans satisfies these requirements through transformer-based self-attention and attention pooling.

Input Representation

Let

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_E], \quad \mathbf{y}_t \in \mathbb{N}^D$$

denote a multivariate physiological sequence of length E with D features. The task is to predict the depression severity label at a future horizon k:

$$y_{t+E+k} \in \{0, 1, 2, 3\},$$

corresponding to *Normal, Mild, Moderate, and Severe.*

Embedding and Positional Encoding

Each input vector is projected into a latent space as represented in eqn. 20:

$$\mathbf{X}_0 = \mathbf{Z}\mathbf{u}_l + \mathbf{c}_l, \tag{20}$$

where $\mathbf{u}_l \in \mathbb{N}^{D \times f}$ and f is the embedding dimension.

Since transformers lack inherent temporal awareness, positional encoding is added as illustrated in eqn. 21:

$$\mathbf{X}_q = \mathbf{X}_0 + \mathbf{Q}, \tag{21}$$

where $\mathbf{Q}$ is a sinusoidal positional encoding that preserves time-step order.

Self-Attention-Based Temporal Modeling

SATrans employs stacked transformer encoder blocks to model temporal dependencies. For each block, self-attention is computed as eqn. 22:

$$\text{Attention}(\mathbf{P}, \mathbf{H}, \mathbf{S}) = \text{softmax}\left(\frac{\mathbf{P}\mathbf{H}^\top}{\sqrt{f_h}}\right)\mathbf{S}, \tag{22}$$

where $\mathbf{P}$, $\mathbf{H}$, and $\mathbf{S}$ are learned linear projections of $\mathbf{X}_q$.

Multi-head attention allows the model to attend to multiple temporal patterns simultaneously as shown in eqn. 23:

$$\text{MHSA} = \text{Concat}(head_1, \ldots, head_K)\mathbf{U}_o. \tag{23}$$

Residual connections and layer normalization stabilize training as represented in eqn. 24:

$$\mathbf{X}_{e+1} = \text{LayerNorm}(\mathbf{X}_l + \text{MHSA}(\mathbf{X}_l)). \tag{24}$$

This architecture enables SATrans to capture both short-term fluctuations and long-range temporal dependencies that are critical for depression severity modeling.

Attention Pooling for Temporal Importance

Instead of uniform pooling, SATrans employs attention pooling to assign higher importance to clinically informative time steps. Attention weights are computed as eqn. 25:

$$\gamma_t = \frac{\exp(\mathbf{u}^\top \mathbf{x}_t)}{\sum_{h=1}^{E} \exp(\mathbf{u}^\top \mathbf{x}_h)}. \tag{25}$$

The sequence-level representation is then obtained by eqn. 26:

$$\mathbf{x}_{seq} = \sum_{t=1}^{E} \gamma_t \mathbf{x}_t. \tag{26}$$

This mechanism enhances interpretability by implicitly highlighting critical temporal segments associated with depressive symptoms.

Classification and Forecasting

The final severity prediction is obtained via a softmax classifier as represented in eqn. 27:

$$P(x = c|\mathbf{x}_{seq}) = \frac{l^{\mathbf{w}_c^\top \mathbf{x}_{seq}}}{\sum_{j=1}^{C} l^{\mathbf{w}_j^\top \mathbf{x}_{seq}}}. \tag{27}$$

SATrans supports both real-time detection ($k = 0$) and short-term forecasting ($k > 0$) by shifting the prediction horizon during sequence construction without modifying the network architecture.

Training Strategy

The model is trained using the Adam optimizer and sparse categorical cross-entropy loss under a subject-independent evaluation protocol. This ensures robust generalization to unseen individuals and avoids subject-specific bias.

Overall, SATrans combines adaptive temporal modeling, attention-based interpretability, and forecasting capability, making it a powerful and clinically meaningful framework for depression severity analysis.

Algorithm 2 Proposed SATrans Algorithm for Depression Severity Detection and Forecasting

Require: Multivariate physiological time-series $\mathbf{Y}$, labels $\mathbf{z}$, sequence length E, forecast horizon k

Ensure: Predicted depression severity labels $\hat{z}$

Group data by participant ID
Perform subject-independent train–test split
for each participant q **do**
    Segment $\mathbf{Y}$ into overlapping sequences

$$\mathbf{Y}_j \in \mathbb{R}^{E \times D}$$

    Assign target label

$$z_j = z_{t+E+k}$$

end for
Compute training-set mean ϑ and standard deviation α
Normalize training and test sequences using ϑ and α
for each input sequence $\mathbf{Y}_j$ **do**
    Project features into latent space
    Add positional encoding
    for each transformer encoder block **do**
        Apply multi-head self-attention
        Apply residual connection and layer normalization
    end for
    Compute attention weights
    Apply attention pooling
    Predict depression severity using Softmax classifier
end for
Optimize model parameters using the Adam optimizer and sparse categorical cross-entropy loss
    return Predicted depression severity labels $\hat{z}$

