

EmoVibe: AI-driven multimodal emotion analysis for mental health via social media dashboards

Deepali Vora, Aryan Sharma, Mudit Garg, Steve Francis

Department of Computer Science and Engineering, Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University), Pune, India

Article Info

Article history:

Received Mar 12, 2025

Revised Sep 8, 2025

Accepted Oct 18, 2025

Keywords:

Affective state estimation

Emotion dashboard

Hybrid transfer learning

Late-fusion

Multimodal emotion recognition

Real-time monitoring

ABSTRACT

Monitoring mental health via social media often utilizes unimodal approaches, such as sentiment analysis on text or single-staged image categorization, or executes early feature fusion. However, in real-world contexts where emotions are conveyed via text, emojis, and images, unimodal approach leads to obscured decision-making pathways and overall diminished performance. To overcome these limitations, we propose EmoVibe, a hybrid multimodal AI framework for emotive analysis. EmoVibe uses attention-based late fusion strategy, where text embeddings are generated from bidirectional encoder representations from transformers (BERT) and visual features are extracted by vision transformer. Subsequently, emoticon vectors linked to avatars are processed independently. Later, these independent data features are integrated at higher levels, enhancing interpretability and performance. In contrast to early fusion methods and integrated multimodal large language models (LLMs) like CLIP, Flamingo, GPT-4V, MentaLLaMA, and domain-adapted models like EmoBERTa, EmoVibe preserves modality-specific contexts without premature fusion. This architecture saves processing cost, allowing for clearer, unambiguous rationalization and explanations. EmoVibe outperforms unimodal baselines and early fusion models, obtaining 89.7% accuracy on GoEmotions, FER, and AffectNet, compared to BERT's 87.4% and ResNet-50's 84.2%, respectively. Furthermore, a customizable, real-time, privacy-aware dashboard is created, supporting physicians and end users. This technology enables scalable and proactive intervention options and fosters user self-awareness of mental health.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Deepali Vora

Department of Computer Science and Engineering, Symbiosis Institute of Technology Pune

Symbiosis International (Deemed University)

Lavale, Mulshi, Pune, Maharashtra 412115, India

Email: deepali.vora@sitpune.edu.in

1. INTRODUCTION

Mental health is a significant component of feeling good health-wise. It influences emotional, psychological, and social factors that dictate how individuals think, feel, and interact. With today's very connected digital era, social media sites are places where people can express themselves and learn [1]. While these sites incite discussion, they also create a different kind of social bonding that forces us to compare ourselves to others. Cyberbullying and emotional distress exacerbate mental health issues such as depression, loneliness, and anxiety. The World Health Organization estimates that almost 970 million people globally are impacted by mental health disorders, and India alone accounts for over 197 million cases, indicating an urgent need for scalable and accessible mental health monitoring solutions [2], [3].

Social media sites are expected to reach a global user base of 6.0 billion by 2028. Being a double-edged sword, they enable users to express emotions and gain approval to enhance social connectedness. On the other hand, they magnify the adverse phenomenon of social comparison, cyberbullying, and reliance on external approval by likes and comments. These relations worsen psychological disorders, such as anxiety, depression, and low self-esteem. According to one study, many individuals experience mental instability and self-doubt because of the culture of social comparison which is encouraged on the internet [4]. These barriers are becoming more and more well-known, but the scope of currently available methods of mental health evaluation remains low. Unimodal information is used in most studies and this approach fails to appreciate the intricacy of expression of emotion [5]. Single-image-based emotion detection and sentiment analysis based on text alone are just two instances of this. Additionally, since these models typically utilize retrospective assessments instead of real-time functionality, they miss significant opportunities for timely intervention [6]. Human feelings are not conveyed in a single sentence due to their internal complexity. Current models cannot efficiently leverage the elaborate multimodal environment generated when a picture or emoji describes a user's emotional response [7].

This work suggests an AI multimodal approach to identifying emotional states from social media emoticons, images, and text. It adopts hybrid transfer learning to match inter-modal features and utilizes stronger models such as vision transformers for images and bidirectional encoder representations from transformers (BERT) for text. This approach can identify complex emotional states, including contradictory emotions or ambiguous expressions, frequently overlooked by current methods. The real-time capability enables proactive mental health screening and treatment, with a personalized behavioral dashboard making it useful for general users and healthcare providers. The model follows ethical standards and includes privacy-enhancing capabilities for managing sensitive user data securely.

The list of key contributions of this work includes:

- EmoVibe is a multimodal AI framework combining text, images, and emoticons through late fusion for real-time and accurate emotion recognition.
- Hybrid transfer learning design, integrating pre-trained models (BERT, ResNet-50) with bespoke long short-term memory (LSTM) and convolutional neural network (CNN) designs to maximise domain-sensitive emotion detection.
- An innovative real-time social media emotional dashboard (SMED) for interactive monitoring of emotional trends and timely intervention.
- Combining privacy-preserving methods and ethical data practices to securely process sensitive mental health data.
- Large-scale benchmarking on GoEmotions, FER, and AffectNet datasets with better performance (accuracy of up to 89.7%) than unimodal and early fusion baselines.

The suggested framework significantly advances comprehending and enhancing psychological well-being in the digital age, utilizing cutting-edge machine learning models and multimodal integration. The following were found to be the principal research questions to guide the study: i) RQ1: design a multimodal platform to integrate and assess emoticons, text, and images on social media for detecting behavioral patterns such as anxiety, anger, and depression; ii) RQ2: hybrid transfer learning is applied to enhance accuracy and reliability in emotion detection in numerous modalities; and iii) RQ3: develop an emotional dashboard in real-time for comprehensive visualization that enables timely and customized mental health care.

This paper is divided into seven sections. Section 2 is the literature review. The methodology is described in section 3. The implementation is described in section 4. Section 5 presents the findings and comments. Section 6 presents the main conclusions and implications. Finally, section 7 outlines upcoming work, and followed by a list of references.

2. LITERATURE REVIEW

The intersection of AI and monitoring of mental health has resulted in massive research in emotion recognition using digital data on a large scale. Earlier research used single-modal data, such as text and images, to identify behavioral traits, but it failed to capture emotional expressions on social media. Advanced natural language processing (NLP) and computer vision models allowed the development of multimodal systems, resulting in more context-based mental health monitoring.

Hybrid transfer learning and data fusion methods address modality-specific shortcomings. Late fusion architectures, for instance, enhance classification accuracy by aligning information in text, images, and emoticons. Issues persist, however, such as computing requirements, privacy, and real-time analysis capability. Table 1 captures influential findings, methodologies, and contributions in prior work, establishing the goals and scope of this research.

