
# Transformer-Based Emotion Recognition Using Multimodal Data Fusion

Submitted By:
**Shivendu Mishra (206124031)**

Guide:
**Dr. R. Bala Krishnan**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI

# 1. PROBLEM STATEMENT, OBJECTIVES, & MOTIVATION

## ➢ Problem Statement

- To develop a robust multimodal emotion recognition system using Transformer-based architecture
- Effectively models the cross-modal dependencies and temporal dynamics across the physiological signals

## ➢ Objectives

- To design a Transformer-based multimodal system
- To fuse the ECG, EDA, and acceleration features to predict the emotions
- To reduce the artifacts in the signals
- To improve the accuracy

## ➢ Motivation

- To enhance the human-computer interaction
- To bridge the gap between the laboratory validations and real-world deployment
- To support affective computing

# 2-A. LITERATURE SURVEY

| Sl. No. | Paper Details | Techniques Used | Gaps Identified |
|---|---|---|---|
| 1. | **Title:** UAED: Unsupervised Abnormal Emotion Detection Network Based on Wearable Mobile Device<br>**Author:** *J. Zhu et al.*<br>***Year:*** *2023* | • Gaussian Mixture VAE<br>• 2D-CNN with stacking<br>• Whitening distance anomaly scoring<br>• Unsupervised training | • Unimodal approach – Uses only ECG signals<br>• No cross-modal attention – Cannot leverage multiple physiological signals<br>• Limited emotion granularity – Only binary anomaly detection<br>• Fixed window processing – No adaptive temporal modeling |
| 2. | **Title:** Self-supervised ECG Representation Learning for Emotion Recognition<br>**Author:** *P. Sarkar et al.*<br>***Year:*** *2022* | • Self-supervised learning<br>• Contrastive pre-training<br>• CNN-LSTM architecture<br>• Transfer learning fine-tuning | • Single modality dependency – ECG-only approach<br>• No multimodal fusion – Misses complementary signal information<br>• Supervised requirement – Needs labeled data for fine-tuning<br>• Short-term focus – Limited long-range temporal context |

