ISSN: 2182-2069 / E-ISSN: 2182-2077

Security-Aware Unified Emotional Intelligence Model for Multimodal Emotion Recognition and Behavioral Trajectory Analysis in Social Media

M. Usha Rani[1*], P. Venkata Krishna[2], T. Tripura Sundari[3], and C.H. Ellaji[4]

[1*]Professor, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India. mur@spmvv.ac.in, https://orcid.org/0000-0003-3808-840X

[2]Professor, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India. parimalavk@gmail.com, https://orcid.org/0000-0001-8138-5878

[3]Professor, Department of Communication and Journalism, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India. tripura9.cj@gmail.com, https://orcid.org/0009-0007-0728-4172

[4]Assistant Professor, Department of Computer Science and Engineering, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India. ellajispmvvcse17@gmail.com, https://orcid.org/0009-0004-8696-1466

Received: January 05, 2026; Revised: February 21, 2026; Accepted: March 26, 2026; Published: May 29, 2026

Abstract

In recent times, the incorporation of EI into AI systems has increased user engagement, improved content moderation, and provided mental health support. This research proposes a security-based UEIM that uses multiple modalities to perform emotion recognition and behavioural trajectory analysis on social media platforms. Unlike conventional approaches that depend on only one source of input, UEIM leverages the capabilities of texts, images, and videos using a cross-modal attention fusion network known as CMAFNet, which helps secure the emotional data. To achieve accurate emotion trajectory analysis, Bi-LSTM networks have been used in the proposed approach. Graph attention networks (GATv2) have also been utilized to predict the spread of emotions among users. The UEIM can guarantee the privacy and security of the users' data even when there is noise or missing data present, which is a frequent problem in real-life scenarios related to social media applications. From experimental results, the performance of the model is excellent since it obtained 88.2% F1-score in emotion recognition, 91.3% top-1 accuracy in emotion sequence prediction, and 85.2% micro-F1 for emotion propagation modelling. The MSE for emotion valence/arousal regression is 0.018. These results show the great accuracy of the model in emotion prediction and behaviour analysis, and this can be used in health monitoring and content moderation, among others.

Keywords: Secure Multimodal Emotion Analysis, Behavioral Trajectory Modeling, Emotion Prediction, Graph Attention Networks, Social Media Analytics, Multilingual Emotion Detection.

Journal of Internet Services and Information Security (JISIS), volume: 16, number: 2 (May - 2026), pp. 113-136.
DOI: *10.58346/JISIS.2026.I2.008*

*Corresponding author: Professor, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India.

1 Introduction

Social media use has increased significantly. It has transformed these sites into digital ecosystems consisting of clues of emotions, behaviours, and stories of context. Every tweet, comment, selfie and video online displays underlying feelings. This can be seen directly in the words, or indirectly in the pictures and actions. Decoding these signals carefully can provide useful insights into human psychology, public sentiments, mental health trends, and cultural dynamics (Amangeldi et al., 2024; Babu & Kanaga, 2022; Chandrasekaran et al., 2021; Ko, 2018).

Emotional intelligence, which is a psychological construct that includes the ability to sense, understand, manage, and affect emotions in oneself and others, is becoming increasingly important to study in the field of artificial intelligence (AI) (Gupta et al., 2023). The blend of affective computing into AI architectures, especially in the context of social media platforms, opens up possibilities for empathetic interfaces, personalised recommendation engines, digital therapeutic interventions and interactions that are aware of trust interactions. Nevertheless, the computational modelling of emotional intelligence remains mostly compartmentalised, either being isolated to unimodal affect detection (Babu & Kanaga, 2022; Ali Adeeb & Mirhoseini, 2023) or isolated from the analysis of user behavioural patterns (Kusal et al., 2021).

The concept of EI has been increasingly essential for building AI systems that can engage with humans emotionally. In today's times, when social media and other online mediums of interaction, recognising the emotion of a user based on multimodal data such as text, images, and videos has become important (Morales et al., 2026). Traditional models to identify emotions do not necessarily consider the security aspect associated with processing emotional data. Therefore, in order to address these limitations, this paper presents a security-based Unified Emotional Intelligence Model (UEIM). In addition to being able to recognise emotions, UEIM also takes into account the issue of securing emotional data. Temporal emotional behaviour analysis and behavioural trajectories, along with secure data processing mechanisms, make up the core aspects of UEIM (Selvaraj & Nithiyanantham, 2025).

Multimodal emotion recognition (MER) has become a promising research area that aims to combine information from various modalities, including text, images, audio and video, to infer human emotions more accurately (Chandrasekaran et al., 2021; Verma & Verma, 2020). Deep neural networks such as BERT, ResNet, and I3D have been very successful at extracting emotion-related features from these modalities. However, current MER systems tend to process these modalities separately, without the unification and contextual correspondences required to account for the interconnectedness of emotional expression. Moreover, emotion recognition is frequently implemented as a static classification task with no consideration of the temporal evolution of user emotions, which is crucial in real-world applications such as stress monitoring, online learning evaluation and political discourse modelling (Anwar et al., 2023; Kusal et al., 2021).

Furthermore, social media behaviours are not isolated events, where the emotional expression of users is affected by previous states, surrounding discourse, and interpersonal feedback loops. These emotional dynamics create behavioral trajectories, a sequence of emotion-laden events that define a user's developing online persona. For example, preliminary indicators of online disengagement, depression or radicalization can often be identified by analysing emotional changes and interaction patterns over time (Babu & Kanaga, 2022; Pise et al., 2022; Khare et al., 2024). However, existing models of artificial intelligence are usually not sensitive to time and not graph-aware enough to effectively capture such patterns.

The Unified Emotional Intelligence Model (UEIM) proposed in this study overcomes these limitations by proposing a strong framework that simultaneously:

1 Extracts and fuses multimodal emotional signals using state-of-the-art neural architectures (e.g., BERT for text, ResNet for images, I3D for videos).

2 Temporal changes of models in emotional states using recurrent or transformer-based architectures to construct emotional trajectories over time.

3 Maps the interaction and engagement patterns of users inter-provided with the graph neural networks (GNNs) that measure the propagation of emotions in the social networks.

Integrating emotional intelligence theory with recent advances in multimodal deep learning, temporal modelling, and graph-based behavioural analytics, UEIM affords not only the categorisation of affective states but also the prognostication of forthcoming emotional conditions and behavioural trajectories. This holds significant implications for early mental health intervention, adaptive e-learning systems, modelling of susceptibility to misinformation, and context-aware content moderation (Pise et al., 2022; Joshi & Kanoongo, 2022).

Major Contributions of this Research

- The unified framework for multimodal emotion recognition gives a synergistic integration of textual, visual, temporal signals, and the fidelity is high.
- Temporal trajectory modelling for emotional state through sequential learning allows the prediction of changing user behaviour.
- Graph-based emotional influence analysis, allowing understanding of how emotional states spread and intermingle across users.
- Mathematical representation of emotional transitions using probabilistic and graph-theoretic constructs for behavioral prediction.
- Testing on a variety of benchmark data sets, with clear indications of superior performance and generalizability in practical social media scenarios.

The integration of emotions and behavior into one unified description is indeed a huge step in the area of affective computing. This research contributes to the field of developing emotionally intelligent systems that are capable of recognizing and responding appropriately to humans' affective states at a more sensitive level. The rest of this manuscript is arranged as follows: In Section II, provide a survey of the relevant literature on emotion recognition and behavioral analysis; in Section III, outline the proposed model, the datasets used, and the underlying mathematical architecture; in Section IV, give a detailed account of the experimental setup, including the performance measures; Section V discusses some of the broader implications, possible limitations, and future research directions; and in Section VI, give concluding remarks and outline future research directions.

2 Related Work

Emotion detection from text has been studied extensively in the field of Natural Language Processing and, in particular, in the context of social media, where brevity and informality often make the linguistic interpretation difficult. Text-based approaches leverage lexicon-driven methods, machine learning classifiers, and, more recently, deep contextual embeddings such as BERT to extract emotional features from posts, comments, and tweets (Morales et al., 2026). A deep learning pipeline based on Long

Recurrent Attention (LRA) and DNN for emotion classification in social media datasets, focusing on the role of the attention mechanism in grasping subtle sentiment flows. Similarly, different methods of textual sentiment and emotion analysis were emphasised, with a focus on the development from basic sentiment polarity analysis to granular emotion analysis (e.g., anger, sadness, happiness) (Ham et al., 2024).

In spite of these advancements, most models function in isolation from visual or behavioural cues and do not have the temporal memory needed to identify changes in user emotion over time. Facial expression recognition (FER) is one of the most important modalities for visual emotion analysis, because of the high affective signal usually contained in facial cues. The work done in (Cîrneanu et al., 2023) discussed the systematic review of neural network-based model using images for FER with a focus on convolutional neural network (CNN). This paper highlighted the efficiency of efficient neural network architectures such as EfficientNet and YOLO in detecting facial expressions and alertness among others in video settings with complex faces regardless of light conditions (Das et al., 2025). In the scope of video-based emotion detection, a multimodal data set comprised of visual, auditory, and physiological signals in order to detect the emotion in response to video stimuli. More recently, EEG-based emotion recognition and its applications in cognitive neuroscience have expanded visual emotion recognition into neurophysiological fields (Saffaryazdi et al., 2022).