Table 1. Summary of key studies on multimodal emotion detection

Ref	Methodology	Dataset used	Performance	Limitations	Modality used	Nosology focused
[8]	BERT-based text summarization with depression detection	DAIC-WOZ	F1-score: 0.81 (validation set)	Token length limitation of BERT models	Text	Depression
[9]	Hybrid deep learning model combining FastText, CNN, and LSTM for depression	Social media (Twitter, Reddit)	Improved accuracy over state-of-the-art models	Limited to text-based data; requires feature engineering	Text	Depression
[10]	Signal-image encoding with deep learning for emotional state recognition	Real-world sensor data	98.5% accuracy for emotion classification	Small training dataset, limited to sensor data	Physiological sensors, image	Mental well-being, emotional states
[11]	Deep learning with contextual emotion detection using image data	EMOTIC, MSCOCO, ADE20K	mAP: 79.6%	Only considers image data, requires contextual understanding	Image (body language, context)	Emotion detection (contextual)
[12]	Functional network connectivity with deep learning for mental health	rs-fMRI data (brain imaging)	Improved accuracy	Lack of interpretability in deep learning models	Neuroimaging (rs-fMRI)	Mental health (depression, stress, anxiety)
[13]	Time-enriched multimodal transformer for depression detection	Twitter, Reddit, multimodal datasets	F1-score: 0.931 (Twitter), 0.902 (Reddit)	Requires precise time information between posts for optimal results	Text, image (EmoBERTa, CLIP embeddings)	Depression
[14]	Machine learning for multimodal mental health detection (passive sensing)	Social media, smartphones, wearable devices, audio, and video	Varies by approach, generally improves with fusion	Requires careful fusion of features from heterogeneous data	Text, audio, video, wearables	Multiple mental health disorders (depression, and anxiety)
[15]	Mobile-based application for preventing and treating mental health issues	Mobile application dataset	No specific performance metrics were provided	Limited to mobile app functionality and access to health departments	Mobile app, behavioral tracking	Mental health disorders in adolescents
[16]	BERT-based classification for phobia subtypes in tweets	Novel tweet dataset (811,569 tweets)	F1-score: 78.44% (binary), 24.01% (multi-class)	Limited to text data; not applicable to all phobia subtypes	Text	Phobia, anxiety
[17]	Machine learning for mental health detection using passive sensing	Multiple passive sensing datasets	Varies with the method applied	Privacy concerns: requires fusion of multi-source data	Text, image, audio, wearables, video	Various mental health disorders (general)
[18]	Machine learning for emotion detection and sentiment analysis	123 papers reviewed	Varies by method	Limited to sentiment analysis; data and application domain-focused	Text	Emotion detection, sentiment analysis
[19]	Mental health analysis in social media posts (survey)	Social media (Twitter, Reddit, Sina Weibo)	Varies by method and dataset	Lacks standardized metrics across studies; significant variations in approaches.	Text	Depression, stress, suicide risk
[20]	Interpretable mental health analysis using large language models (LLMs)	Social media (Twitter, Reddit)	State-of-the-art interpretability and accuracy	Requires domain-specific fine-tuning; limited open-source datasets	Text, image (via LLM)	Mental health (general)
[21]	Multimodal learning with transformers for mental health analysis	Social media and multimodal datasets	Varies by application	Needs extensive data and multiple modalities; challenges with inter-modality	Text, audio, image	General mental health analysis
[22]	Multimodal analysis for depression detection in social media	Twitter (8,770 annotated users)	Improved F1-score over unimodal approaches	Challenges with real-time deployment; need for ethical data usage	Text, image, video	Depression, suicidal tendencies

Recent multimodal LLM frameworks have greatly improved vision-language comprehension. CLIP [23] constructs joint embeddings via contrastive pretraining; Flamingo [24] introduces visual contexts into LLMs through gated cross-attention; GPT-4V [25] provides a unified approach for multimodal reasoning; and MentaLLaMA [19] focuses on interpretability for mental health evaluation. While these approaches are valuable, most models utilize early, or token-level fusion, which tends to diminish modality-specific

attributes and suffer from high computational costs. In contrast, EmoVibe utilizes attention-guided late fusion to maintain independent high-level embeddings, align them, and neutralize conflicts before merging. This also applies to emoticon representations, which enables better interpretability, reduced resource usage, and real-time responsiveness for clinician-facing dashboards.

Current research highlights several essential challenges in multimodal AI systems for mental health. Unimodal approaches fail to consider the complexity of emotional expressions conveyed through several data types leading to imprecise or lacking judgments. Furthermore, privacy and ethical concerns make data gathering and analysis much more complex, and the lack of real-time monitoring has made it more difficult to act quickly. Furthermore, although existing multimodal models in [14], [15], [17], [19] exhibit potential, they have difficulty striking a compromise between computing economy and accuracy. Since models frequently encounter constraints because of the scarcity of high-quality, open-source training data as in [16], [18], [20] there is a notable void in integrating large-scale, multimodal datasets. Furthermore, many techniques are still hindered by the difficulty of fine-tuning LLMs using domain-specific data [21]. These shortcomings highlight the need for a multimodal framework that addresses the scalability and interpretability of AI systems in real-time scenarios for a trustworthy, scalable, and morally sound mental health evaluation.

The EmoVibe model, a post-fusion methodology, preserves modality-specific properties while resolving cross-modal conflicts by using attention. It combines pre-trained models such as BERT and ResNet-50 with a customized LSTM and CNN architecture with high performance and low computational complexity. Interestingly, it incorporates emoticons, text, and images to provide high contextual sensitivity towards emotional cues. The framework is optimized for real-time deployment with optimized pipelines and light configurations, enabling scalable and responsive emotion tracking. EmoVibe remedies deficits in fusion strategy, modality fusion, interpretability, and scalability, providing a robust, ethical, and full-fledged solution for AI-based mental health analysis on the internet.

Multimodal emotion recognition has come a long way, yet the existing models have modality conflict, expensive computational cost, limited real-time performance, and insecure privacy. In early methods of fusion, modality-specific knowledge is likely to be diluted and performance degraded. Most models do not focus on emoticons, which are needed to decode nuanced emotional responses in social media posts. To mitigate these restrictions, the present report proposes a scalable, ethical and real-time multimodal model on late fusion, hybrid transfer learning and emoticon robust integration to enhance emotional analysis.

3. METHOD

A multimodal mental health assessment framework is envisioned to transcend the shortcomings of unimodal solutions by combining social media text, images, and emoticons. Leveraging cutting-edge machine learning/deep learning methods, the system will offer real-time user mental health insights in terms of indicators of anxiety, anger, and depression. The resulting SMED will visually represent emotional trends among users and professionals, enabling early detection and ongoing monitoring. The data collection module collects multimodal data through social media APIs, grouping it by emotional categories and applying standardized formats. Preprocessing operations are applied to all modality types and strict quality control guarantees data reliability for accurate model predictions. The overall structure of the envisaged EmoVibe framework is depicted in Figures 1 and 2, showing two different stages of the pipeline. Figure 1 illustrates the first phase, starting from the data collection module. It is where text content, emoticons, and images are mined from social media using API integration. The pre-processing operations are modality-dependent: text data is tokenized and padded, images resized and augmented, while emoticons are translated into sentiment classes. Each modality must be cleaned and normalized here in preparation for parallel processing.