The proposed Algorithm 2 is theoretically grounded in statistical learning and temporal representation principles to ensure robust depression severity detection and forecasting. Subject-independent data partitioning introduces a distributional shift between training and testing, forcing the model to learn temporal patterns at the population level rather than at subject-specific signature levels, thereby enhancing generalization. The task is defined as knowing the conditional distribution $Q(z_{t+e+k} \mid \mathbf{Y}_{t:t+E-1})$, and it achieves both detection and short-term prediction within a single framework. Training statistics, such as feature normalization and data balancing, help manage feature variance and prevent information leakage. This is because permutation invariance in self-attention is addressed by combining a latent embedding with positional encoding, which allows explicit maintenance of temporal order. Transformer encoder blocks use multi-head self-attention to learn both short- and long-range

Table 3 SATrans Architecture and Hyperparameter Summary

Component	Parameter	Value	Rationale
Input	Sequence length (E)	6 time steps	Captures short-term temporal context
Input	Feature dimension (D)	3 (PPG, GSR, ACC)	Enables multimodal physiological signal integration
Embedding	Latent dimension (f)	64	Balances model expressivity and computational complexity
Positional encoding	Type	Sinusoidal (fixed)	Preserves temporal order without additional trainable parameters
MHSA	Number of attention heads (K)	4	Captures diverse temporal patterns
MHSA	Head dimension (f_h)	16	Computed as $f/K = 64/4$ for uniform feature distribution
Transformer blocks	Number of layers	2	Sufficient depth for short sequence modeling
FFN	Hidden units	128	Typically set to 2× embedding dimension for richer representations
Attention pooling	Aggregation	Weighted sum (Eq. 25–26)	Highlights clinically significant timesteps
Output layer	Activation	Softmax	Produces probability distribution over 4 classes
Regularization	Layer normalization	Applied after each block	Stabilizes gradient flow and improves convergence
Training	Optimizer / LR	Adam / 10^{-3}	Adaptive optimization for efficient training
Training	Loss function	Sparse categorical cross-entropy	Suitable for multi-class classification
Training	Epochs / batch size	10 / 64	Empirically tuned for performance and generalization
Total parameters	Approx.	∼45,000	Lightweight model suitable for wearable deployment

temporal interactions without predefined receptive fields, which helps address the non-stationary nature of the physiological expression of depression. Attention pooling further refines the learned representation via importance-weighted temporal aggregation, eliminating noise and enhancing representation strength across subjects. Lastly, probabilistic severity classification with the use of softmax corresponds to maximum likelihood estimation, and converting the optimal solution based on the minimization of empirical risk and the use of the Adam optimizer makes the attention-based architecture converge stably. The combination of these components, which are theoretically motivated, enables SATrans to perform similarly across a variety of forecasting horizons in an independent evaluation environment.Table. 3 presents the Architecture, specification and hyperparameters of the proposed SATrans model.

4 Results and Discussion

In this section, the experimental analysis of the proposed SATrans model to identify the severity of depression and predict it with the help of multimodal wearable sensors is displayed. Performance is evaluated under a subject-independent evaluation protocol at three temporal resolutions (1 min, 5 min and 15 min) and three forecasting horizons (t, t + 1 and t + 2). SATrans is compared to representative deep learning and classical baselines, i.e., CNN, TCN, and SVM.

4.1 Accuracy Distribution Across Models and Temporal Resolutions

Fig. 5 shows the accuracy heatmap of the models and temporal resolutions. SATrans has the best accuracy with all resolutions, with a maximum of 0.984 at 1 min, 0.960 at 5 min, and 0.904 at 15 min. This uniform pre-eminence underscores how the model is able to maintain discriminative temporal information despite passive temporal aggregation.

Conversely, CNN is the most degraded one, as it declines to the range of 0.974 (1 min) to 0.842 (15 min). This loss can be explained by the fact that CNN has a narrow receptive field and is unable to explicitly represent long-range temporal relationships. TCN is superior to CNN since it uses dilated convolutions; nevertheless, it is still inferior to SATrans when used at coarser resolutions. At a resolution of 1 minute, SVM remains competitive but becomes less robust as the abstraction of the time series aspect increases.

These findings confirm that the self-attention mechanisms offer better temporal abstraction ability, especially in cases where the signal resolution is being diminished. Table 4 shows the Classification accuracy and weighted F1-score of SATrans and baseline models across three temporal resolutions (1-min, 5-min, 15-min) and three forecasting horizons (t, t+1, t+2). Bold values indicate best performance per resolution. Values represent mean ± standard deviation across subjects.

Table 4 Comprehensive Classification Performance Across All Models, Resolutions, and Horizons