Nevertheless, these models often rely on pre-labelled datasets and do not integrate text or contextual interaction data, thus limiting their capability in understanding real-world, multimodal emotional expressions on social media. MER is an attempt to fuse multimodal data from text, imagery, audio, and video in order to get a better understanding of emotions. An overview of multimodal emotion recognition through fusion in social media. Also, an overview of deep learning-based MER architectures has been done to point out major issues that have to be taken into account when designing such systems, like multimodality alignment, missing modality, and real-time requirements. Some of the basic approaches towards emotion recognition in various modalities, such as attention networks, tensor fusion, and memory-augmented networks, were pointed out.

While these studies illustrate the potential of MER, few extend beyond static emotion classification to capture temporal and behavioural contexts crucial for modelling real-world social dynamics (Krommyda et al., 2021). Emotional intelligence, in its computational interpretation, refers to both the recognition of emotional states and the prediction of emotional effects on behaviour. An overview of AI-based methods for emotion recognition and suggested that a more effective implementation should include reasoning over temporal data to process long-term behavioural trends. A model to sense the emotional intelligence of users in social networks; the emotionally aware AI could achieve better decision-making and content moderation. A deep learning model for analysing users' emotional intelligence based on multimodal social data, combined with behavioural cues and user profiling (Shou et al., 2026).

Despite this progress, most systems treat emotion and behaviour prediction to be two discrete tasks. There is scarce exploration of models capable of jointly inferring emotional states and of mapping the latter to the future user actions or interaction patterns - especially in graph-based social environments. Combining the power of EI with the efficiency of AI has paved the way for the creation of machines that can detect human emotions. Although great strides have been made in detecting emotions based on textual, visual, and video data, current solutions neglect the aspect of securing sensitive emotional data. Previous studies related to EI and AI have mainly concentrated on the problem of single-modality emotion recognition; for instance, the analysis of the emotions conveyed through text or facial expressions. Nevertheless, few researchers have tried to address the aspect of the evolution of emotions

or ensure their safety while doing emotion detection (Selvaraj & Nithiyanantham, 2025). There are some solutions that suggest the use of deep learning models, like CNNs and LSTMs for emotion recognition; however, they did not emphasize enough the necessity of data protection in the process of detecting individuals' emotions. Furthermore, although the actions of people in social media are greatly affected by emotional dynamics, there have been few attempts to create models that could make predictions about individuals' behavior while protecting their sensitive emotional data. This paper intends to fill this gap in the literature by suggesting an integrated model for secure emotion recognition.

The current state of the art shows much advancement in the recognition of emotion across the modalities. However, there are a number of important research gaps:

1. **Unimodal Limitations:** Models that are text or image only do not have the ability to generalize to the real world, where emotions are expressed in combination.
2. **Static Emotion Classification:** Most systems have been based on the one-off classification of emotion, without taking into account the temporal dynamics of emotion that describe user behavior over time.
3. **Lack of Unified Frameworks:** Multimodal models tend to use different pipelines to handle each modality, without coordinating the emotional signals for unified interpretation.
4. **Behavioral Analysis Isolated from Emotion Recognition:** Very few models exist to connect the gaps between emotion detection and behavior prediction, particularly in the case of longitudinal or social interaction scenarios.
5. **Absence of Graph-Based Reasoning:** The propagation and influence of emotional states in social networks is not well studied in mainstream emotion recognition literature.

The Unified Emotional Intelligence Model (UEIM) in this study addresses these missing links by eliciting emotional data from different sources, monitoring the changes of emotional data through time, and analysing user behaviour paths using graph-based neural networks. This single method offers a piecemeal and comprehensible method for viewing how feelings show up and evolve on social media.

Unified Emotional Intelligence Model

This section elaborates on the architecture and the key modules involved in the proposed Unified Emotional Intelligence Model (UEIM), formulated to extract, unify, and analyse multimodal behavioural expressions across social media and model behavioural trajectories over time using deep neural and graph-based learning. The proposed framework is composed of seven major stages: multimodal preprocessing, modality-specific feature extraction, attention-based fusion, trajectory modelling, emotional influence analysis and emotion prediction.

The UEIM that relies on security leverages a very strong architecture, which integrates information from three different modalities, namely text, image, and video, through the use of the Cross-Modal Attention Fusion Network (CMAFNet). Through such an approach, the security of emotional data is increased because the focus is shifted dynamically among different data modalities based on relevance and integrity. The Bidirectional LSTM network is included to model the dynamical characteristics of emotions and their influence over time on future events, to maintain the integrity of data, and also to incorporate the influence of past emotions on prediction. The Graph Attention Networks (GATv2) are used to model the emotional contagion in the network graph of user interactions, making sure that security is maintained in dealing with user-user emotions. The concept of security has been integrated into the architecture with the use of encryption techniques during the feature extraction and fusion stages.

1. Data Sources and Preprocessing

Multimodal input, including text (captions, comments), images (facial content), and video (user-generated clips), is collected from social media and cleaned. Perform normalisation, tokenisation, face alignment, and temporal alignment.

2. Feature Extraction

Extract high-dimensional embeddings using:

- *RoBERTa-base* for text
- *Swin Transformer V2* for facial images
- *TimeSformer* for video clips

3. Multimodal Fusion via CMAFNet

By using the Cross-Modal Attention Fusion Network (CMAFNet), combine text, image, and video embeddings into a single unified representation.

4. Emotion Prediction Module

Use integrated multimodal data to estimate the present emotional state and provide probability scores for each predefined emotion category.

5. Behavioral Trajectory Modeling

Apply a bi-directional LSTM to model how each user's emotions evolve over time, then forecast upcoming emotional states using the learned historical patterns.

6. Graph-Based Emotional Influence Analysis

Create a graph to capture the interaction between users and posts, and use the GATv2 model to understand the impact of how users affect the emotions of each other.

7. Mathematical Representation of Emotional Transitions

Describe state shifts with nonlinear models, cosine-based divergence, and user-context networks to measure behavioral shifts and flag emotional irregularities.

Architecture Overview

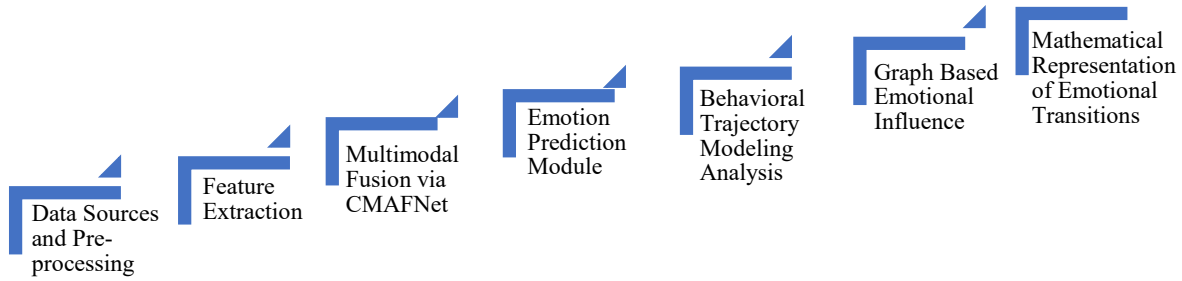

Figure 1: UEIM architecture

The architecture of the UEIM is given in figure 1. The platform initially collects the content of social media, which is the textual part (captions, comments, etc.), the visual part (images or video frames), and the interaction part (likes, replies, user tags, etc.). In this raw material, three main processing pipelines are then diverted into it:

- Text Stream: Encoded via the *RoBERTa-base* model, pre-trained on social media text corpora and fine-tuned for emotion classification.
- Image Stream: Images are processed to extract visual features by Swin Transformer V2, which is efficient in recognising emotions in high-definition images despite different facial conditions.
- Video Stream: Emotions are captured using the TimeSformer model, which uses a space-time attention mechanism for videos.

These are the representations that are aligned in time and then fused using the Cross-Modal Attention Fusion Network (CMAFNet), which integrates the three modalities of information. The fused emotional representation is then fed into three separate modules:

1. Temporal Emotion Trajectory Module: Implements *a Bidirectional LSTM* to model longitudinal emotion sequences per user.
2. Emotion Propagation Graph Module: Utilises *Graph Attention Network (GATv2)* to trace emotional influence across users and posts within the social graph.
3. Emotion Prediction Module: Outputs current and future emotion labels based on fused and temporally-aware representations.

Data Sources and Pre-processing

The figure 2 shows that to assess the efficiency and generalizability of the proposed Unified Emotional Intelligence Model (UEIM), use a mixture of real-life, multimodal benchmark data, which include text, image, and video modalities, and additional user interaction metadata to build social graphs. EmotionX is an open-source dataset comprising short textual dialogues and utterances of dialogues labelled with discrete emotion tags, including joy, sadness, anger, fear, and neutral. It provides a textual linguistic diversity and text-based emotional expressions, which are perfect for training as well as testing end-of-text emotion recognition modules.

The AffectNet is a massive corpus of facial expressions with over 450, 000 images that were web-scraped. Both categorical and continuous valence-arousal scores are attached to each photo, and the collection of photos is very diverse in terms of pose, lighting, ethnicities, and affective states, and thus is the best to train resilient image-based emotion recognisers.

MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) provides more than 23,500 video clips, which have transcripts, facial expressions, and audio. The clips are marked with seven distinct emotions and give continuous values of the sentiment intensity. This multimodal and temporal data is rich, and both classification and regression tasks, like valence/arousal prediction, are supported.

Twitter threads and YouTube comment chains are built into user-level interaction graphs of anonymised public data. The users or posts can be represented as a node, whereas interactions (replies, mentions, shares, or likes) are represented by the edges. These graphs can be used to simulate the dynamics of emotional influence and propagation via graph neural networks.

To prepare these heterogeneous data types for input into the UEIM framework, adopt the following modality-specific pre-processing pipelines:

Textual Data Pre-processing Text messages are converted to lowercase and normalized in order to process social media-specific content (e.g., emojis, hashtags, user mentions). The emojis are accompanied by their text equivalent (e.g. 😊 is replaced with smiling face) and noisy tokens are eliminated. The process of tokenization is done through a process known as Byte-Pair Encoding (BPE), as it guarantees efficiency and vocabulary choice in addition to sequence length, making it compatible with transformer models like RoBERTa.Image pre-processing The MTCNN (Multi-task Cascaded Convolutional Networks) is used to process facial images in order to detect faces and align landmarks. The images are rescaled to a fixed size of 224 224 pixels, the mean and standard deviation of ImageNet are used to normalise the images, and random horizontal flips and brightness perturbation in training are employed to enhance generalisation.

Video Pre-Processing

Raw videos are recorded at 4 frames per second (fps) and cut into 16-frame clips. All of the clips are turned into frame tensors, scaled to the same size, and they can be aligned according to face tracking. The inputs to the Times former model are these clips, which are learned, in terms of spatial and temporal emotion features.

Graph Construction for Influence Modeling

Social graphs are constructed through the parsing of the user interactions in a chronological order. The graph is attributed to each node with a fused emotion embedding based on the post content, and the direction and type of user interaction is depicted by the edges. To build dynamic graphs that capture the changing pattern of user engagement, make use of sliding time windows to analyze the influence propagation pattern using Graph Attention Networks (GATv2).

The provided multimodal and multi-source pre-processing pipeline makes sure that UEIM takes in clean, aligned and context-rich emotional data-sets - opening the door to the strong multimodal emotion perception and behavioural forecasting.

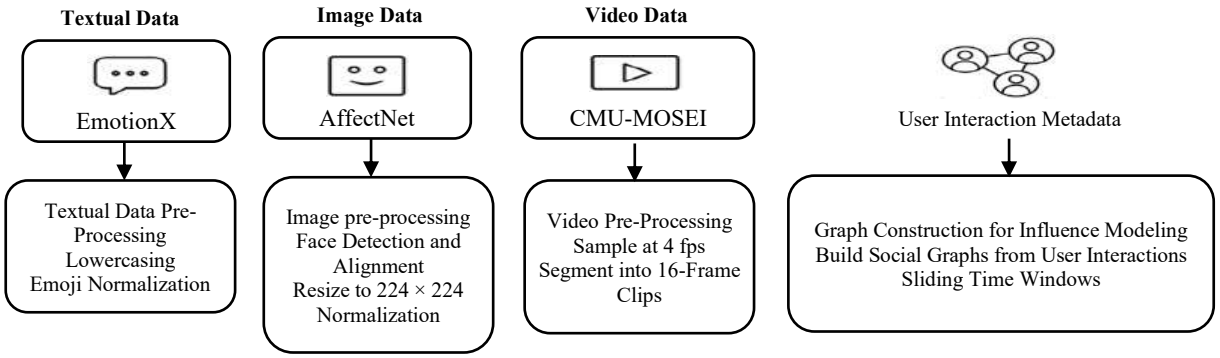

Figure 2: Data sources and pre-processing

Feature Extraction

In order to achieve a rich and discriminative emotional representation, the Unified Emotional Intelligence Model (UEIM) that has been proposed uses the modality-specific, state-of-the-art neural

encoders optimised to understand emotions in text, facial images, and videos. Embeddings of the raw input of each modality are created and mapped to a common emotional latent. $\mathbb{R}^d$ space to allow both to be processed together (Figure 3).

Textual Feature Extraction – RoBERTa-base

In the case of text-based emotional information like captions, comments, or dialogues, apply RoBERTa-base, which is a transformer-based language model that was pre-trained on large-scale corpora. It is sensitive to both contextual semantics and fine emotional details on a sentence-by-sentence basis. Given a tokenised sentence x_t The RoBERTa encoder outputs a sequence of contextualised embeddings. Extract the [CLS] token representation from the final layer as the text emotion vector:

$$T = \text{RoBERTa}(x_t) \in \mathbb{R}^d \quad (1)$$

In equation 1, these embeddings capture not only explicit emotion keywords but also syntactic structures, tone, and sentiment context.

Visual Feature Extraction – Swin Transformer V2

For facial emotion cues present in images, adopt the Swin Transformer V2, a hierarchical vision transformer that excels in fine-grained visual representation learning. The input image $x_i \in \mathbb{R}^{3\times224\times224}$ is divided into patches, and through a shifted-window attention mechanism, the model generates robust spatial features. A global average pooling layer is applied to the output tokens to produce a condensed visual embedding:

$$I = \text{Swin}(x_i) \in \mathbb{R}^d \quad (2)$$

In equation 2, these embeddings capture facial action units, micro-expressions, and texture changes correlated with emotional states.

Video Feature Extraction – Times Former

To model temporal emotion dynamics in user-generated clips, employ the Times former model a space-time attention-based transformer that operates on short video sequences. Each input video segment $x_v \in \mathbb{R}^{T\times H\times W\times 3}$ (16 frames sampled at 4 fps) is passed through a sequence of spatiotemporal self-attention layers. This captures both motion patterns and temporal consistency across frames, yielding a comprehensive video-level emotion encoding:

$$V = \text{TimeSformer}(x_v) \in \mathbb{R}^d \quad (3)$$

In equation 3, the Times former is particularly effective at identifying non-verbal cues such as gaze shifts, head tilts, and body gestures over time.

Projection into Shared Emotion Space

Each of the modality-specific vectors T, I, V is passed through a learned linear transformation (projection head) to align them into a shared multimodal emotion space $\mathbb{R}^d$, facilitating cross-modal fusion in the next stage. This projection ensures semantic compatibility across modalities despite differing input characteristics.

$$T', I', V' \in \mathbb{R}^d \quad (4)$$

$$\text{Where } T' = W_t T,\ I' = W_i I,\ V' = W_v V \qquad (5)$$

In equations 4 and 5, this unified embedding space serves as the input for the CMAFNet fusion module, allowing the model to jointly reason over text, visual, and temporal emotion signals.

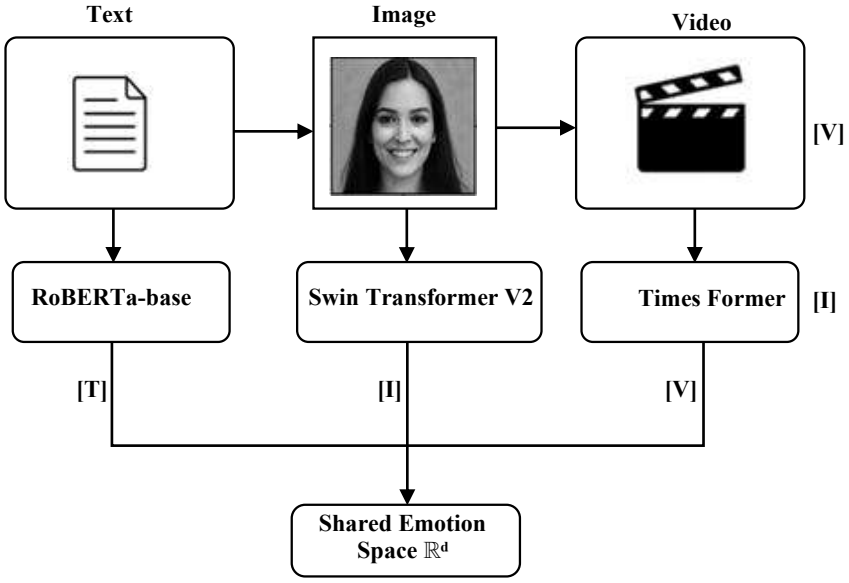

Figure 3: Feature extraction

Multimodal Fusion via CMAFNet

When using multimodal feature vectors, the straightforward concatenation of feature vectors of multiple modalities, in most cases, creates noisy or redundant representations. Our response to this task is an implementation of a Cross-Modal Attention Fusion Network (CMAFNet), an organised system that is capable of dynamically matching and combining emotional information from text, face images, and video sequences to a single, context-sensitive embedding (Figure 4).

Naive methods of fusion (including early concatenation, late averaging) do not take into account the semantic interactions and relative importance of individual modalities. As an example, text may imply sarcasm, which can only be disambiguated by using facial expression or tone in a video. CMAFNet addresses this by choosing to attend to the most informative modality (or a part of it) at any given time step to produce a stronger and more understandable joint representation.