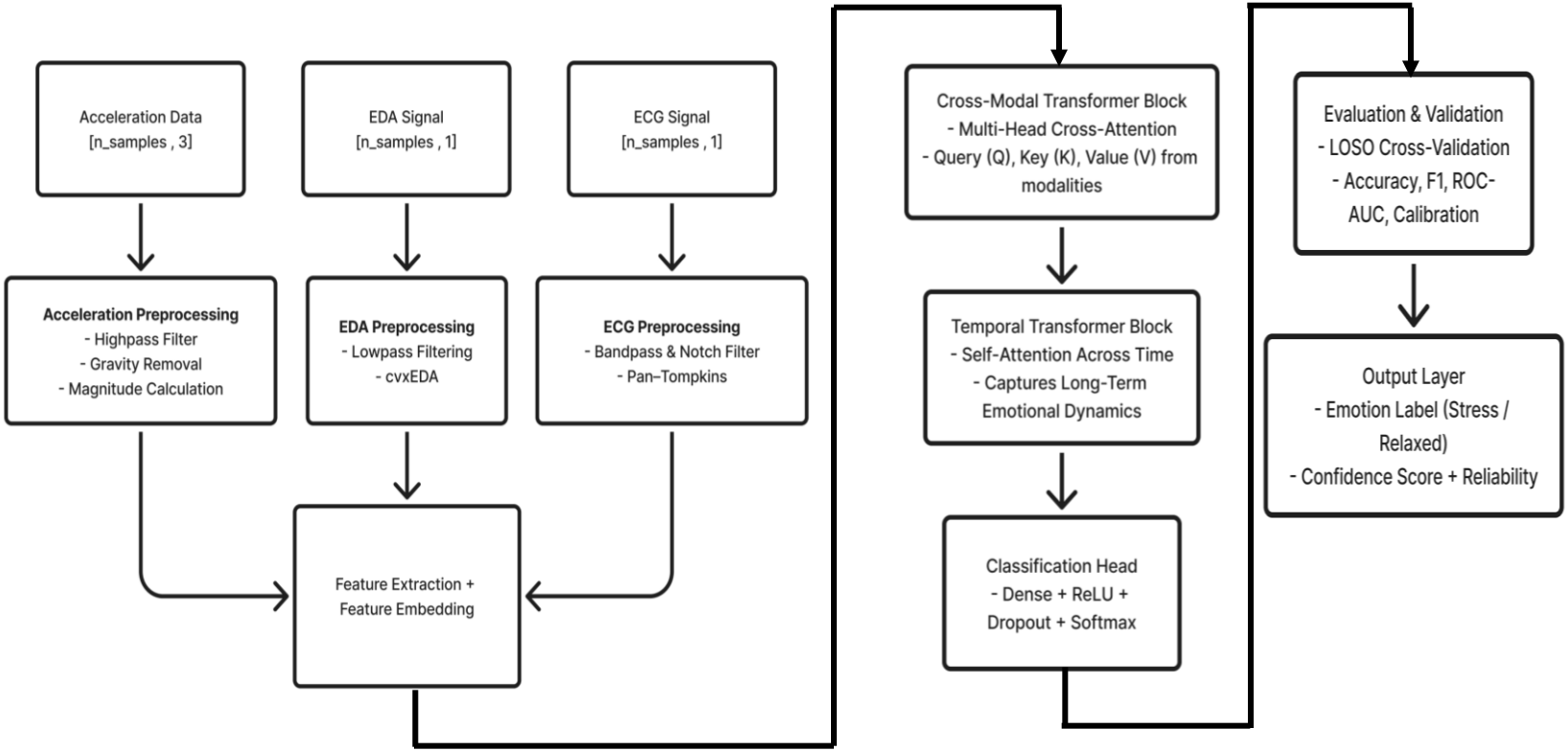
# 2-B. LITERATURE SURVEY CONTD.

| Sl. No. | Paper Details | Techniques Used | Gaps Identified |
|---|---|---|---|
| 3. | **Title:** Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition<br>**Author:** *K. Yang et al.*<br>***Year:*** *2023* | • LSTM with attention<br>• Multimodal fusion<br>• Mobile deployment optimization<br>• Behavioral + physiological signals | • Limited temporal context – LSTM struggles with long sequences<br>• Sequential processing – Inherent RNN limitations<br>• Shallow fusion – Basic attention on temporal dimension only<br>• Limited physiological coverage – Focuses more on behavioral sensors |
| 4. | **Title:** EmotionSense: An Adaptive Emotion Recognition System<br>**Author:** *Z. Wang et al.*<br>***Year:*** *2020* | • Random Forest classifiers<br>• Wearable sensor integration<br>• Context-aware recognition<br>• Personalization approach | • Shallow models – Limited feature learning capacity<br>• Traditional ML limitations – Poor scalability with complex data<br>• Feature engineering dependency – Manual feature extraction required<br>• No deep temporal modeling – Cannot capture the complex temporal patterns |

# 2-C. LITERATURE SURVEY CONTD.

| Sl. No. | Paper Details | Techniques Used | Gaps Identified |
|---|---|---|---|
| 5. | **Title:** Attention Is All You Need<br>**Author:** A. Vaswani et al.<br>***Year:*** *2017* | • Transformer Architecture<br>• Scaled Dot-Product Attention<br>• Multi-Head Attention<br>• Positional Encoding<br>• Encoder-Decoder Structure | • High Computational Complexity ($O(n^2)$ with sequence length)<br>• No Native Multimodal Support – Designed for NLP<br>• Large Data Requirements for effective training<br>• No Domain Adaptation for physiological signals |