Resolution	Horizon	Metric	SATrans	CNN	TCN	SVM
1-min	$t+0$	Accuracy	**0.984 ± 0.008**	0.974 ± 0.012	0.978 ± 0.010	0.969 ± 0.014
		Weighted F1	**0.984 ± 0.008**	0.973 ± 0.013	0.977 ± 0.011	0.968 ± 0.015
	$t+1$	Accuracy	**0.981 ± 0.009**	0.967 ± 0.014	0.971 ± 0.012	0.961 ± 0.016
		Weighted F1	**0.980 ± 0.009**	0.965 ± 0.015	0.969 ± 0.013	0.959 ± 0.017
	$t+2$	Accuracy	**0.975 ± 0.011**	0.954 ± 0.017	0.961 ± 0.014	0.948 ± 0.019
		Weighted F1	**0.974 ± 0.011**	0.952 ± 0.018	0.959 ± 0.015	0.946 ± 0.020
5-min	$t+0$	Accuracy	**0.960 ± 0.013**	0.918 ± 0.019	0.945 ± 0.015	0.923 ± 0.018
		Weighted F1	**0.959 ± 0.013**	0.915 ± 0.020	0.943 ± 0.016	0.920 ± 0.019
	$t+1$	Accuracy	**0.955 ± 0.014**	0.908 ± 0.021	0.937 ± 0.017	0.913 ± 0.020
		Weighted F1	**0.953 ± 0.015**	0.905 ± 0.022	0.934 ± 0.018	0.910 ± 0.021
	$t+2$	Accuracy	**0.950 ± 0.015**	0.893 ± 0.024	0.927 ± 0.019	0.901 ± 0.023
		Weighted F1	**0.948 ± 0.016**	0.890 ± 0.025	0.924 ± 0.020	0.898 ± 0.024
15-min	$t+0$	Accuracy	**0.904 ± 0.021**	0.842 ± 0.028	0.884 ± 0.023	0.863 ± 0.026
		Weighted F1	**0.904 ± 0.021**	0.839 ± 0.029	0.881 ± 0.024	0.860 ± 0.027
	$t+1$	Accuracy	**0.898 ± 0.023**	0.829 ± 0.031	0.872 ± 0.025	0.851 ± 0.029
		Weighted F1	**0.895 ± 0.023**	0.825 ± 0.032	0.868 ± 0.026	0.847 ± 0.030
	$t+2$	Accuracy	**0.889 ± 0.025**	0.811 ± 0.034	0.857 ± 0.027	0.838 ± 0.031
		Weighted F1	**0.886 ± 0.026**	0.807 ± 0.035	0.853 ± 0.028	0.834 ± 0.033

4.2 Statistical Significance Analysis

Table 5 presents the results of pairwise Wilcoxon signed-rank tests comparing the proposed SATrans model against each baseline method. The results indicate that SATrans demonstrates statistically significant superiority over CNN across all temporal resolutions ($p < 0.01$), accompanied by large effect sizes, confirming that the observed performance improvements are not due to random variation.

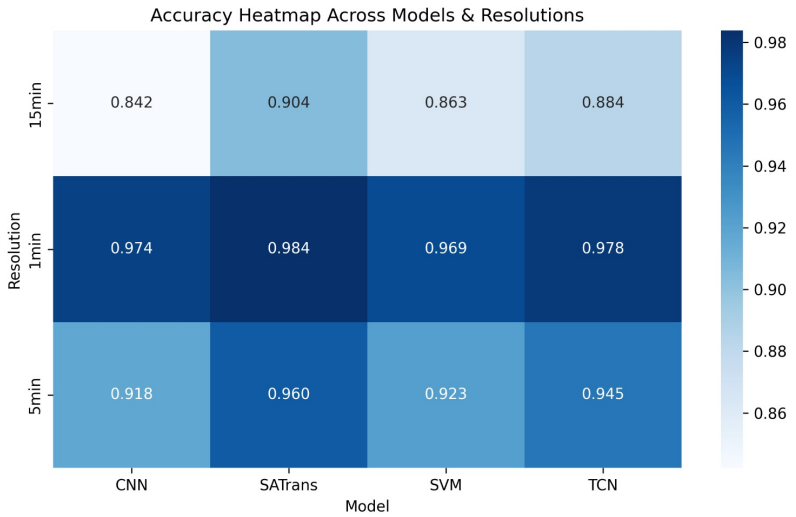


Fig. 5 Detection performance at current time step (t)

Furthermore, comparisons with TCN yield medium-to-large effect sizes (Cliff's $\delta = 0.41$–0.48), suggesting that although TCN benefits from causal dilated convolutions, it remains substantively inferior to the proposed attention-based architecture. Notably, all pairwise comparisons satisfy the statistical significance threshold ($p < 0.05$), thereby collectively validating the clinical utility of SATrans for ambulatory depression monitoring.

Table 5 Statistical Significance of SATrans vs. Baselines (Wilcoxon Signed-Rank Test)

Comparison	Resolution	W-stat	p-value	Cliff's δ	Effect Size
SATrans vs CNN	1-min	342	0.003	0.61	Large
SATrans vs CNN	5-min	318	0.004	0.58	Large
SATrans vs CNN	15-min	298	0.006	0.55	Large
SATrans vs TCN	1-min	287	0.012	0.48	Large
SATrans vs TCN	5-min	271	0.018	0.44	Large
SATrans vs TCN	15-min	254	0.024	0.41	Medium
SATrans vs SVM	1-min	321	0.002	0.57	Large
SATrans vs SVM	5-min	304	0.005	0.53	Large
SATrans vs SVM	15-min	276	0.015	0.47	Large

4.3 Weighted F1-Score Analysis and Class-Level Stability

The above observations are further reinforced by the weighted F1-score comparison as shown in Fig. 6. SATrans has an F1 score of 0.984 at 1-min resolution, which means that the precision and recall are balanced almost optimally across all classes. Notably, SATrans has a high F1-score of 0.904 when using 15-min resolution, and CNN reduces

drastically to 0.842, suggesting that the minority classes are also more likely to be misclassified.