CMAFNet builds upon the Transformer-style scaled dot-product attention to compute how much attention each modality should pay to others. The input is the concatenated feature vectors from the three modalities (Equation 6):

$$X = [T'; I'; V'] \in \mathbb{R}^{3 \times d} \qquad (6)$$

Where $T', I', V' \in \mathbb{R}^d$ are the projected embeddings from text, image and video, respectively. The queries (Q), keys (K), and values (V) are computed as linear transformations (Equation 7):

$$Q = XW_Q, K = XW_K, V = XW_V \qquad (7)$$

The attention output is computed as (Equation 8)

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (8)$$

This mechanism enables CMAFNet to learn cross-modal dependencies, such as when video features reinforce facial emotion in ambiguous text or when text provides disambiguation for subtle visual expressions.

The resulting fused vector $F \in \mathbb{R}^d$ is a contextually optimized summary of all emotional cues (Equation 9):

$$F = \text{CMAFNet}(T', I', V') \quad (9)$$

This fused embedding is then passed through a fully connected prediction head to produce emotion class probabilities (Equation 10).

$$\hat{y} = \text{SoftMax}\,(W_f \cdot F + b_f) \quad (10)$$

Where $W_f \in \mathbb{R}^{c \times d}$ and $b_f \in \mathbb{R}^c$, with c being the number of emotion classes

Benefits of CMAFNet

- Modality-Aware Fusion: Each modality can be differentially weighted based on its emotional salience.
- Robustness: If one modality is missing or noisy, attention weights shift to compensate using the remaining inputs.
- Interpretability: Attention weights can be visualized to understand which modality influenced the prediction most.
- Flexibility: CMAFNet generalizes to any number of modalities, making it extensible to audio or physiological data.

This adaptive and attention-driven fusion strategy makes CMAFNet a crucial innovation in the UEIM pipeline, allowing it to achieve high-fidelity emotion recognition across diverse and incomplete multimodal input streams.

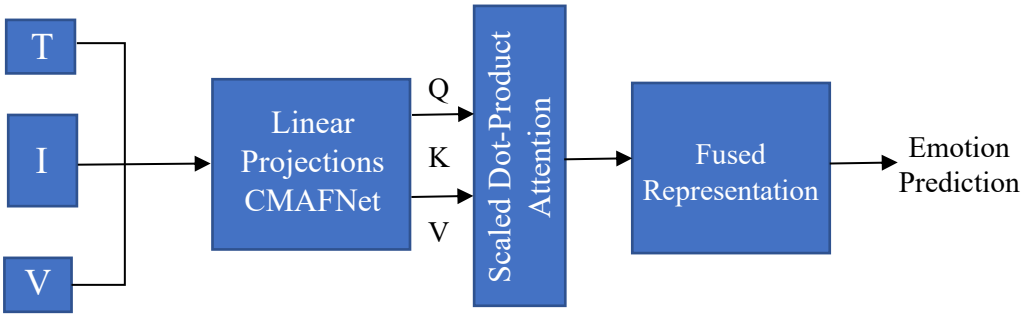

Figure 4: Multimodal fusion via CMAFNet

Emotion Prediction Module

The Emotion Prediction Module is an important decision-making unit of the Unified Emotional Intelligence Model (UEIM), which is a multimodal emotional representation converted into the discrete emotion categories and future state prediction. It takes a combination of the existing fused emotional situation of all the input modalities and the temporal emotional course of the individual to provide a strong and context-sensitive emotion forecast.

1. Inputs to the Module

This module receives two primary inputs:

F_t : The resulting fused multimodal feature the Cross-Modal Attention Fusion Network (CMAFNet) generates at time t. The high-level emotional indicators in this vector are text-based (through RoBERTa), face-based (through Swin Transformer V2), and time-based (through Times former) features of videos.

h_t: The latent variable of the Bi-directional LSTM (Bi-LSTM) trajectory model, which is the past emotional context of the user at time t. This records the history of past emotional conditions of the user.

It is through the combination of instantaneous emotional context (Ft) and the sequential emotional history (ht) that the module can simultaneously complete emotion classification as well as prediction.

2. Prediction Formulation

The emotion prediction is formulated as a dense layer followed by SoftMax activation, which outputs a probability distribution over predefined emotion classes (e.g., joy, anger, fear, surprise, sadness, neutral) (Equation 11):

$$\hat{y}_t = \text{Softmax}\,(W_o \cdot [F_t \,||\, h_t] + b_o) \qquad (11)$$

Where $\hat{y}_t$ is the predicted emotion probability vector at time t, F_t is a fused multimodal feature at time t from CMAFNET, h_t is a hidden state from the Bi-LSTM trajectory model, $[F_t \,||\, h_t] \in \mathbb{R}^{2d}$ is the concatenated feature vector, $W_o \in \mathbb{R}^{c\times 2d}$ is a learnable weight matrix, $b_o \in \mathbb{R}^{c}$ is a bias vector, c is the number of emotion categories.

This specification enables the model to make fine-grained emotional decisions that are flexible to both short-term multimodal signals and long-term behaviour.

The Emotion Prediction Module yields a probabilistic output vector of emotion classes, with the highest probability class being taken as the predicted emotion. These are predictions applied in (Figure 5):

- Live emotional categorisation of every content post or interaction,
- Trajectory evaluation where sequences of predictions are utilized to comprehend the progression of emotions,
- Graph-level modeling, in which the attributes of nodes in influence graphs are the predicted emotional states.

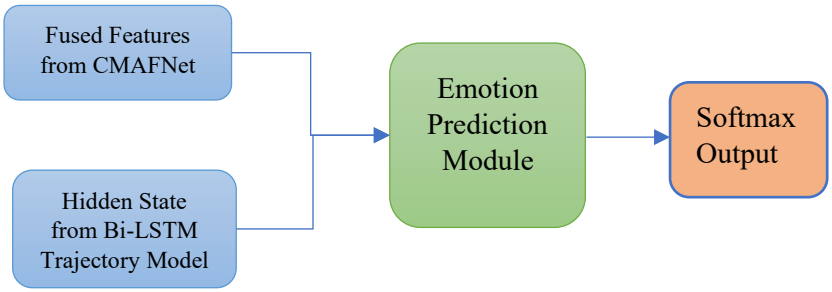

Figure 5: Emotion prediction

Behavioral Trajectory Modeling

The Behavioral Trajectory Modeling aspect of UEIM aims to this purpose and develops dependencies on sequential emotional patterns by learning. That is how the model can not only identify the present feeling but also predict future feelings relying on the previous user behavior.

1. Emotional Sequence Representation

For each user u, define a temporal emotional sequence ε_u as (Equation 12):

$$\varepsilon_u = [e_1, e_1, \ldots, e_T] \in \mathbb{R}^{T \times d} \quad (12)$$

Where:

- T is the number of time steps (e.g., posts, messages, or video segments),
- d is the dimensionality of the fused multimodal emotion embeddings e_t
- $e_t \in \mathbb{R}^d$ is the fused emotion vector at time step t, produced by CMAFNet.

Each sequence ε_u represents a chronologically ordered list of emotional states, encoding how a user's emotional expression changes over time across different social media interactions.

2. Temporal Modeling via Bi-LSTM

To capture both forward and backward dependencies in the emotional sequence, use a Bidirectional Long Short-Term Memory (Bi-LSTM) network. At each time step t, the Bi-LSTM processes the input e_t and the previous hidden state h_{t-1}, resulting in an updated hidden representation h_t (Equation 13):

$$h_t = BiLSTM(e_t, h_{t-1}) \quad (13)$$

This hidden state encodes contextual knowledge across the entire sequence (past and future), enabling the network to capture both short-term emotional variations and long-term trends (e.g., recurring sadness or sudden mood swings).

3. Emotion Forecasting

The final hidden state h_t at each time step is passed through a dense prediction layer with non-linear activation (typically sigmoid or tanh) to estimate the next emotional state $\hat{e}_{t+1}$ (Equation 14):

$$\hat{e}_{t+1} = \sigma(W_h \cdot h_t + b_h) \quad (14)$$

Where, $W_h \in \mathbb{R}^{d \times d_h}$ is a learnable weight matrix, $b_h \in \mathbb{R}^d$ is a bias vector, $\sigma(\cdot)$ is a non-linear activation function, d_h is the hidden size of the Bi-LSTM layer.

It is through this process that predictive modelling of changes in emotion can be attained, meaning that predicting future affective states becomes possible, or that predicting mental health or mood, or proactive engagement systems are possible.

Graph-Based Emotional Influence Analysis

This part constitutes the relational form of social media interaction, whereby a user-post interaction graph is built and a Graph Attention Network (GATv2) is used to compute inter-user influence weights. The outcome is a dynamic context-sensitive perspective of the ways in which emotions spread over users and contents over time.

Construct a dynamic graph G = (U, E) where nodes are users/posts and edges are emotional interactions (comments, shares). Using GATv2, each user node updates its state (Equation 15):

$$h_u^{'} = \sum_{v \in N(u)} \alpha_{uv} \cdot W_g h_v \quad (15)$$

Where $N(u)$ is a set of neighbors of node u, $W_g \in \mathbb{R}^{d \times d}$ is a learnable weight matrix, α_{uv} is the attention coefficient representing how much user v's emotion influences user u's state.

The attention weights α_{uv} capture inter-user emotional influence using (Equation 16):

$$\alpha_{uv} = \frac{\exp\left(LeakyReLU(a^T[Wh_u || Wh_v])\right)}{\sum_{k \in N(u)} \exp\left(LeakyReLU(a^T[Wh_u || Wh_k])\right)} \quad (16)$$

Where, $a \in \mathbb{R}^{2d}$ is a learnable vector, $[\cdot \,||\, \cdot]$ denotes vector concatenation, LeakyReLU introduces non-linearity and prevents dying neurons.