# 3-A. BLOCK SCHEMATIC



# 4-A. ALGORITHMS

**ALGORITHM:** Pan-Tompkins + Heart Rate Variability (HRV) Feature Extraction

➢ **INPUT**

Raw ECG signal x(t) sampled at 700 Hz

➢ **PROCESS**

**Step 1:** Apply a bandpass filter to remove the baseline wander and high-frequency noise

**Step 2:** Use the Pan-Tompkins algorithm to detect R-peaks

- Differentiate → Square → Moving window integration → Threshold
- Output is a list of detected R-peak indices

**Step 3:** Compute R–R intervals to estimate beat-to-beat timing

**Step 4:** Derive Heart Rate Variability (HRV) features, such as mean RR, SDNN, RMSSD, etc.

➢ **OUTPUT**

Clean ECG signal with noise removed

# 4-B. ALGORITHMS CONT.

**ALGORITHM:** Convex Optimization (cvxEDA Decomposition)

➢ **INPUT**

Raw EDA signal $y(t)$y(t) collected from wrist sensor

➢ **PROCESS**

**Step 1:** Apply a low-pass filter (cutoff = 2 Hz) to remove high-frequency fluctuations

**Step 2:** Use cvxEDA (Convex Optimization-based Decomposition) to split the signal

- Tonic component (T): Slow baseline skin conductance
- Phasic component (P): Fast changes linked to emotional arousal

**Step 3:** Detect Skin Conductance Response (SCR) peaks from the phasic component $r$

**Step 4:** Extract statistical features — Mean, Standard Deviation, and Number of SCR Peaks

➢ **OUTPUT**

Clean ECG signal with noise removed

# 4-C. ALGORITHMS CONT.

**ALGORITHM:** Convex Optimization (cvxEDA Decomposition)

- **INPUT**

  3-axis acceleration signal $a_x(t)$ , $a_y(t)$ , $a_z(t)$

- **PROCESS**

  **Step 1:** Apply a high-pass filter (cutoff = 0.5 Hz) to remove the gravitational components
  **Step 2:** Compute the magnitude of the resultant acceleration
  **Step 3:** Segment the signal into small time windows (for example, 4 to 6 s)
  **Step 4:** Extract the activity-level features such as mean, variance, energy, and entropy

- **OUTPUT**

  Clean ECG signal with noise removed

# 4-D. ALGORITHMS CONTD..

**ALGORITHM:** Cross-Modal Transformer with Multi-Head Attention (Fusion Stage)

➢ **INPUT**

Embedded feature sequences for each modality: $e_{ECG}$, $e_{EDA}$, $e_{ACC}$ obtained from CNN encoders or preprocessing

➢ **PROCESS**

**Step 1:** Prepare attention inputs

- For each modality $m$, create: Query ($Q_m$) , Key ($K_m$) , Value ($V_m$)

**Step 2:** Compute attention weights

- Use scaled dot-product attention to decide how strongly one signal should focus on another

**Step 3:** Use multiple attention heads

- Several parallel “heads” learn different kinds of relationships

**Step 4:** All heads are concatenated and linearly combined

➢ **OUTPUT**

Model outputs a single fused vector $F_{fusion}$

# 4-E. ALGORITHMS CONTD..

**ALGORITHM:** Temporal Transformer Block (Sequential Modeling)

➢ **INPUT**

Sequence of fused embeddings $F_{fusion,t}$ obtained from the Cross-Modal Transformer

➢ **PROCESS**

**Step 1:** Add positional information

- Since Transformers don't know time order naturally, add positional encodings

**Step 2:** Compute self-attention across time

- For every time window, compute attention

**Step 3:** Residual + Layer Normalization

- Maintains stable gradients and keeps temporal context consistent

**Step 4:** A small network refines each time-step's contextual embedding, strengthening the temporal coherence

➢ **OUTPUT**

Final time-dependent emotional embedding $F_{temporal}$

# 5. EXPERIMENTS & PERFORMANCE METRICS

## Experimental Setup

- Dataset: WESAD (primary) and AffectiveROAD (supplementary)
- Training:
  - ECG/EDA CNN: 100 epochs × 32 batch size
  - Acceleration CNN: 50 epochs × 64 batch size
  - Optimizer: AdamW (Weight Decay = 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e\text{-}8$)
  - Validation: LOSO Cross Validation