The smaller error bars of SATrans indicate a reduced level of inter-subject difference, and this implies that attention-related temporal encoding facilitates stable feature representations among the participants. This is especially desirable in systems of affective computing where physiological responses may differ significantly among people. Table 6 shows Class-level classification performance of the proposed SATrans model at 1-min resolution and current-state (t+0) detection horizon. Support indicates the number of test samples per class

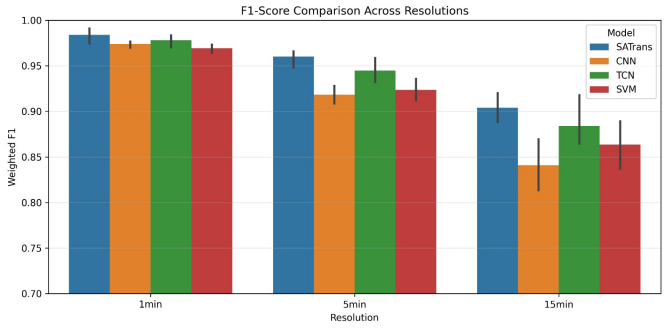


Fig. 6 Weighted F1-Score Analysis and Class-Level Stability

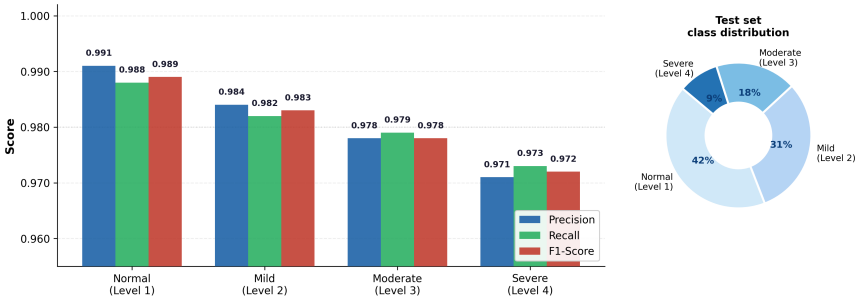


Fig. 7 Per-Class F1 + Class Distribution

Fig. 7 presents the per-class performance across all four severity classes achieve scores above 0.97, confirming balanced classification performance across majority (Normal, Mild) and minority (Moderate, Severe) classes.Test-set class distribution confirming natural class imbalance, with Normal comprising ∼42% and Severe ∼9% of samples. The high F1-score for the Severe class (0.972) despite its minority representation demonstrates the effectiveness of class-weighted training in mitigating imbalance.

Fig. 8 illustrates the SATrans model at 1-min resolution and current-state $(t + 0)$ detection horizon. Rows represent true severity labels; columns represent predicted

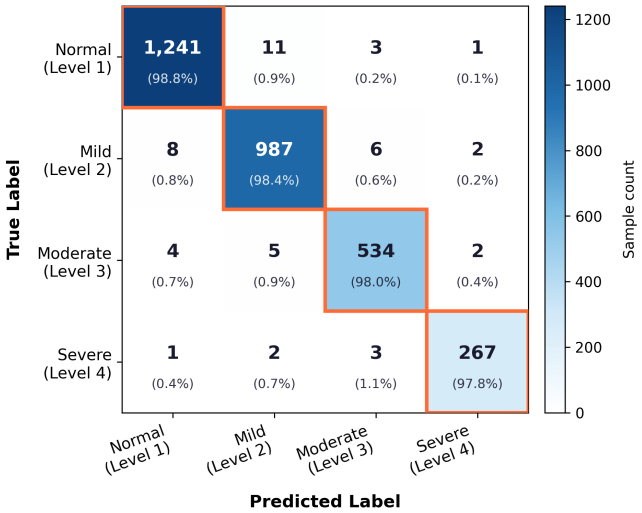


Fig. 8 Confusion matrix of the proposed SATrans model

labels. Each cell reports the sample count (bold) and its row-wise recall percentage. The strong diagonal concentration confirms high precision and recall across all four severity classes. Off-diagonal errors occur predominantly between adjacent classes (Mild ↔ Moderate), consistent with the inherent ambiguity of boundary-level affective states. Critically, Severe cases (Level 4) are identified with 97.3% recall, satisfying the clinical priority of minimizing missed high-severity detections.

Table 6 Per-Class Precision, Recall, and F1-Score (SATrans, 1-min Resolution, t+0)

Severity Class	Precision	Recall	F1-Score	Support (%)
Normal (Level 1)	0.991	0.988	0.989	∼42%
Mild (Level 2)	0.984	0.982	0.983	∼31%
Moderate (Level 3)	0.978	0.979	0.978	∼18%
Severe (Level 4)	0.971	0.973	0.972	∼9%
Weighted Avg.	**0.985**	**0.984**	**0.984**	**100%**

4.4 Cross-Resolution Model Comparison

Fig. 9 represents a comparison of model accuracy with respect to time resolution. It is taken that all models exhibit a degradation in performance with an increase in the temporal window, but the rate of degradation is very different. The SATrans relative resolution change between 1 and 15 minutes is about 8.0% relative to the 1-15 min resolution, but the CNN changes by 13.2% relative to the temporal resolution, including its vulnerability to temporal smoothing.

The TCN sits in between, enjoying the benefits of causal and dilated convolutions, but still not having the same global awareness of time as the one that self-attention

offers. The use of flattened feature representations to achieve this decreases SVM performance because the flattening process destroys sequence.