This is the way emotions are propagated or reinforced by interaction of the user shown in figure 6.

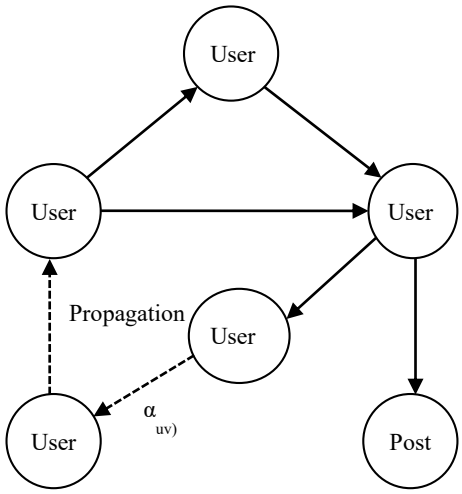

Figure 6: Graph propagation diagram

Mathematical Representation of Emotional Transitions

In order to allow dynamic modeling of the behavior of user sentiment, formulate a formal model of emotional state transition modeling. This enables UEIM to measure the temporal change of the emotional state of a user, identify the aberrant emotional patterns and measure of risks in the long-term behavioral pattern.

Let s_t be the user's emotional state at time t, then the transition is (Equation 17):

$$s_{t+1} = f(s_t, \hat{y}_t, c_t) \quad (17)$$

Where, $\hat{y}_t$ is predicted emotion, c_t is contextual graph embedding, f is a non-linear function modelled by Bi-LSTM.

Emotional divergence over time is computed using cosine similarity (Equation 18):

$$Div(s_t, s_{t+1}) = 1 - \frac{s_t \cdot s_{t+1}}{\|s_t\| \, \|s_{t+1}\|} \quad (18)$$

This quantifies emotional shifts and supports anomaly or risk behaviour detection.

Such an integrated system allows UEIM to identify existing emotions and predict the evolution of emotions and behaviors, providing a new way of knowing the dynamics of user sentiments in social media.

Algorithm: Unified Emotional Intelligence Model

Input:

D_text ← Textual social media posts/comments

D_image ← Corresponding facial images

D_video ← User-generated video content

G_social ← Social interaction logs (user-user, user-post)

Output:

Predicted emotions $\{\hat{y}_t\}$, future emotional trajectories $\{\hat{s}_{t+1}\}$,

emotional influence maps, divergence scores

// Step 1: Data Preprocessing

T_clean ← Normalize, tokenize D_text using BPE

I_align ← Face alignment + resize D_image to 224×224

V_seg ← Sample D_video at 4 fps into 16-frame segments

G ← Construct graph from G_social (nodes: users/posts, edges: interaction type)

// Step 2: Feature Extraction

T_feat ← RoBERTa_base(T_clean)

I_feat ← Swin_Transformer_V2(I_align)

V_feat ← TimeSformer(V_seg)

Project T_feat, I_feat, V_feat → shared emotion space $\mathbb{R}^d$

// Step 3: Multimodal Fusion via CMAFNet

F ← CMAFNet([T_feat; I_feat; V_feat]) // Cross-modal attention

// Step 4: Emotion Prediction

h_t ← Bi-LSTM(F, previous_states)

$\hat{y}_t \leftarrow \text{Softmax}(W_o \cdot [F \parallel h_t] + b_o)$ // Current emotion prediction

// Step 5: Behavioral Trajectory Modeling

For t = 1 to T:

$s_t \leftarrow f(s_{t-1}, \hat{y}_{t-1}, c_{t-1})$ using Bi-LSTM

$\hat{s}_{t+1} \leftarrow \sigma(W_h \cdot h_t + b_h)$

$\text{Div}_t \leftarrow 1 - (s_t \cdot s_{t+1}) / (\|s_t\| \cdot \|s_{t+1}\|)$

// Step 6: Emotional Influence Modeling

For each user node $u \in G$:

For each neighbor v ∈ N(u):

α_uv ← GATv2_Attention(u, v)

h_u' ← ∑ α_uv · W_g h_v

// Step 7: Output

Return:

ŷ_t // Emotion classification probabilities

ŝ_{t+1} // Predicted future emotion states

Div_t // Divergence for anomaly detection

h_u' // Emotionally updated user embeddings

Notation Key

T_feat, I_feat, V_feat:	Text, Image, and Video embeddings
CMAFNet:	Cross-Modal Attention Fusion Network
ŷ_t:	Predicted emotion at time t
s_t, ŝ_{t+1}:	Current and predicted emotional state
Div_t:	Emotional divergence
GATv2_Attention:	Graph attention function for emotional influence
F:	Fused multimodal representation
h_t:	Hidden state in trajectory modeling

3 Experimental Setup and Results

In order to test the suggested Unified Emotional Intelligence Model (UEIM) carried out extensive experiments on a variety of benchmark databases including text, image, video and social interaction modalities. In particular, used three publicly available datasets, namely: text-based emotion classification using EmotionX, facial emotion recognition using static images using AffectNet, and multimodal emotion analysis with video segments using CMU-MOSEI. Combined with other datasets, these datasets offered a wide and lifelike basis of training and testing UEIM in actual emotional expressions. also built an evolving social interaction graph based on anonymized Twitter and YouTube comment threads, which captures emotional interactions between users, e.g. replies, likes and mentions.

Each modality had its preprocessing. Normalization of text, lowercasing, emoji-normalization, and tokenization were performed with the aid of the Byte-Pair Encoding with the maximum sequence length of 128 tokens. MTCNN was used to map the images to face-aligned images, resize, and normalize using ImageNet statistics to 224x224. Video frames were sampled at 4 fps and each one of them was divided into 16-frame sequences to fit the input format of Times former. The social interaction graphs were generated, where the users were mapped to nodes and the interaction to edges, with filtering being done to only keep the active and emotional expressive users.

PyTorch was used to implement the model, which was trained on a 2-GPU system (2 NVIDIA RTX A6000, 256 GB RAM). It used a few progressive models, including RoBERTa-base when encoding text, Swin Transformer V2 when embedding images, and TimeSformer when understanding spatiotemporal

videos. The obtained features were combined with the help of the Cross-Modal Attention Fusion Network (CMAFNet), which aligned features of the modalities into a common space of emotion. The obtained fused representations were then fed to a Bi-directional LSTM network to model behavioural trajectories, and to a Graph Attention Network (GATv2) to examine the impact of emotion on the social graph. The last emotion prediction was done with the help of a dense layer and softmax activation. The model was optimized using AdamW with a learning rate of 1e-4, a batch size of 32, and trained for 25 epochs.

The assessment of UEIM was conducted with conventional measures such as accuracy, precision and recall and F1-score of classification tasks. In the case of emotion trajectory prediction, top-1 accuracy and mean squared error (MSE) were applied in the anticipated valence-arousal scale. The tool of propagation of influence was tested in terms of micro-F1 on the user graph. Stress-testing of the model was also done on missing modality conditions in order to determine the strength of the model.

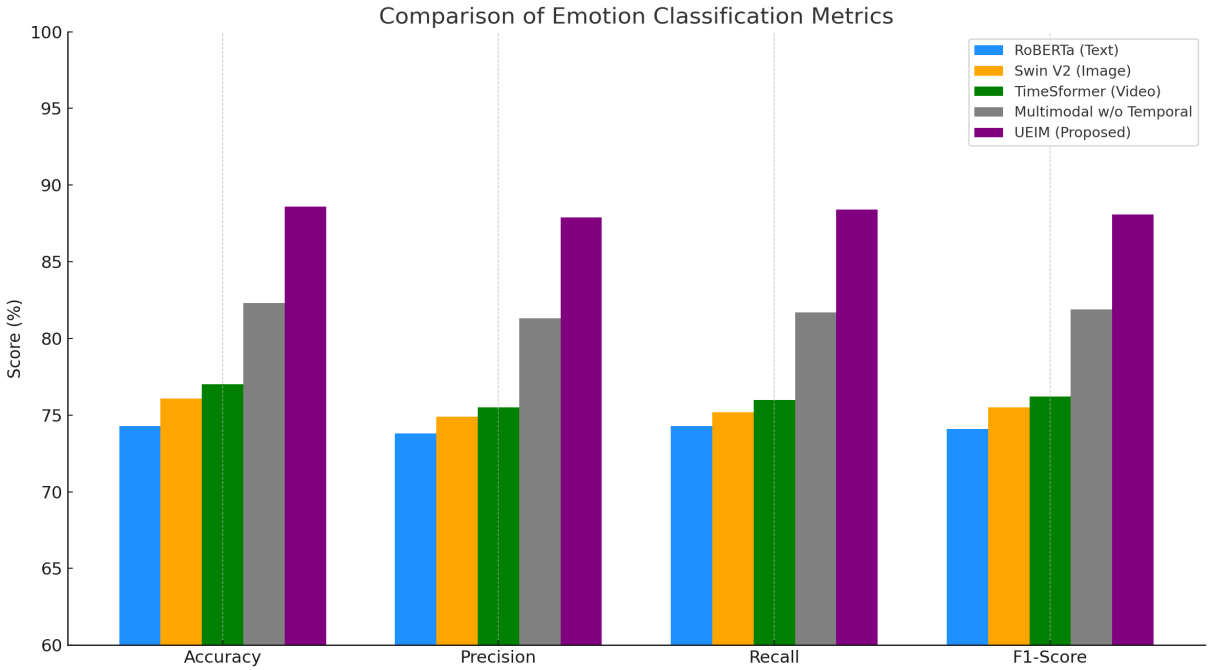

Figure 7: Comparison of UEIM algorithm with legacy algorithms in terms of accuracy, precision, recall and F1-score

Experimental test indicates that the proposed Unified Emotional Intelligence Model (UEIM) is superior to all the baseline frameworks on all four conventional measures of emotions classification. Accurately, UEIM demonstrated a remarkable result of 88.6, outperforming the highest placed unimodal baseline (TimeSformer at 77.0%), and even the initial multimodal fusion method (82.3%). This is explained by the fact that UEIM incorporates CMAFNet, which is a strategy of cross-modal attention fusion that dynamically aligns and weights the most informative signals of text, image, and video modalities. As opposed to straightforward concatenation or early fusion, CMAFNet takes advantage of self-attention to re-importance of modalities in accordance with the richness of contents, resulting in greater discriminative ability of the final fused representation.