The primary analysis module combines an LSTM text analysis model and a CNN image analysis model. Training each model individually and fusing their outputs in a cross-modal interaction layer improves the system's ability to make accurate emotional predictions. Pre-trained models are optimized using hybrid transfer learning, ensuring maximum system performance on real-world data. The model can capture fine-grained emotional patterns that unimodal systems tend to miss due to combining separate modality-specific models and a fusion component. Due to their state-of-the-art capacity to perceive semantic nuances and contextual relationships, pre-trained models like BERT and DistilBERT were utilized for text processing [26].

The models were fine-tuned for domain-specific use to enhance their performance on datasets such as GoEmotions. Pre-trained models such as VGG-16 and ResNet-50 were employed for image data to leverage their robust feature extraction ability, particularly for facial expression recognition [27]. Custom CNN and LSTM models were also created to strike a compromise between computational effectiveness and flexibility to specific data properties. To improve domain-specific performance, hybrid configurations were created by fusing the advantages of pre-trained models with unique designs.

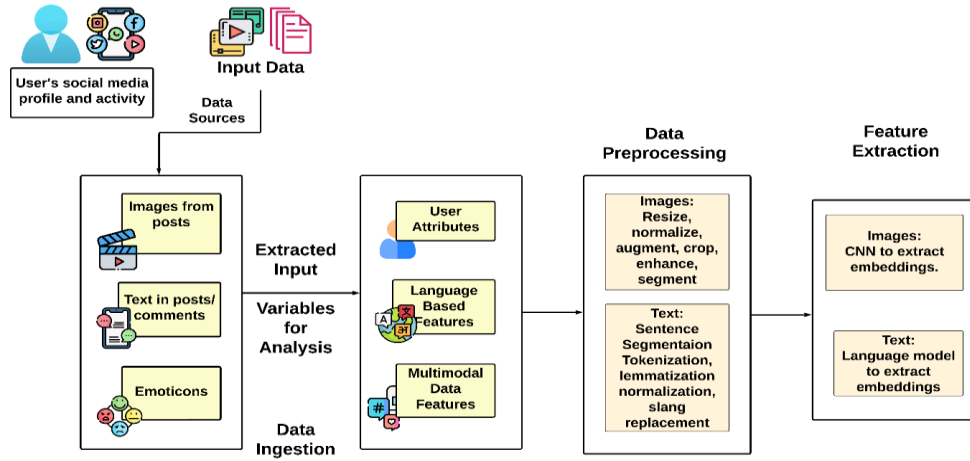


Figure 1. Proposed system architecture-phase 1

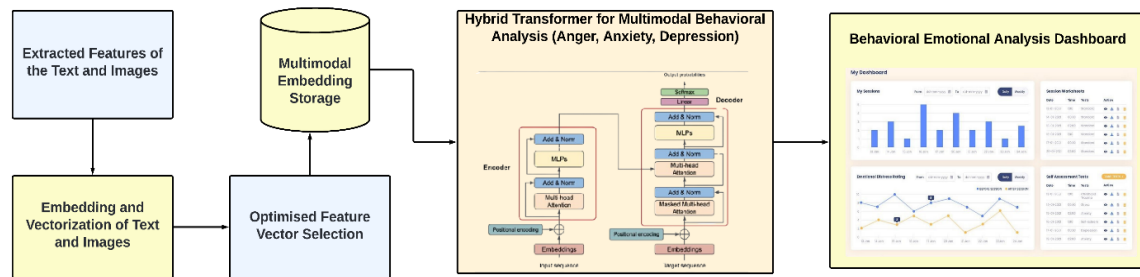


Figure 2. Proposed system architecture-phase 2

Figure 2 illustrates the second stage, in which processed inputs are input to specialized model streams: an LSTM-based stream for sequential text data, a CNN-based stream for visual input, and a symbolic mapping stream for emoticons. These streams' outputs are input to a multimodal fusion layer, where attention-based late fusion is performed to resolve conflicts across modalities and dynamically weigh contributions from every stream. Our fusion layer implements transformer-based scaled dot-product attention. Given modality-specific query Q , key K , and value V matrices, we compute as in (1).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Where d_k is the dimensionality of the key vectors. We employ multi-head attention with h parallel heads—concatenating and linearly projecting their outputs—to allow the model to jointly attend to information from different representation subspaces [28]. This combination mechanism ensures that every modality's high-level semantic features are aligned contextually before final classification. Currently, when conflicting emotion scores occur for a text, vision, and emoticon stream, EmoVibe calculates their final prediction as the average of the three modality scores, while maintaining equal weight for each score. Although this approach prevents over- or under-emphasizing any one contribution, it may minimize using the most dependable modality under particular circumstances.

As the system's user interface, the dashboard module combines data insights into an attractive design that displays emotional shifts and real-time behavioral trends. In addition, the dashboard can track metrics across specific timeframes like daily, weekly, or monthly. For consumers and mental health experts, tools like line charts, bar graphs, and trend lines depict emotional states across time, making them an indispensable decision-support tool.

Since mental health information is personal, the architecture has robust privacy and security controls to protect user information. Anonymization systems ensure data privacy by eliminating personal identifiers during collection [29]. Role-based access controls limit access to data to the approved users, while encryption

secures data when it is transmitted and stored. Aside from offering technical safeguards, the system addresses broader ethical concerns regarding its usage. The approach employs advanced preprocessing methods that ensure sound data handling to avert misinterpretation of diligently selected content or inflated bias. By shunning an over-reliance on any modality, modal contextual analysis decreases the likelihood of misclassification. An ethical and transparent environment for dealing with sensitive mental health information is established by honest and ethical standards maintained through open user consent processes [30].

A systematic evaluation procedure is applied to the framework to ensure its reliability and strength. A stratified training-validation-testing ratio ensures proper representation of all emotional classes, and cross-validation is employed to prevent overfitting and provide balanced model evaluations. Confusion matrices are analyzed to identify and correct common misclassifications. To confirm the system's usefulness in actual environments, its real-time performance is also tested in different data loads, measuring latency and scalability [31].

The design and application of the framework rest on some assumptions. The framework assumes publicly available social media data can inform users' mental health statuses, as user-generated content carries useful information about emotional health. The system also assumes access to high-performance computing resources in order to deal with large multimodal datasets' training and real-time deployment. The project developed on an HP Z8 GPU workstation with two AMD Radeon PRO W6800 GPUS and 32 GB RAM to accelerate training. The system presumes that multimodal data improves detection of subtle mental health markers by conveying more information about emotions than any single modality. Lastly, the system depends on strict privacy regulations to guarantee moral compliance without sacrificing openness in data processing. The warning thresholds of the dashboard and personalization features enable experts and users to customize their experience.

4. IMPLEMENTATION DETAILS

The proposed multimodal framework integrates various data modalities and late fusion is employed in the multimodal fusion process to preserve modality-specific features, allowing text, images, and emoticons to be processed separately before their outputs are merged [32]. Late fusion was chosen because it prevents feature dilution and makes more accurate and subtle emotional predictions [33]. LSTMs are good at dealing with contextual and sequential data their use for text analysis was supported. However, the BERT's high computing needs made it impractical for real-time use. As CNNs work effectively with small datasets, they were selected for image analysis. Normalization in text processing is taken care of by natural language toolkit (NLTK) and TensorFlow. OpenCV transforms images and maps emoticons to categories of sentiments. Using parallel processing and highly optimized pipeline designs, the system reduces latency even when working with large volumes of data. The implemented system is organized as follows.