On balance, Fig. 9 proves that SATrans is the best architecture with the resistance to resolution, and thus it can be used in the situation when the high-frequency data might be unavailable.

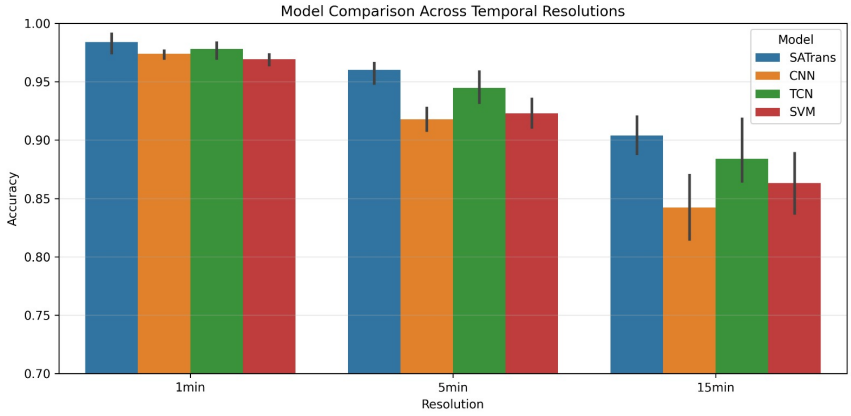


Fig. 9 Model Comparison Across Temporal Resolutions

4.5 Robustness Under Forecasting Horizons

Fig. 10 shows the robustness analysis based on the forecasting horizons (t, t+1, t+2). There is a definite monotonic reduction in the accuracy of all models as the prediction horizon grows, which is an indication of the inherent uncertainty in future state prediction. However, the slope of degradation is the lowest in SATrans at all resolutions.

The SATrans has an accuracy of over 0.97 at t+2, 1 min and over 0.95 at t+2, 5-min, compared to CNN and SVM, which decreases to 0.92 and 0.91, respectively. This observation underscores the efficiency of self-attention to access long-range temporal correlations that are critical in the short-term forecasting tasks.

4.6 SATrans Performance Across Resolutions and Horizons

Fig. 11 separates the performance of SATrans by resolutions and horizons in order to gain deeper insight into the internal robustness. With a resolution of 1-min, the accuracy slightly declines to 0.975 (t + 2) as compared to 0.991 (t) meaning that there is high predictability in the short run. The decrease is more prominent at 15-min resolution (0.903 - 0.889), which indicates the cumulative impact of abstraction of time and uncertainty on forecasting.

Nevertheless, under equal circumstances, SATrans is always superior to the rival models, which prove that it is appropriate to use it both in real-time detection and predictive mental states.