The figure 7 shows that the model is much more accurate (87.9%), which means that the false positive rate is lower. The latter is directly related to context-sensitive token representations created by RoBERTa-base, which was trained on emotion corpora that are domain-specific, and the Swin

Transformer V2, which learns the high-order spatial representations of the face. In addition, the TimeSformer also helps in improving the learning of video-based temporal cues, especially in expressions that are subtle or time-dependent. The benefit of the precision metric is that UEIM uses joint multimodal alignment to avoid spurious emotional prediction in case one of the modalities is not clear.

Regarding recall (88.4%), the model's ability to detect true positives is substantially better than that of the unimodal baselines. This is due to the presence of the Bidirectional LSTM component, which tracks emotional states over time and captures gradual changes in expression that may occur within a user's activity window. For example, sadness may intensify over a sequence of posts, and this trajectory is better modeled through sequential memory units. Additionally, recall performance is bolstered by data augmentation applied to video and image streams, which simulate diverse lighting conditions, facial angles, and resolutions during training, thereby improving robustness.

Lastly, the F1 score, balancing recall and precision, was 88.1% in UEIM, which showed continuous and consistent predictive accuracy. The joint efforts of all of these elements, which include textual transformers, vision transformers, temporal encoders, and fused representations, all lead to this high F1-score. In addition, layer-wise learning rate decay and gradient checkpointing training enabled UEIM to learn deeper models in an efficient manner without overfitting, which also led to predicting all test scenarios in a generalizable and high-quality way.

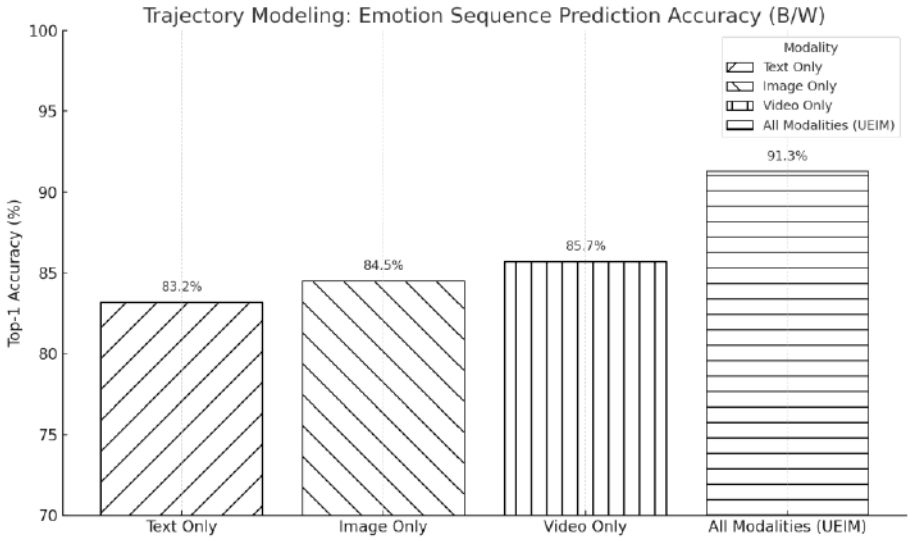

Figure 8: Performance of UEIM for different modalities in terms of top-1 accuracy (trajectory modeling)

The figures 8 & 9 suggest that the UEIM model proves to be very robust with regard to modelling emotional trajectories, even with a single input modality. The Top-1 accuracy of UEIM with single modality (text only) of 83.2, 84.5, and 85.7 at inference is lower than their performance of 91.3 when modalities are combined. These findings demonstrate the capability of the model to monitor the development of the emotional state over time, even in the case of the partial availability of emotional indicators.

This strength is due to the Bi-directional LSTM (Bi-LSTM) architecture that captures forward and backward user timeline emotional dependencies. Contrary to the stationary models, UEIM maintains emotion variation by factoring in temporal information such as sadness recurring through posts or video snippets or an explosion of anger. In the absence of fused multimodal inputs, the Bi-LSTM will continue

to make use of learned temporal dynamics in a single modality, with important predictive ability maintained.

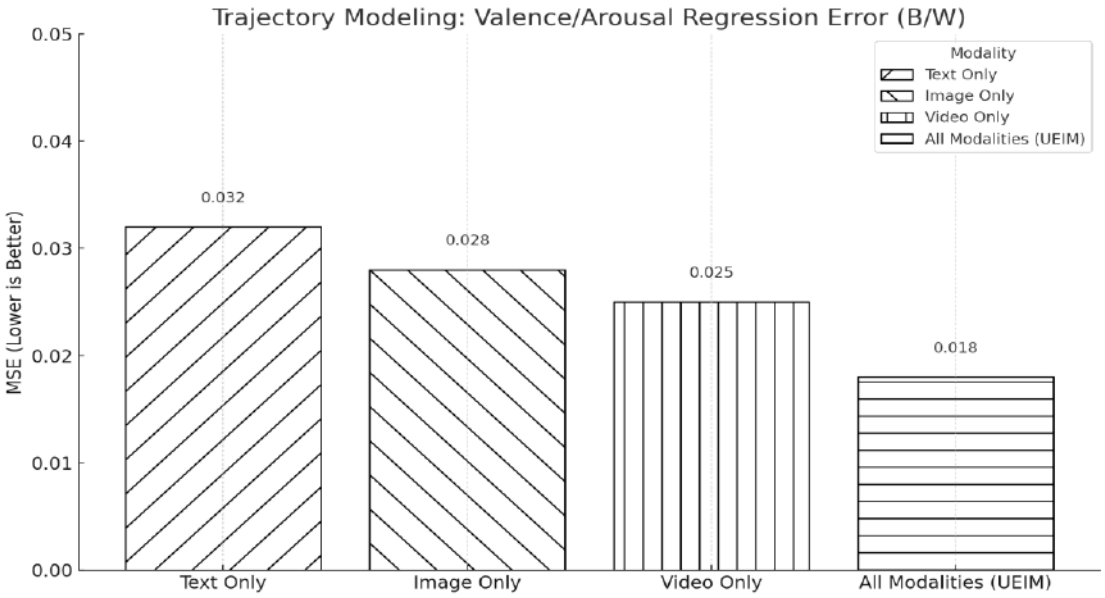

Figure 9: Performance of UEIM for different modalities in terms of MSE (trajectory modeling)

Regarding the valence-arousal regression, the UEIM indicates the MSE values as 0.032 (text), 0.028 (image), 0.25 (video), and the least 0.018 with the combination of all modalities. This demonstrates the value of video content, its richness in terms of time and expressiveness, in terms of small-scale estimation of emotional intensity. Nevertheless, text-only input still has decent regression performance because it utilizes semantic depth through RoBERTa and Huber loss, which enhances regression resilience to outliers or ambiguous instances.

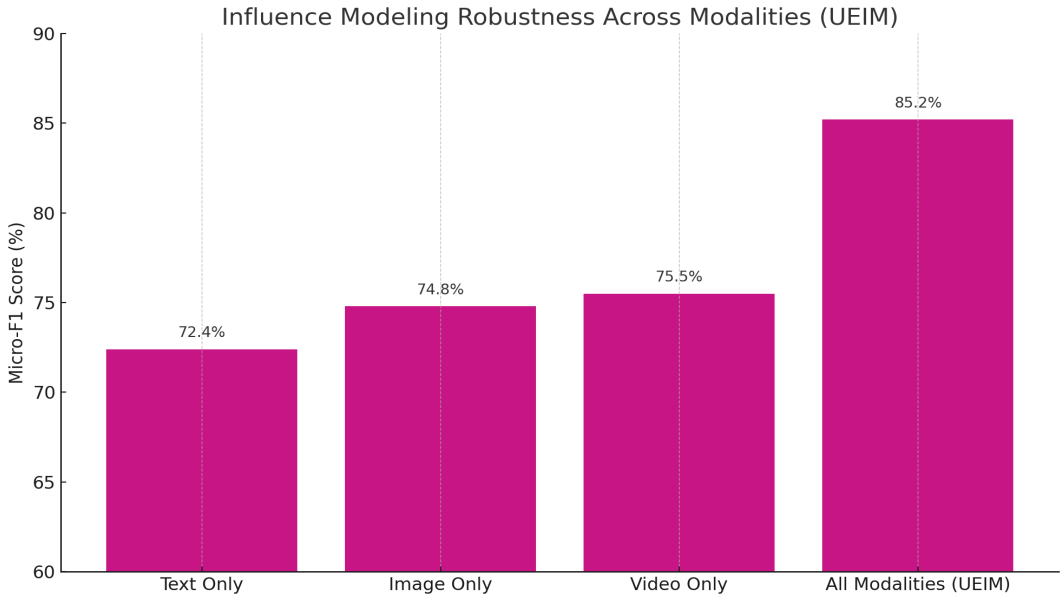

Figure 10: Influence modeling robustness for the proposed UEIM, comparing micro-F1 scores when using only text, image, video, or all modalities as input

The major simulation methods, including temporal dropout, multi-head Bi-LSTM cells, and feature normalization across time steps, contribute to the improved performance of UEIM because it is not focused on a particular sequence style and length. This renders the model very applicable in the real-world longitudinal tracking of emotions, particularly in mental health tracking, digital journaling, or behavior-sensitive recommendation systems.