4.1. Data collection and preparation

Regarding dataset selection, the framework employs datasets like GoEmotions [34] for text-based emotion labels and FER [35]/AffectNet [36] for image-based emotion recognition. The GoEmotions dataset, developed by Google Research, includes text data that captures complex emotional nuances. The dataset distribution and examples are illustrated in Figure 3.

•cant escape •anxiety is something you cant escape. You wish you could but you really cant 🙄
#cantstop #anxious #anxiety #anxietyquotes #cantescape

Figure 3. Text and emoticon data sample

The FER and AffectNet datasets contain images of facial expressions corresponding to various emotional states. Figure 4, which appears as follow, provides representative samples of facial expressions. The Figures 4(a) to 4(f)—anger, digested, fearful, happy, neutral, and sad, present the range of emotions analyzed, highlighting the diversity of emotional expressions captured via the subfigures. The framework further examines emotions from a multidimensional perspective, making it more sensitive to subtle emotional cues within content generated on social media. To achieve dataset efficiency, it's sorted by emotion and modality, with images categorized by tagged expressions and text files by emotion which helps in rapid data access when the model is being trained and tested. The system updates the data with the latest relevant material in social media in real-time, keeping pace with changing patterns of language usage, slang, and visual modes of expression. Time-based relevance entails regular updating of data representing real-time

emotional states. It is capable of handling the most recent trends of user activities to remain quick to enable useful monitoring in real time. Figure 4 presents a visual sample from these datasets, highlighting the diversity of emotional expressions captured via Figures 4(a) to 4(f).

In this research, all social media data employed were gathered from publicly accessible sources by the platform's terms of service. To ensure ethical considerations, we ensured data use complied with relevant guidelines for responsible AI research, such as respect for user privacy and consent where necessary. In addition, we tested for language, geography, and demographic representation biases, and we will continue to reduce such biases in subsequent work through more balanced and diverse data sampling approaches.



Figure 4. Sample image dataset labels showing emotions through facial expressions of (a) anger, (b) disgusted, (c) fearful, (d) happy, (e) neutral, and (f) sad

4.2. Data preprocessing

Special preprocessing is needed for each type of data. The Keras tokenizer is used to convert words into integer indices. Further processing sanitizes the text, eliminating special characters and stop words. Sequences are padded to a predetermined length of 100 tokens for normalizing input shapes for processing in batches, optimizing computing efficiency, and enhancing model performance. Picture data is preprocessed by resizing images to 150×150 pixels for custom models and 224×224 pixels for ResNet-50 and VGG-16. Scaling pixel values between 0 and 1 enhances gradient descent efficiency and model convergence. Various data augmentation techniques improve generalization and strengthen the dataset, allowing optimal model performance in real-world applications. Emoticons are necessary markers of emotional content. To adequately express their meaning, they are associated with pre-defined sentiment tags. For instance, the "😊" emoticon is used for happiness, while the "😞" emoji is used for melancholy. By providing context to the text data, this mapping allows the algorithm to consider cases where emoticons reinforce or refute the sentiment expressed in the text. The system offers a more complete and intricate emotional profile by including emoticons in the emotional analysis, significantly enhancing its ability to interpret social media user expressions.

4.3. Model architecture development

The model structure separates text and image processing before fusion to extract useful patterns by each submodel, and the fusion layer merges insights to produce the overall output. The text model employs LSTM layers for sequence data. A 128-unit LSTM learns sentence word relationships as is optimal for recognizing mood changes and emotional tone because it can maintain long text sequences. A dense layer with ReLU activation boosts interpretative power of weak emotional signals. Output is projected to depression, anxiety, anger, and neutral categories. A 3-layer CNN model with increasing filter sizes (32, 64, 128) and MaxPooling layers is used for spatial information. It produces hierarchical facial expression features from edges to larger structures like smiles or frowns. A dense layer with ReLU activation fine-tunes the features, resulting in data for the final classification layer, producing three emotion classes: happy, sad, and angry. The hybrid BERT+LSTM architecture combined the sequential processing power of LSTM with the contextual advantages of BERT. Figure 5 illustrates the hybrid text-processing model, which uses BERT embeddings and an LSTM layer to learn semantic richness and sequential dependencies.

To adjust the features to the particular domain requirements of the FER/AffectNet datasets, the ResNet-50+CNN hybrid model first extracted high-level features from picture data using ResNet-50. These features were then sent via a bespoke CNN layer. Using pre-trained information and adjusting to domain-specific tasks were balanced by these hybrid arrangements. Figure 6 illustrates the hybrid image-processing model that combines ResNet-50's pre-trained features with a custom CNN head to improve facial emotion detection accuracy.

The multimodal fusion layer triggers a hybrid transfer learning process to combine the image and text model outputs thereby improving understanding across the modalities. Additionally, the fusion layer improves interpretability through combining text, images, and emoticon data into a single entity, thus providing a more integrated explanation of emotional states that is useful in distinguishing complex emotional patterns. An attention layer dynamically balances features resolving conflicts arising when modalities present conflicting views while highlighting contextually relevant features. The attention mechanism supports the system in handling complex emotional states, thus ensuring fair predictions across all the modalities. The multimodal fusion architecture of Figure 7 fuses outputs of the text and image models (as well as sentiment-mapped emoticons) based on attention-based late fusion. This architecture enables the system to weigh inputs adaptively based on contextual relevance, which is most effective in conflicting emotional signals between modalities.