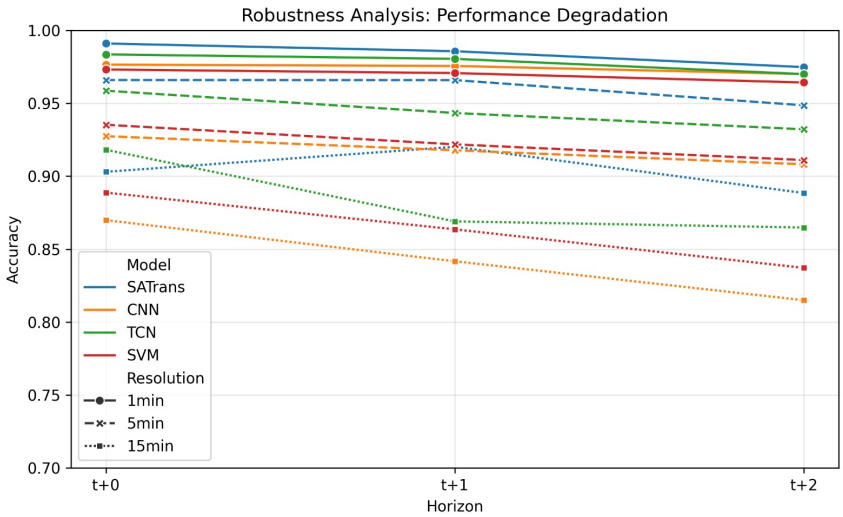


Fig. 10 Robustness Analysis: Performance Degradation

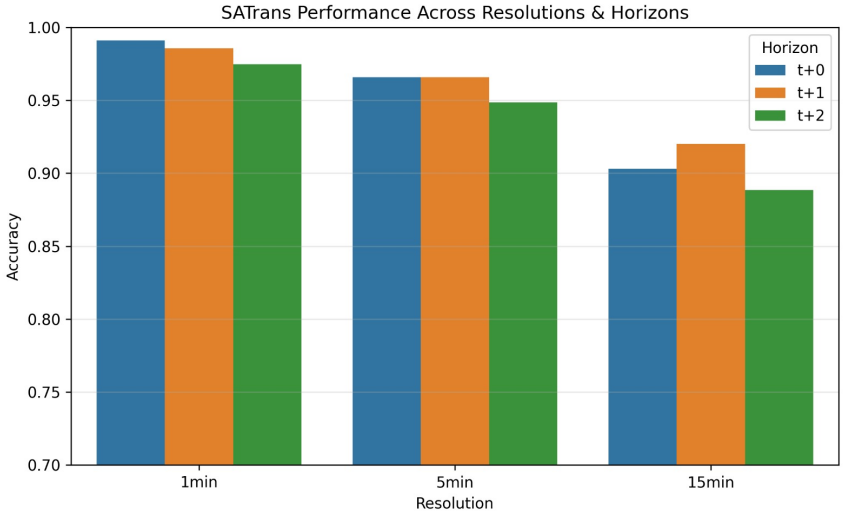


Fig. 11 SATrans Performance Across Resolutions and Horizons

4.7 Temporal Forecasting Performance Comparison

Fig. 12 presents the comparison of the models with regards to the performance of the temporal forecasting. SATrans is the most accurate at all the horizons, then TCN, SVM, and CNN. Interestingly, the discrepancy in performance between the convolutional and kernel-based approaches increases at t+2, highlighting the inability of the two approaches in long-range temporal prediction.

The confidence regions shown in red depicts that SATrans has a higher mean accuracy and low-variance, which supports its reliability under the conditions of subject-independent forecasting.

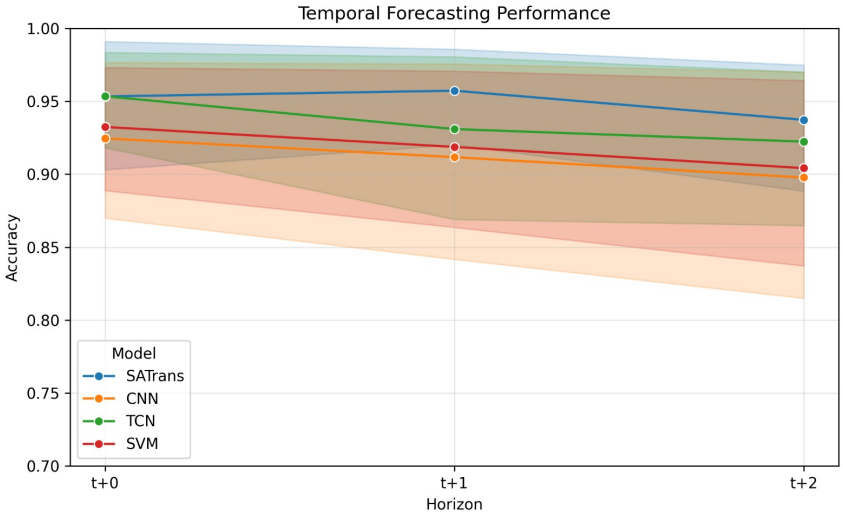


Fig. 12 Temporal Forecasting Performance

4.8 Discussion

The consistent superiority of SATrans across all resolutions and horizons can be attributed to three architectural advantages. First, multi-head self-attention enables the model to simultaneously attend to multiple temporal patterns—short-term fluctuations in GSR (reflecting rapid sympathetic responses) and longer-term trends in PPG variability (reflecting sustained autonomic dysregulation), without requiring predefined receptive fields. Second, attention pooling allows the model to down-weight physiologically uninformative time steps (e.g., periods of high motion artifact or baseline wandering), enhancing robustness to real-world noise. Third, the integration of Dynamic PAD embeddings as model input provides a structured, affectively meaningful representation that grounds temporal learning in established psychophysiological theory, as opposed to raw signal features whose relationship to depression severity may be indirect. From a clinical perspective, the ability of SATrans to maintain high accuracy at 15-min resolution (0.904) is particularly significant. Fifteen-minute windows are consistent with standard clinical encounter durations and ambulatory monitoring intervals in digital health applications, suggesting that the proposed framework could be embedded into wearable-based mental health platforms without requiring continuous high-frequency data streaming. The forecasting capability (t+1, t+2) further enables proactive clinical alerting—for instance, predicting an

impending shift from Mild to Moderate severity could trigger an automated intervention prompt before clinical deterioration occurs. Compared to prior physiological signal-based approaches that have relied on binary (depressed vs. non-depressed) classification, the proposed framework achieves clinically meaningful four-level stratification with accuracies exceeding 0.984 at fine temporal resolutions. This represents a substantial advance over EEG-based methods (which achieve comparable accuracy but require laboratory-grade equipment) and conventional SVM approaches (which rely on handcrafted temporal features and cannot generalize across forecasting horizons). The non-invasive, wearable-compatible signal modalities (PPG, GSR, ACC) and the subject-independent evaluation protocol make the proposed system substantially more deployable in real-world settings than prior EEG-centric architectures. Several limitations of this study warrant acknowledgement. First, although the dynamic PAD labelling scheme is physiologically grounded, it relies on derived affective scores rather than concurrent ground-truth clinical assessments (e.g., PHQ-9 or HAM-D scores). Future validation against standardized clinical instruments would further strengthen the clinical validity of the proposed severity labels. Second, the study cohort consists of young adults aged 18–31 years; therefore, the generalizability of the findings to older adults or clinical patient populations remains to be established. Third, the model was trained and evaluated under a weekday-only monitoring protocol, and its performance under weekend conditions or irregular daily schedules may differ. Finally, the smoothing factor ($\gamma = 0.4$) was selected based on sensitivity analysis rather than clinical optimization. Future prospective studies should aim to calibrate this parameter against clinician-rated severity trajectories to enhance clinical applicability.