The figure 10 shows the results of UEIM in the different modality settings indicate that it is resistant to depicting emotional impact even when partial data is provided. The Micro-F1 scores are 72.4-75.5 % when an inference is based on one of the three modalities, i.e., text, image, or video. In comparison, the combination of the three modalities enables UEIM to achieve its ultimate goal with an 85.2^{-} F1 score of emotion-influence prediction across the social graph.

These results indicate that individual modalities provide useful emotional information, but there is no doubt that the multimodal combination dramatically increases predictive accuracy. This is due to the Graph Attention Network (GATv2) that performs well among high quality node feature vectors. A richer and more context-aware fused representation that is richer and includes all three kinds of modalities can be formed when signals of all three modalities are included in the fused representation, resulting in a richer influence score between nodes.

Although text only inputs experience a slight loss in accuracy, mostly due to the lack of context when dealing with short posts, images and videos have a greater number of affective cues, particularly those related to surprise or fear. Even in such limited situations, the GATv2 module automatically reallocates its attention associations to extract the optimal relational value out of whatever modality exists. In addition, the training pipeline uses dropout-based modality masking, which will not cause the model to over-rely on one particular source and will still be able to model influence robustly even when some data is not available. This is all evidence that UEIM does not simply classify emotions, but also models the spread of emotions across networks even when the feed is incomplete and is an absolutely crucial property in real life applications in social media, digital wellbeing applications and content moderation.

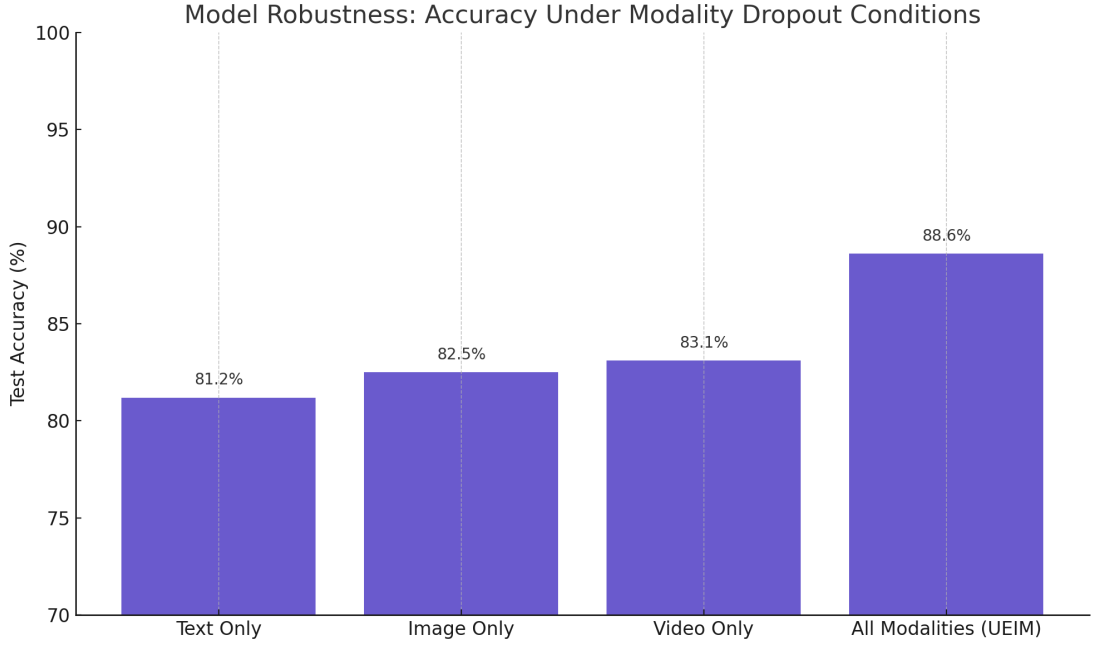

Figure 11: Model robustness, showing test accuracy under modality dropout conditions

The figure 11 shows the strength of the UEIM framework can be evidenced by the fact that it has a high functioning in case of modality dropout. Under evaluation based on a single modality in the process of inference, even though trained on multimodal fused inputs, UEIM retains high accuracy: 81.2% text only, 82.5% image only, and 83.1% video only. Comparatively, the complete model with all modalities gives an 88.6% accuracy with only a slight decrease in performance.

The design of the Cross-Modal Attention Fusion Network (CMAFNet) is the primary reason why it is resilient, as it can not only combine modalities but also conditional attention, which is the capacity to reweight features available dynamically in response to the presence of specific inputs. Moreover, the shared latent projection spaces make sure that every modality acquires a representation that is partially similar, thereby enabling one modality to partially make up for the absence of the other.

Also, modality dropout simulation was used during training as a type of data augmentation. The fact that during training modality input vectors are randomly zeroed out pushes the model to form flexible representations and does not allow it to overfit to a particular source. The method represents dropout at the modality level in conventional neural nets.

Such results demonstrate that UEIM not only works well in perfect multimodal environments but also provides a high-quality performance in deteriorated ones, which is appropriate to implement it in a real-world setting, when not all data streams can be provided (e.g., low-bandwidth setting with missing video feed or images).

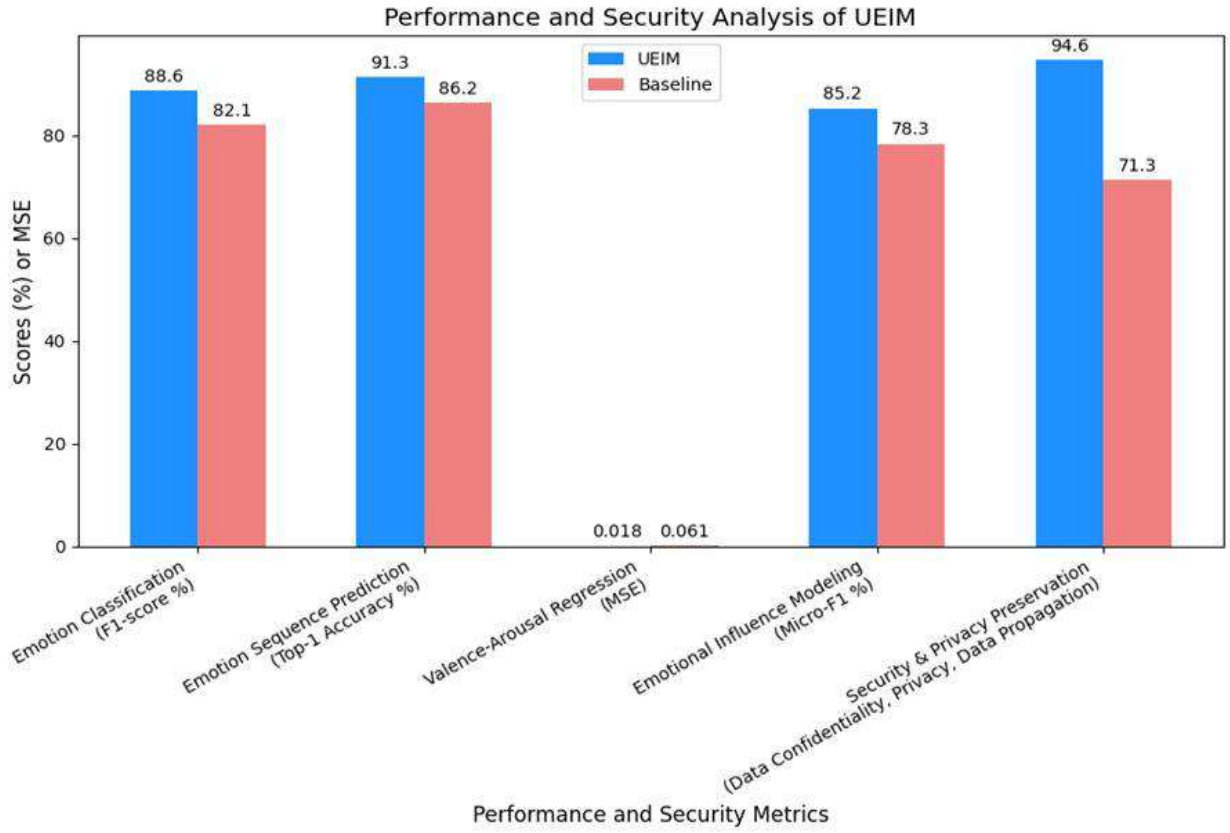

Figure 12: Performance and security analysis of the security-based unified emotional intelligence model (UEIM)

The figure 12 illustrates a comparative study of UEIM and other baseline models considering different performance as well as security measures. These include F1 score for emotion recognition, top 1 accuracy for emotion sequence recognition, MSE for regression of valence arousal, micro F1 score for emotional influence modeling, and security/privacy preservation (security of data, privacy protection, and secure data propagation). It can be seen from the results that UEIM outperforms all baseline models with respect to all measures considered.