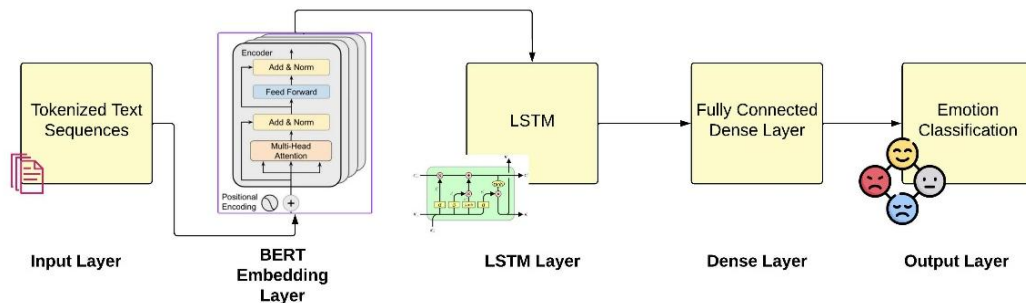


Figure 5. Hybrid text-based model to determine and map emotion from text

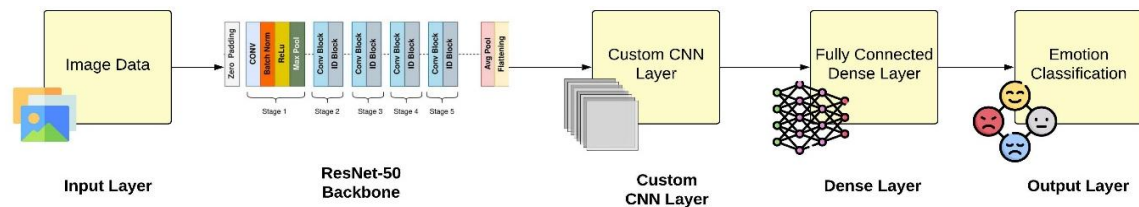


Figure 6. Hybrid image-based model for facial emotion detection and mapping

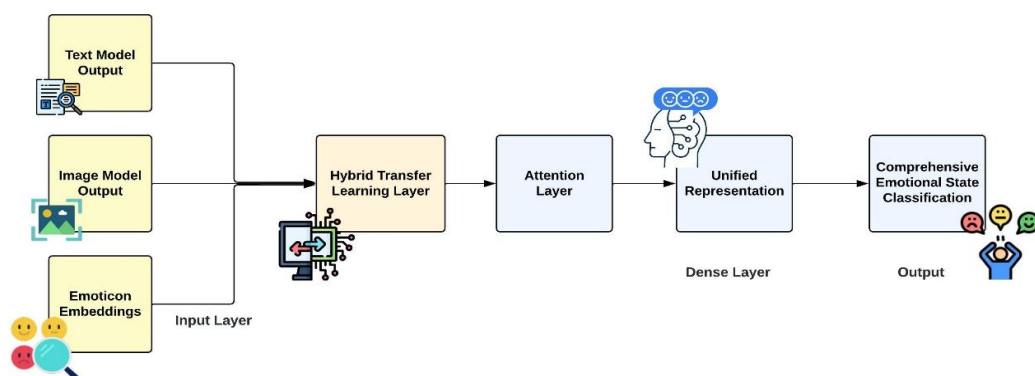


Figure 7. Proposed multimodal fusion model to combine textual, image and emoticon model outputs

4.4. Model training and evaluation

Training and testing are in phases. Initially, all modality-specific models are trained independently. Later, they are collectively refined through fusion layer. The text model is supervised fine-tuned for 16 epochs with AdamW optimizer, initial learning rate 2×10^{-5} , weight decay 0.01, batch size 64, dropout 0.1, and validation split 20%. Early stopping with 3 epochs of no improvement. The image-based models are trained for up to 32 epochs under the same conditions, initial learning rate of optimizer 1×10^{-4} , batch size 32, dropout 0.1, validation split 20%, and early stopping at 3 static epochs. High-level embeddings are concatenated after every backbone checkpoint in the late-fusion layer. Joint fine-tuning is conducted for 10 epochs using AdamW optimizer, learning rate 1×10^{-5} , weight decay 0.01, dropout 0.1, and validated after 2 epochs with no improvement. Model performance is measured and supplemented with tracking using appropriate metrics like mean absolute error, categorical cross-entropy loss, accuracy and confusion matrices. This setup achieves industry-standard resource utilization and adaptive model convergence for multimodal emotion analysis.

4.5. Dashboard integration and real-time analysis

SMED offers lay users and mental health practitioners' immediate access to emotional trends. The dashboard displays emotional data in various charts to identify emotional patterns and necessary interventions by practitioners. Figure 8 presents the SMED, the visual interface layer. It offers real-time visualizations using line graphs, emotion distribution charts, and alert systems for significant emotional shifts. The dashboard's architecture supports real-time updates, making it a practical tool for clinicians and end-users. This dashboard also includes alert mechanisms that may signal significant emotional changes on some predefined thresholds. For example, suppose the user's emotional state changes rapidly. If they have a significant increase in their anxiety or depression levels, the system alerts them almost immediately so that they can seek adequate support or intervention in good time.

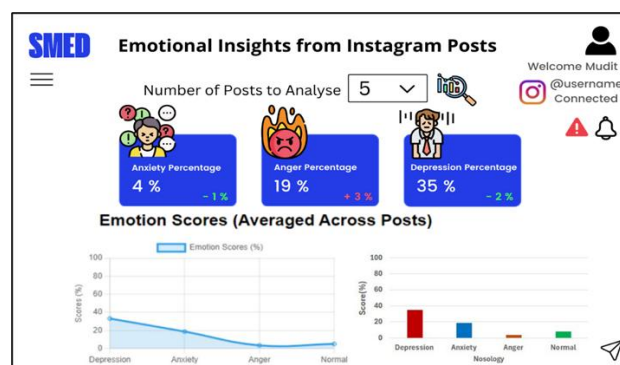


Figure 8. SMED prototype

5. RESULTS AND DISCUSSION

Two of the pre-trained models, BERT and DistilBERT, were employed for text data analysis, and an LSTM model as well as a hybrid model using an LSTM architecture with BERT embeddings. BERT performed outstanding semantic understanding and contextual nuances on the GoEmotions dataset, with an accuracy of 87.4%. DistilBERT was slightly lower, with an accuracy of 85.9%, but with less computational resources. The LSTM model in-house had robust sequential processing capabilities, achieving an accuracy of 83.6%, but was not capable of detecting complex semantic relationships. Blending sequence modelling with pre-trained embeddings, the hybrid architecture of BERT embeddings coupled with an LSTM, enhanced performance, leading to a final accuracy of 88.2%.

ResNet-50 and VGG-16 were used to train the dataset, and a CNN layer and a custom CNN model were experimented with. VGG-16, with its plain architecture, fared poorly with 82.9% accuracy on the FER dataset, with ResNet-50, at 84.2%. The custom CNN optimally weighed feature extraction and efficiency at 81.5%. The hybrid ResNet-50+CNN configuration, based on pre-trained features and domain adaptation, yielded 85.1%. A late fusion method dynamically combines multiple models and achieves better performance. Contextual attention networks balance modalities and settle disputes and report a precision of 89.7%, higher than early and individual fusion models (78.4%). The performance metrics for each configuration on the FER and GoEmotions datasets are compiled in Table 2, and the graphical representation is in Figure 9.

Table 2. Model performance on GoEmotions and FER datasets

Model	Modality	Accuracy (%)	Precision	Recall	F1-score	Dataset
Pre-trained BERT	Text	87.4	0.85	0.6	0.85	GoEmotions
Pre-trained DistilBERT	Text	85.9	0.83	0.5	0.84	GoEmotions
Custom LSTM	Text	83.6	0.81	0.2	0.81	GoEmotions
Hybrid BERT+LSTM	Text	88.2	0.87	0.8	0.87	GoEmotions
Pre-trained ResNet-50	Image	84.2	0.82	0.3	0.82	FER
Pre-trained VGG16	Image	82.9	0.8	0.1	0.8	FER
Custom CNN	Image	81.5	0.78	0.79	0.79	FER
Hybrid ResNet-50+CNN	Image	85.1	0.84	0.85	0.84	FER
Early fusion	Multimodal	78.4	0.77	0.76	0.76	Fused Dataset: GoEmotions and FER
Custom late fusion (Proposed)	Multimodal	89.7	0.88	0.9	0.89	Fused Dataset: GoEmotions and FER