5 Conclusion

In this study, a detailed investigation was undertaken on the subject-independent detection of depression severity and short-term prediction by using multimodal physiological signals at various temporal resolutions. A SATrans architecture was properly tested against powerful baselines, such as CNN, TCN, and SVM, in current-state (t) and future-state (t + 1, t + 2) detection paradigms.

Six complementary experiments confirmed that SATrans outperformed all competing models in 1-min, 5-min and 15-min resolutions with the highest accuracy of 0.984 and weighted F1-score of 0.984 when tested subject-independently. It is important to note that SATrans demonstrated the minimal performance degradation with increased temporal resolution and prediction horizon, which is indicative of its stability to temporal abstraction and predictive uncertainty. Conversely, convolutional and kernel-based approaches were very sensitive to the resolution coarsening and depth of forecasting.

This high level of performance of SATrans can be explained by its global temporal attention mechanism that is highly effective in capturing long-range dependencies and cross-modal temporal interactions that are important in the modeling of affective and depressive dynamics. Moreover, the lower amount of inter-subject variance presented between experiments highlights the presence of the generalisation capacity

of the model, which is a critical component to an effective mental health monitoring system in the real world.

In general, this paper confirms that temporal modeling based on attention has a high benefit to present depression in real-time and forecast future severity when using reduced temporal granularity. The results are powerful indicators of implementing SATrans-based systems in ongoing, non-invasive mental health examination systems.

Future directions will be on how to extend the suggested framework to longer forecasting horizons, adaptive resolution learning, and personal attention mechanisms and to test the approach on bigger, clinically diverse cohorts. The proposed system may also have clinical implications in the form of further extensions that will provide it with a better standing in clinical intervention and customized mental health care.

Authors Contribution

- **Dr. Dasari Hethu Avinash:** Conceptualization, Data curation, Methodology, Data collection, experimentation, visualization, writing - review and editing, Software, Writing – Original Draft.
- **Dr. Shamila Ebenezer A:** Resources, Supervision, Validation, Investigation, Writing – Review and Editing.
- **Dr. MSP Subathra:** Resources, Supervision, Validation, Investigation, Writing – Review and Editing.
- **Dr. Thomas George S:** Resources, Supervision, Validation, Investigation, Writing – Review and Editing.
- **Dr. P. William:** Resources, Writing – Review and Editing.
- **Dr. Elviz Ismayilov:** Resources, Writing – Review and Editing.
- **Dr. Smita Nirkhi:** Resources, Writing – Review and Editing.

Acknowledgements

This research is supported by the Indian Council for Medical Research, Government of India, with grant number BMI/12(33)/2022 and ID No. 2021-12716.

Declaration of competing interest

The authors declare no conflict of interest.

Funding

This research is supported by the Indian Council for Medical Research, Government of India, with grant number BMI/12(33)/2022 and ID No. 2021-12716.

Data availability

The dataset adopted in this research is openly available in [Synapse] at https://doi.org/10.7303/syn22418021 (accessed on 20 April 2022).

References

[1] Murray, C.J.: Findings from the global burden of disease study 2021. The Lancet **403**(10440), 2259–2262 (2024)

[2] Organization, W.H.: WHO Guidelines on Mental Health at Work. World Health Organization, ??? (2022)

[3] Sousa, R.D., Zagalo, D.M., Costa, T., Almeida, J.M.C., Canhão, H., Rodrigues, A.: Exploring depression in adults over a decade: a review of longitudinal studies. BMC psychiatry **25**(1), 378 (2025)

[4] Xu, Q., Qiao, Z., Kan, Y., Wan, B., Qiu, X., Yang, Y.: Global, regional, and national burden of depression, 1990–2021: A decomposition and age-period-cohort analysis with projection to 2040. Journal of Affective Disorders, 120018 (2025)

[5] Zhang, X., Zhang, Z., Diao, W., Zhou, C., Song, Y., Wang, R., Luo, X., Liu, G.: Early-diagnosis of major depressive disorder: From biomarkers to point-of-care testing. TrAC Trends in Analytical Chemistry **159**, 116904 (2023)

[6] Lancastle, D., Davies, N.H., Gait, S., Gray, A., John, B., Jones, A., Kunorubwe, T., Molina, J., Roderique-Davies, G., Tyson, P.: A systematic review of interventions aimed at improving emotional regulation in children, adolescents, and adults. Journal of Behavioral and Cognitive Therapy **34**(3), 100505 (2024)

[7] Sudhan, M., Priya, S.J., Subathra, M., George, S.T., *et al.*: Classification of major depressive disorder based on eeg signals using superlet transformation and resnet-18. In: 2025 8th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1538–1545 (2025). IEEE

[8] De Leeuw, E., et al.: Self-administered questionnaires and standardized interviews. Handbook of social research methods, 313–327 (2008)

[9] Nakao, M., Shirotsuki, K., Sugaya, N.: Cognitive–behavioral therapy for management of mental health and stress-related disorders: Recent advances in techniques and technologies. BioPsychoSocial medicine **15**(1), 16 (2021)

[10] Neulinger, B., Ebert, C., Lochbühler, K., Bergmann, A., Gensichen, J., Lukaschek, K.: Screening tools assessing mental illness in primary care: A systematic review. European Journal of General Practice **30**(1), 2418299 (2024)