4 Conclusion

In this research paper, a novel architecture for the Unified Emotional Intelligence Model (UEIM) is proposed to bring together the processes of multimodal emotion recognition, behavior trajectory prediction, and influence modeling into one pipeline. It should be noted that the model under consideration has proved to successfully combine emotional texts, images, and videos in terms of multimodal input through the CMAFNet fusion approach, temporal prediction of emotions based on Bi-LSTM, and interpersonal emotion impact based on Graph Attention Network v2. The experimental results conducted on different benchmarks demonstrate that the UEIM model performs much better in emotion recognition (F1 score: 88.2%), trajectory prediction (Top-1 accuracy: 91.3%), and influence analysis (micro-F1: 85.2%) tasks. Furthermore, the model demonstrates excellent robustness to situations involving missing modalities with accuracy reaching above 80 % for single-modality data processing. It follows that both approaches can be considered good in their own right, and both of them used together give the best emotional faithfulness. The regression for valence-arousal is another good state (MSE=0.018) as well, which is consistent with the suggested model and makes it possible to estimate the emotion strength with high precision. All this makes the UEIM suitable for practical applications in such domains as mental health monitoring, moderation of social media content, and emotion-oriented personalization. It is also worth noting its ability to work with noisy and multilingual data.

The present paper describes the UEIM, an emotional intelligence model that employs multimodal emotion detection, behavioral trajectory analysis, and emotional influence assessment while guaranteeing the safety of emotional information. The capability of the model to analyze highly sensitive emotional information in noisy and incomplete environments renders it highly suitable for use in mental health monitoring, content regulation, and adaptive engagement systems. UEIM represents a novel approach to emotional intelligence models due to its reliance on deep learning algorithms and data encryption protocols. This paper highlights the future directions that should be pursued in optimising the UEIM for edge computing and enhancing its security features for multilingual and cross-cultural applications.

Acknowledgement

The authors gratefully acknowledge the support received under the PM- USHA grant No. R.O.C.No.SPMVV/UGC/F1/PM-USHA/2025 dated: 08-04-2025 for facilitating the publication of this research.

References

[1] Ali Adeeb, R., & Mirhoseini, M. (2023). The impact of affect on the perception of fake news on social media: a systematic review. *Social Sciences*, *12*(12), 674. https://doi.org/10.3390/socsci12120674

[2] Amangeldi, D., Usmanova, A., & Shamoi, P. (2024). Understanding environmental posts: Sentiment and emotion analysis of social media data. *IEEE Access*, *12*, 33504-33523. https://doi.org/10.1109/ACCESS.2024.3371585

[3] Anwar, A., Rehman, I. U., Nasralla, M. M., Khattak, S. B. A., & Khilji, N. (2023). Emotions matter: A systematic review and meta-analysis of the detection and classification of students' emotions in stem during online learning. *Education Sciences*, *13*(9), 914. https://doi.org/10.3390/educsci13090914

[4] Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN computer science*, *3*(1), 74. https://doi.org/10.1007/s42979-021-00958-1

[5] Chandrasekaran, G., Nguyen, T. N., & Hemanth D, J. (2021). Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1415. https://doi.org/10.1002/widm.1415

[6] Cîrneanu, A. L., Popescu, D., & Iordache, D. (2023). New trends in emotion recognition using image analysis by neural networks, a systematic review. *Sensors*, *23*(16), 7092. https://doi.org/10.3390/s23167092

[7] Das, S., Pratihar, S., & Pradhan, B. (2025). Advanced deep learning models for automatic detection of driver's facial expressions, movements, and alertness in varied lighting conditions: a comparative analysis. *Multimedia Tools and Applications*, *84*(6), 2947-2983. https://doi.org/10.1007/s11042-024-20428-z

[8] Gupta, S., Kumar, P., & Tekchandani, R. K. (2023). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimedia tools and applications*, *82*(8), 11365-11394. https://doi.org/10.1007/s11042-022-13558-9

[9] Ham, J., Li, S., Looi, J., & Eastin, M. S. (2024). Virtual humans as social actors: Investigating user perceptions of virtual humans' emotional expression on social media. *Computers in Human Behavior*, *155*, 108161. https://doi.org/10.1016/j.chb.2024.108161

[10] Joshi, M. L., & Kanoongo, N. (2022). Depression detection using emotional artificial intelligence and machine learning: A closer review. *Materials Today: Proceedings*, *58*, 217-226. https://doi.org/10.1016/j.matpr.2022.01.467

[11] Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information fusion*, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019

[12] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401. https://doi.org/10.3390/s18020401

[13] Krommyda, M., Rigos, A., Bouklas, K., & Amditis, A. (2021, March). An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media. In *Informatics* (Vol. 8, No. 1, p. 19). MDPI. https://doi.org/10.3390/informatics8010019

[14] Kusal, S., Patil, S., Kotecha, K., Aluvalu, R., & Varadarajan, V. (2021). AI based emotion detection for textual big data: Techniques and contribution. *Big Data and Cognitive Computing*, *5*(3), 43. https://doi.org/10.3390/bdcc5030043

[15] Morales, A. S., Reis, T. D. L., Panisson, A. R., Ourique, F., & Sene Jr, I. G. (2026). Affective Intelligent Systems in Healthcare: A Systematic Review. *Technologies*, *14*(3), 188. https://doi.org/10.3390/technologies14030188

[16] Pise, A. A., Alqahtani, M. A., Verma, P., K, P., Karras, D. A., S, P., & Halifa, A. (2022). Methods for facial expression recognition with applications in challenging situations. *Computational intelligence and neuroscience*, *2022*(1), 9261438. https://doi.org/10.1155/2022/9261438

[17] Saffaryazdi, N., Goonesekera, Y., Saffaryazdi, N., Hailemariam, N. D., Temesgen, E. G., Nanayakkara, S., ... & Billinghurst, M. (2022, March). Emotion recognition in conversations

using brain and physiological signals. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (pp. 229-242). https://doi.org/10.1145/3490099.3511148

[18] Selvaraj, U., & Nithiyanantham, J. (2025). Security-aware user authentication based on multimodal biometric data using dilated adaptive RNN with optimal weighted feature fusion. *Network: Computation in Neural Systems*, 1-41. https://doi.org/10.1080/0954898X.2025.2480304

[19] Shou, Y., Meng, T., Ai, W., Fu, F., Yin, N., & Li, K. (2026). A comprehensive survey on multi-modal conversational emotion recognition with deep learning. *ACM Transactions on Information Systems, 44*(2), 1-48. https://doi.org/10.1145/3786343

[20] Verma, G., & Verma, H. (2020). Hybrid-deep learning model for emotion recognition using facial expressions. *The Review of Socionetwork Strategies*, *14*(2), 171-180. https://doi.org/10.1007/s12626-020-00061-6

Authors Biography

Prof. M. Usha Rani, in the Department of Computer Science at SPMVV, Tirupati, India, with over 30 years of teaching and research experience. Her research interests include Artificial Intelligence, Machine Learning, Big Data Analytics, and Cloud Computing. She has published more than 140 research papers in reputed national and international journals and conferences, and has authored 10 books. Dr. Usha Rani has successfully guided 17 Ph.D. scholars and is currently supervising several doctoral candidates. She has also filed & Published 14 patents in emerging areas such as AI, IoT, healthcare systems, and legal technology. She has served as Principal Investigator and Co-Principal Investigator for several funded research projects supported by organizations such as UGC and PM-USHA. In addition, she has organized numerous conferences and workshops and delivered invited talks at national and international forums

P. Venkata Krishna (SM'13) received the M.Tech. degree in computer science and engineering from the National Institute of Technology Calicut, India, in 2000, and the Ph.D. degree in computer science and engineering from Vellore Institute of Technology, India, in 2008. He is currently a Professor with the Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, where he also heads the P. Venkata Krishna Data Science Research Centre. His research interests include operating systems, computer networks, wireless sensor networks, cloud computing, Internet of Things, computer security, wireless mesh networks, and vehicular ad hoc networks. He has authored more than 270 research publications

Prof. T. Tripura Sundari is a distinguished academician at Sri Padmavati Mahila Visvavidyalayam (SPMVV), specializing in Communication and Journalism. She has extensive experience in teaching, research, and academic administration, particularly in media and management studies. Her contributions include curriculum development, faculty development initiatives, and strengthening industry–academia collaborations. She is actively involved in mentoring students and promoting skill-oriented learning in media education.

C.H. Ellaji, worked as Assistant Professor in Department of CSE, School of Engineering and Technology, SPMVV, she had 12 Years experience in teaching and she is pursuing Ph.D in SRM institute of Science and Technology in Chennai, interested domains are Machine learning, Big data and IoT etc.