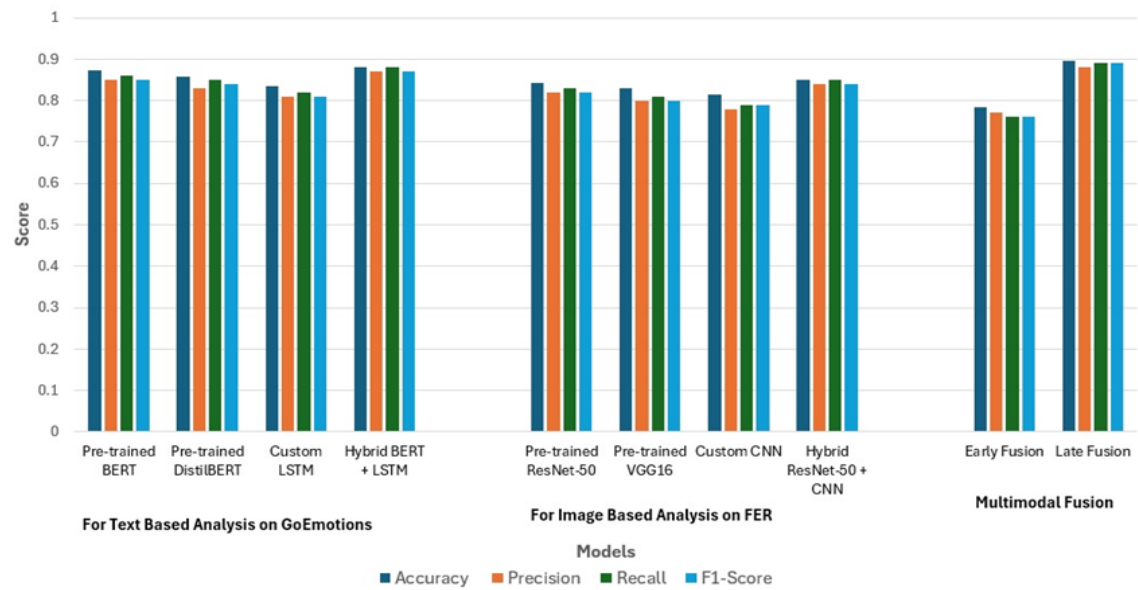


Figure 9. Comparative performance of models on GoEmotions and FER datasets

To assess generalizability, the framework was evaluated on a second dataset, including AffectNet for images and GoEmotions for text. Consistent patterns emerged in the results, with the late fusion model continuing to perform better. The assessment measures for this dataset are shown in Table 3, and its graphical representation is in Figure 10.

Table 3. Model performance on GoEmotions and AffectNet datasets

Model	Modality	Accuracy (%)	Precision	Recall	F1-Score	Dataset
Pre-trained BERT	Text	87.4	0.85	0.86	0.85	GoEmotions
Pre-trained DistilBERT	Text	85.9	0.83	0.85	0.84	GoEmotions
Custom LSTM	Text	83.6	0.81	0.82	0.81	GoEmotions
Hybrid BERT+LSTM	Text	88.2	0.87	0.88	0.87	GoEmotions
Pre-trained ResNet-50	Image	82.3	0.81	0.82	0.81	AffectNet
Pre-trained VGG-16	Image	81.2	0.79	0.8	0.79	AffectNet
Custom CNN	Image	79.8	0.76	0.77	0.77	AffectNet
Hybrid ResNet-50+CNN	Image	83.4	0.82	0.83	0.82	AffectNet
Early fusion	Multimodal	76.5	0.75	0.74	0.75	Fused Dataset: GoEmotions and AffectNet
Custom late fusion (Proposed)	Multimodal	88.9	0.87	0.88	0.87	Fused Dataset: GoEmotions and AffectNet

SMED is also useful in interpreting these results. The late fusion model allows an actionable, relevant, and accurate forecast by removing technical output-versus-real-world-applications mismatches. It registers features between modalities dynamically, boosted by pre-trained and task-specific models, and performs well consistently. The method gathers complementary and contextually related information from

every modality to perform robustly on different datasets and includes SMED. The framework has real-time and scalable mental health monitoring potential. We acknowledge that EmoVibe exhibits limitations in accurately detecting emotions in posts with sarcasm or irony. These cases often involve a mismatch between literal word meanings and intended sentiment, leading to misclassification. Future model iterations could incorporate context-aware mechanisms such as pretrained sarcasm detection modules or transformer-based architectures fine-tuned on sarcasm-annotated datasets to address this. Integrating user-specific historical data and conversational context may also improve disambiguation in such nuanced emotional expressions.

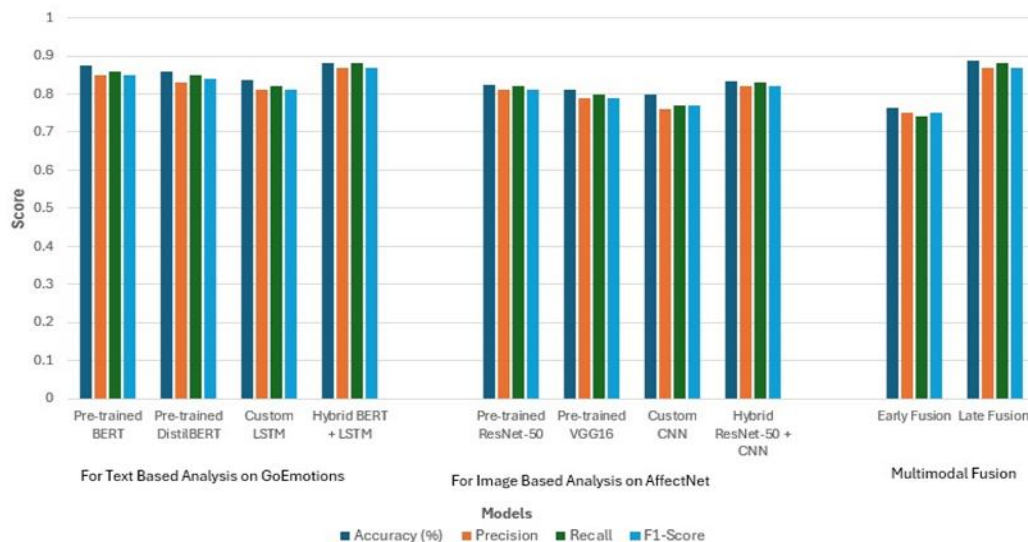


Figure 10. Comparative performance of models on GoEmotions and AffectNet datasets

6. CONCLUSION

This paper introduces EmoVibe, a robust multimodal AI framework that combines text, images, and emoticons to provide subtle and real-time analysis of emotion on social media. EmoVibe is able to address significant limitations of existing unimodal and early fusion strategies, and offers a richer interpretation of emotional expressions through a hybrid transfer learning framework and late fusion strategy. Model generalizability and effectiveness were demonstrated because of its higher accuracy (89.7% on different datasets such as GoEmotions, FER, and AffectNet). The social media SMED developed within the framework of this work also offers an interactive and real-time interface to users and mental health professionals to track the emotional trends, thus enabling early and direct actions. Despite the framework performing well, cultural bias in the dataset used, ethical concerns, and the limitation of access to computational resources can be considered a challenge. The future studies will focus on adopting additional modalities (e.g., audio and video), creating culturally adaptive models, and integrating privacy-preserving AI techniques like federated learning to increase user confidence. EmoVibe provides a scalable, ethical, and operational solution to proactive mental health screening in the age of digital technology through the design of multimodal AI systems to address mental health issues.