## Performance Metrics

- Core Metrics: Accuracy, ROC-AUC, PR-AUC
- Calibration Metric: Expected Calibration Error (ECE) – Assesses the reliability of predictions
- Subject-Level Metric: Inter Subject Variability (ISV) – Measures the generalization

## Model Setup

- Modality specific CNN encoders for ECG, EDA, and Acceleration
- Cross-Modal Transformer Fusion + Temporal Transformer for sequence modeling

# 6. IMPLEMENTATION ENVIRONMENT

- **Hardware**
  - **Processor:** Intel i7 / AMD Ryzen 7 (8-core, 3.0GHz+)
  - **RAM:** 16 GB minimum
  - **Storage:** 500 GB NVMe SSD
  - **GPU:** NVIDIA RTX 4060 (12 GB VRAM) or equivalent CUDA-enabled GPU
  - **Operating System:** Windows 11 (64-bit)
- **Software & Libraries**
  - **Language:** Python 3.10.x
  - **IDE:** Visual Studio Code
  - **Core Framework:** PyTorch (with CUDA support)
  - **Deep Learning Models**
    - NeuroKit2/SciPy Essential for robust ECG R-peak detection, cvxEDA decomposition, and HRV analysis
    - Transformers (for implementing attention mechanisms)
  - **Data Handling & Processing:** NumPy, Pandas, OpenCV, Open3D
  - **Visualization Tools:** Matplotlib, TensorBoard, Seaborn

# 7. EXPECTED MILESTONE / WORK PLAN

| Milestone | Period | Works to be Completed |
|---|---|---|
| MS0 | 14 Aug 2025 – 26 Oct 2025 | Literature survey, tool & language selection, dataset setup (WESAD/AffectiveROAD), sensor level study, bio-signal study, algorithm selection, and project planning |
| MS1 | 29 Oct 2025 – 23 Nov 2025 | Implement preprocessing for ECG, EDA, and acceleration (filtering, HRV, SCR, activity features) and begin 1D-CNN feature extraction |
| MS2 | 26 Nov 2025 – 21 Dec 2025 | Integrate feature embeddings and Transformer-based fusion, temporal modeling, and perform preliminary training and validation |
| MS3 | 24 Dec 2025 – 31 Dec 2025 | Conduct full LOSO evaluation, performance analysis and finalize document |

# 8. REFERENCES


[1] J. Zhu, J. Jiang, and X. Zhao, *"UAED: Unsupervised Abnormal Emotion Detection Network Based on Wearable Mobile Device," IEEE Transactions on Affective Computing, vol. PP, no. 99, pp. 1–12, 2023.*

[2] K. Yang, X. Huang, and Y. Liu, *"Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition," IEEE Transactions on Mobile Computing, vol. PP, no. 99, pp. 1–10, 2023.*

[3] P. Sarkar, S. Dey, and A. Dutta, *"Self-Supervised ECG Representation Learning for Emotion Recognition," IEEE Sensors Journal, vol. 22, no. 18, pp. 17456–17464, 2022.*

[4] Z. Wang, X. Chen, and Y. Zhang, *"EmotionSense: An Adaptive Emotion Recognition System,"* IEEE Internet of Things Journal, vol. 7, no. 9, pp. 8550–8563, 2020.

[5] D. Pollreisz and N. TaheriNejad, *"Detection and Removal of Motion Artifacts in PPG Signals,"* Mobile Networks and Applications, vol. 27, pp. 728–738, Aug. 2019.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *"Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.*

[7] A. Greco, G. Valenza, M. Nardelli, M. Bianchi, L. Citi, and E. P. Scilin*go, "Force–Velocity Assessment of Caress-Like Stimuli Through the Electrodermal Activity Processing: Advantages of a Convex Optimization Approach,"* IEEE Transactions on Human-Machine Systems, vol. 47, no. 1, pp. 91–100, Feb. 2017.


# THANK YOU