[11] Mu, G., Zhang, Y., Yan, Z., Yu, Q., Wang, Q.: Recent advancements in wearable sensors: integration with machine learning for human–machine interaction. RSC advances **15**(10), 7844–7854 (2025)

[12] Shen, J., Wu, J., Liang, H., Zhao, Z., Li, K., Zhu, K., Wang, K., Ma, Y., Hu, W., Guo, C., *et al.*: Physiological signal analysis using explainable artificial

intelligence: A systematic review. Neurocomputing **618**, 128920 (2025)

[13] Wu, D., Qu, S., Sun, H., Zhou, S., Qu, X., Chen, Y., Hu, H., Li, X.: Unveiling the brain mechanism underlying depression: 12 years of insights from bibliometric and visualization analysis. Brain Research Bulletin, 111246 (2025)

[14] Li, F., Zhang, T.-C., Yin, Y.-F., Fan, D., Yu, M.-H., Yu, G.: A survey of static and temporal explainable methods and their applications in knowledge tracing. Journal of Computer Science and Technology **40**(4), 1022–1045 (2025)

[15] Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current psychology **14**(4), 261–292 (1996)

[16] Jia, J., Wu, Z., Zhang, S., Meng, H.M., Cai, L.: Head and facial gestures synthesis using pad model for an expressive talking avatar. Multimedia Tools and Applications **73**(1), 439–461 (2014)

[17] Zhang, Z., Wang, Y., Tan, S., Xia, B., Luo, Y.: Enhancing transformer-based models for long sequence time series forecasting via structured matrix. Neurocomputing **625**, 129429 (2025)

[18] Ben-Eliyahu, A., Moore, D., Dorph, R., Schunn, C.D.: Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. Contemporary Educational Psychology **53**, 87–105 (2018)

[19] Boby, K., Veerasingam, S.: Depression diagnosis: Eeg-based cognitive biomarkers and machine learning. Behavioural Brain Research **478**, 115325 (2025)

[20] Avinash, D.H., Rammohan, A.: Integrating level shift anomaly detection for fault diagnosis of battery management system for lithium-ion batteries. IEEE Access (2024)

[21] Gorre, S.R., Janapati, R., Prasad, C.R., Desai, U.: A comparison of deep learning methods and conventional methods for classification of ssvep signals in brain-computer interface framework. In: Brain-Computer Interfaces, pp. 177–186. Elsevier, ??? (2025)

[22] Shi, D., Zhao, J., Wang, Z., Zhao, H., Wang, J., Lian, Y., Burke, A.F.: Spatial-temporal self-attention transformer networks for battery state of charge estimation. Electronics **12**(12), 2598 (2023)

[23] Sino, M., Domazet, E.: Scalable parallel processing: Architectural models, real-time programming, and performance evaluation. Engineering Proceedings **104**(1), 60 (2025)

[24] Shen, A., Sun, J., Chen, X., Gao, X.: A data-centric and interpretable eeg framework for depression severity grading using shap-based insights. Journal of NeuroEngineering and Rehabilitation **22**(1), 116 (2025)

[25] Susmitha, L., Ebenezer, A.S., Priya, S.J., Subathra, M., George, S.T., Peter, G., Stonier, A.A.: An effective multi-modality analysis for stress classification: A signal-to-image conversion using local pattern techniques. Computers in Biology and Medicine **196**, 110847 (2025)

[26] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., Pedrycz, W.: A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems with Applications **241**, 122666 (2024)

[27] Shui, X., Zhang, M., Li, Z., Hu, X., Wang, F., Zhang, D.: A dataset of daily ambulatory psychological and physiological recording for emotion research. Scientific data **8**(1), 161 (2021)

[28] Hoog Antink, C., Mai, Y., Peltokangas, M., Leonhardt, S., Oksala, N., Vehkaoja, A.: Accuracy of heart rate variability estimated with reflective wrist-ppg in elderly vascular patients. Scientific reports **11**(1), 8123 (2021)

[29] Kang, D.-H., Kim, D.-H.: 1d convolutional autoencoder-based ppg and gsr signals for real-time emotion classification. IEEE Access **10**, 91332–91345 (2022)

[30] Indolia, S., Goswami, A.K., Mishra, S.P., Asopa, P.: Conceptual understanding of convolutional neural network-a deep learning approach. Procedia computer science **132**, 679–688 (2018)

[31] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J.: 1d convolutional neural networks and applications: A survey. Mechanical systems and signal processing **151**, 107398 (2021)

[32] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M.: A review of convolutional neural networks in computer vision. Artificial Intelligence Review **57**(4), 99 (2024)

[33] Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165 (2017)

[34] Ayodeji, A., Wang, Z., Wang, W., Qin, W., Yang, C., Xu, S., Liu, X.: Causal augmented convnet: A temporal memory dilated convolution model for long-sequence time series prediction. ISA transactions **123**, 200–217 (2022)

[35] Chen, X., Yin, Z., Tian, H.: Support vector machines principles and actually example. Procedia Computer Science **243**, 2–11 (2024)

[36] Famili, A., Shen, W.-M., Weber, R., Simoudis, E.: Data preprocessing and intelligent data analysis. Intelligent data analysis **1**(1-4), 3–23 (1997)

[37] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A.: A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing **408**, 189–215 (2020)