7. FUTURE WORK

By establishing such vast data sources and model flexibility in various scenarios, more work for this multimodal AI system can be extended. The framework can be extended by incorporating multimodal inputs like audio (tone, pitch, and prosody) and short video clips. This extension will enable the model to capture a more diverse range of emotional and contextual information, thus enhancing the robustness and accuracy of the model's comprehension in dynamic, real-world environments. A fuller picture of mental health could be gained by extending the system to include offline data or at least indicators beyond social media. Creating multilingual and culturally sensitive models also needs to be enhanced for the framework to operate effectively and appropriately across different linguistic and cultural clusters to improve its utility and effectiveness across the globe. Future enhancements will also explore dynamic, confidence-aware fusion—adjusting attention weights according to per-modality reliability—and adaptive attention mechanisms to boost performance when modalities are incongruent or ambiguous. Additionally, we will incorporate cross-cultural corpora reflecting diverse emotional norms to mitigate cultural bias in emotion recognition further.

Computational efficacy and real-time processing will also be essential to scale up to large datasets and more complex data streams. The user data confidentiality and trust in applications as sensitive as mental health assessment will be enhanced by more effort in developing responsible AI practices, such as advanced privacy-preserving techniques. These will improve the use of the framework as a critical tool for individual and professional users of mental health to gain ethical, accessible, and nuanced mental health knowledge.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Deepali Vora	✓	✓		✓				✓	✓	✓	✓	✓	✓	✓
Aryan Sharma		✓	✓	✓	✓	✓		✓	✓	✓	✓			
Mudit Garg		✓	✓	✓	✓		✓	✓	✓	✓	✓			
Steve Francis		✓	✓	✓		✓	✓		✓	✓	✓			

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The author states there is no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.




REFERENCES

- [1] J. Hancock, S. X. Liu, M. Luo, and H. Mieczkowski, "Psychological well-being and social media use: A meta-analysis of associations between social media use and depression, anxiety, loneliness, eudaimonic, hedonic and social well-being," *SSRN Electronic Journal*, Mar. 2022, doi: 10.2139/ssrn.4053961.
- [2] WHO, "Mental health," *World Health Organization*. 2024. Accessed: Nov. 18, 2024. [Online]. Available: https://www.who.int/health-topics/mental-health#tab=tab_1
- [3] S. Dattani, L. R. Guirao, H. Ritchie, and M. Roser, "Mental health," *Our World in Data*, 2024. Accessed: Nov. 18, 2024. [Online]. Available: <https://ourworldindata.org/mental-health>
- [4] H. Wang, P. Miao, H. Jia, and K. Lai, "The dark side of upward social comparison for social media users: an investigation of fear of missing out and digital hoarding behavior," *Social Media + Society*, vol. 9, no. 1, Jan. 2023, doi: 10.1177/20563051221150420.
- [5] R. Beniwal and P. Saraswat, "A hybrid BERT-CPSO model for multi-class depression detection using pure Hindi and Hinglish multimodal data on social media," *Computers & Electrical Engineering*, vol. 120, Dec. 2024, doi: 10.1016/j.compeleceng.2024.109786.
- [6] A. Radwan, M. Amameh, H. Alawneh, H. I. Ashqar, A. AlSobeh, and A. A. A. R. Magableh, "Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis," *International Journal of Web Services Research*, vol. 21, no. 1, pp. 1–22, Feb. 2024, doi: 10.4018/IJWSR.338222.
- [7] L. P. Hung and S. Alias, "Beyond sentiment analysis: a review of recent trends in text based sentiment analysis and emotion detection," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27, no. 1, pp. 84–95, 2023, doi: 10.20965/jaciii.2023.p0084.
- [8] H. S. Gavalan, M. N. Rastgoo, and B. Nakisa, "A BERT-based summarization approach for depression detection," *arXiv:2409.08483*, 2024.
- [9] M. Ajith and others, "A deep learning approach for mental health quality prediction using functional network connectivity and assessment data," *Brain Imaging and Behavior*, vol. 18, no. 3, pp. 630–645, Feb. 2024, doi: 10.1007/s11682-024-00857-y.
- [10] K. Woodward, E. Kanjo, and A. Tsanas, "Combining deep learning with signal-image encoding for multi-modal mental wellbeing classification," *ACM Transactions on Computing for Healthcare*, vol. 5, no. 1, pp. 1–23, Jan. 2024, doi: 10.1145/3631618.
- [11] F. Limami, B. Hdioud, and R. O. H. Thami, "Contextual emotion detection in images using deep learning," *Frontiers in Artificial Intelligence*, vol. 7, pp. 1386753, Jun. 2024, doi: 10.3389/frai.2024.1386753.
- [12] V. Tejaswini, K. S. Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, pp. 1–20, Jan. 2024, doi: 10.1145/3569580.





- [13] M. S. Rahmadiannisa, F. S. Firdaus, I. M. Manaf, and K. W. P. Putra, "Dopamind+: mobile-based application to prevent and treat mental health disorders in adolescents," *International Journal of Software Engineering and Computer Science*, vol. 4, no. 1, pp. 321–338, Apr. 2024, doi: 10.35870/ijsecs.v4i1.2391.
- [14] M. K. A. Das and J. A. Hughes, "Exploring BERT-based classification models for detecting phobia subtypes: a novel tweet dataset and comparative analysis," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024.
- [15] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's just a matter of time: detecting depression with time-enriched multimodal transformers," *Advances in Information Retrieval: 45th European Conference on Information Retrieval*, 2023, pp. 200–215, doi: 10.1007/978-3-031-28244-7_1.
- [16] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches," *Sensors*, vol. 24, no. 2, Jan. 2024, doi: 10.3390/s24020348.
- [17] A. Alsiaity and R. Orji, "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions," *Behaviour & Information Technology*, vol. 43, no. 1, pp. 139–164, Jan. 2024, doi: 10.1080/0144929X.2022.2156387.
- [18] M. Garg, "Mental health analysis in social media posts: a survey," *Archives of Computational Methods in Engineering*, vol. 30, no. 3, pp. 1819–1842, Apr. 2023, doi: 10.1007/s11831-022-09863-z.
- [19] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "MentaLLaMA: interpretable mental health analysis on social media with large language models," in *Proceedings of the ACM Web Conference 2024*, Singapore, Singapore: ACM, May 2024, pp. 4489–4500. doi: 10.1145/3589334.3648137.
- [20] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 12113–12132, 2023, doi: 10.1109/TPAMI.2023.3275156.
- [21] A. H. Yazdavar *et al.*, "Multimodal mental health analysis in social media," *PLOS ONE*, vol. 15, no. 4, Apr. 2020, doi: 10.1371/journal.pone.0226248.
- [22] J. D. Haltigan, T. M. Pringsheim, and G. Rajkumar, "Social media as an incubator of personality and behavioral psychopathology: symptom and disorder authenticity or psychosomatic social contagion?," *Comprehensive Psychiatry*, vol. 121, Feb. 2023, doi: 10.1016/j.comppsy.2022.152362.
- [23] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748–8763.
- [24] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, pp. 23716–23736.
- [25] OpenAI *et al.*, "GPT-4 technical report," *arXiv:2303.08774*, 2023.
- [26] J. Oh, M. Kim, H. Park, and H. Oh, "Are you depressed? analyze user utterances to detect depressive emotions using DistilBERT," *Applied Sciences*, vol. 13, no. 10, May 2023, doi: 10.3390/app13106223.
- [27] S. Lu, Y. Ding, M. Liu, Z. Yin, L. Yin, and W. Zheng, "Multiscale feature extraction and fusion of image and text in VQA," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, Apr. 2023, doi: 10.1007/s44196-023-00233-6.
- [28] A. Vaswani *et al.*, "Attention is all you need," *31st Conference on Neural Information Processing Systems*, 2017, pp. 1–11.
- [29] O. Vovk, G. Pihio, and P. Ross, "Methods and tools for healthcare data anonymization: a literature review," *International Journal of General Systems*, vol. 52, no. 3, pp. 326–342, Apr. 2023, doi: 10.1080/03081079.2023.2173749.
- [30] P. Wang and others, "OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," *Proceedings of the 39th International Conference on Machine Learning*, 2022, pp. 23318–23340.
- [31] A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "ConfusionVis: comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowledge-Based Systems*, vol. 247, Jul. 2022, doi: 10.1016/j.knosys.2022.108651.
- [32] L. M. Pereira, A. Salazar, and L. Vergara, "A comparative analysis of early and late fusion for the multimodal two-class problem," *IEEE Access*, vol. 11, pp. 84283–84300, 2023, doi: 10.1109/ACCESS.2023.3296098.
- [33] P. He *et al.*, "Domain-separated bottleneck attention fusion framework for multimodal emotion recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 4, pp. 1–21, Apr. 2025, doi: 10.1145/3711865.
- [34] D. Demszky, D. M.-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: a dataset of fine-grained emotions," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054, doi: 10.18653/v1/2020.acl-main.372.
- [35] I. J. Goodfellow *et al.*, "Challenges in representation learning: a report on three machine learning contests," *International Conference on Neural Information Processing*, 2013, pp. 117–124, doi: 10.1007/978-3-642-42051-1_16.
- [36] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: a database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/TAFFC.2017.2740923.

BIOGRAPHIES OF AUTHORS







Dr. Deepali Vora    completed her Ph.D. in Computer Science and Engineering from Amity University, Mumbai. Currently working as Professor, Computer Science, and Engineering, Symbiosis Institute of Technology Pune, Symbiosis International University (Deemed), Pune, India. She has more than 25 years of experience in total in teaching, research, and industry. She has published more than 75 research papers in reputed national and international conferences and journals. She has co-authored four books and numerous book chapters and delivered various talks in data science and machine learning. She received grants from government bodies such as DST, AICTE, ISTE and industry. She is acting as a reviewer for many international conferences and journals like IEEE Access, IGI Global, Springer, and Inderscience. She has organized many value-added courses for the benefit of the students. More than 20 students have completed their post-graduate studies under her guidance from Mumbai University. In addition to that, eight students are pursuing research (Ph.D.) under her guidance at Symbiosis International University, Pune. Her course developed on deep learning is currently available on the unschool platform for all, and two technical blogs are available on the KnowledgeHut.com website. She can be contacted at email: deepali.vora11@gmail.com.







Aryan Sharma     is a Computer Science graduate with honors in Artificial Intelligence and Machine Learning from Symbiosis Institute of Technology, Pune. The Semester Exchange at the University of Liverpool brought him to work under faculty supervisors in the Summer of 2024. His research fields include AI, ML, and neural networks, through which he conducted projects focusing on early risk prediction during his semester exchange period, together with anomalous traffic detection, glaucoma detection, and social media emotion analysis. He has explored generative AI as well as natural language processing and deep learning applications through his research interests. At the Symbiosis Centre for Applied AI, he performed research on generative AI applications for legal documents and summaries as a research intern. He is currently an AI Developer Intern at Inventronics, contributing to cutting-edge AI solutions. He has continuously developed AI-driven projects that span healthcare and cybersecurity systems and NLP solutions during his education and work experience, including a mental health chatbot based on retrieval-augmented generation and a glaucoma detection system built with deep learning models. He has certifications in deep learning, NLP, and AI from Stanford University and deep learning. He can be contacted at email: 0dark30coc@gmail.com.



Mudit Garg     is pursuing a B.Tech. in Computer Science and Engineering with honors in AIML from Symbiosis Institute of Technology, Pune. He was also a research intern at SCAAI. He has worked extensively on multiple AI-driven projects, leveraging machine learning, natural language processing, and deep learning techniques to solve real-world problems. His projects span multimodal learning, accessibility, and AI-driven automation, including sight beyond sight, an AI-powered website enhancing content accessibility for visually impaired users, people's lens, a face recognition system for identifying domain experts at meetups, and CropPal, a deep learning-based crop identification and growth stage detection tool for precision agriculture. He also developed a RAG-based multimodal and multilingual chatbot for domain-specific AI conversations. Mudit has demonstrated his innovation and problem-solving approach by winning multiple hackathons, including _VOIS Innovation Marathon 2022, MCCIA Edu-Fest 2022, and IEEE India Council Hack 2.0. His research interests include generative AI, large language models (LLMs) and vision-language models (VLMs), multimodal and multilingual AI, computer vision and AI for good. His research achievements include the best paper award at IEEE ICACTA 2023 and a published article in Bharat @75: Multiple dimensions to empower India for a better tomorrow. He can be contacted at email: gargmudit2708@gmail.com.



Steve Francis     is pursuing a B.Tech. in Electronics and Telecommunication Engineering with Minors in AI/ML from Symbiosis Institute of Technology, Pune. He worked as a research intern at the University of Waterloo under the Mitacs Globalink Research Internship Program and contributed to algal density estimation using MATLAB image processing, collaborating with renowned researchers and peers. His research interests include image processing, AI/ML, embedded systems, signal processing, and applying engineering concepts to solve real-world challenges. He has worked on several multidisciplinary and branch-centric projects, publishing two conference papers focusing on power system design for efficient energy transmission without physical connectors, optimizing power efficiency, signal alignment, and system stability for electric vehicle charging. He also used regression techniques to analyze large-scale environmental datasets, predicting trends in temperature, rainfall, and carbon emissions. During his studies, he gained industry experience through various internships and collaborations, developing expertise in Troubleshooting, MATLAB, Python, IoT, circuit analysis, and embedded systems. As part of his Minor in AI/ML, he explored AI-driven hand motion gesture detection, focusing on deep learning models for recognizing and classifying hand movements, with potential applications in human-computer interaction and assistive technologies. He can be contacted at email: stevenfrancis28@gmail.com.