MoodElevate: A Multimodal AI-Driven Mental Health Companion Using Text and Audio-Based Emotion Recognition

Piyush Gawali[1], Mansi Konde[2], Suchet Mahamuni[3], Akanksha Kuvhare[4]
[1,2,3,4] *Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India*

***Abstract*—MoodElevate is an AI-powered mental and emotional health application that helps users identify their emotional state, track mood patterns, and receive personalised support. The system takes text or voice as input, then applies machine learning models which classify emotions, and delivers customised content such as music, videos, and articles, to help users regulate their mood. A digital journalling feature analyses entries and gives mood-tracking reports. It uses DistilBERT model fine-tuned on the GoEmotions dataset achieved 85–90% accuracy (F1=0.86) for text-based emotion detection, a CNN–LSTM hybrid model trained on CREMA-D dataset showed 79% accuracy for speech-based detection. Multimodal fusion improved overall accuracy by 15%, proving the benefit of combining text and audio inputs for real-world affective computing.**

***Index Terms*—Emotion Detection, Emotional Wellness, Artificial Intelligence, Sentiment Analysis, Journalling System, Mood Tracking, DistilBERT, CNN–LSTM, Multimodal Fusion**

I. INTRODUCTION

Over recent years the importance of mental and emotional health has gained prominent attention. This heightened focus stems from the escalating pressures of academic, professional, and personal life, which frequently manifest as stress, anxiety, and emotional instability. Although a range of mental-health support services, counselling, psychotherapy, peer networks exist, many individuals do not utilise them due to accessibility barriers such as cost, stigma, and geographic availability. This gap has created demand for automated, accessible services that provide on-demand emotional support.

Progress in artificial intelligence and machine learning has given us various applications capable of analysing verbal and non-verbal inputs, speech, and written language to detect emotional states. NLP and deep-learning based systems have changed the method of interaction between humans and computers [1].

Digital journalling has also become a powerful practice: writing about daily experiences helps individuals understand their thought processing and identify repeating emotional patterns. Yet consistent journaling is difficult without motivation. Platforms that couple journalling with intelligent feedback both motivate users and yield structured emotional data amenable to further analysis.

Nevertheless, most current technologies have limited scope since few are capable of performing multimodal emotion detection or offering integrated reflective tools. MoodElevate addresses this gap through (i) text or voice inputs, (ii) classification of emotional states via task-specific ML models, (iii) personalised content recommendations, and (iv) a longitudinal mood-tracking dashboard through digital journalling. The application is designed to complement, not replace, professional mental-health care.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the system architecture. Section 4 details the methodology. Section 5 presents experimental results. Section 6 concludes with future directions.

II. LITERATURE REVIEW

A. Emotion Recognition in Human–Computer Interaction

In early emotion research within HCI, attempts were made to decipher emotional cues speech patterns and facial expressions to create more empathic human–computer interactions. Cowie et al. [2] established

theoretical foundations showing how intelligent systems capable of recognising and responding to affective signals could significantly improve user experience. Pantic and Rothkrantz [3] followed with comprehensive frameworks for automated facial expression analysis, providing taxonomies and benchmarks that have guided subsequent recognition systems.

B. Emotion Datasets for Machine Learning

High-quality annotated datasets are fundamental to robust emotion recognition. The GoEmotions corpus [4] dataset contains more than 58,000 Reddit comments classified in 27 emotion classes, helping NLP models to detect minor affecting factors as well. For multimodal speech and visual data, CREMA-D [5] dataset gives thousands of audio and video clips in multiple emotion categories, while RAVDESS [6] supports speech-emotion detection.

C. Deep Learning Models for Emotion Recognition

Convolutional Neural Networks (CNNs) are highly effective for image-based emotion detection, specially for facial expression analysis. Srivastava et al. [7] demonstrated the power of deep convolutional architectures for large-scale visual recognition, results that have since transferred to affective computing pipelines. For sequential data such as speech and text, Long Short-Term Memory (LSTM) networks [8] solve the vanishing-gradient problem and model long-range temporal dependencies. Transformer-based systems [1] advanced NLP through attention mechanisms. DistilBERT [9] became widely used due to its compact size and competitive language-understanding performance.

D. AI Systems for Emotional Well-being

The role of AI has evolved from emotion recognition to emotion support: systems now combine recognition with conversational agents and personalised feedback to provide accessible mental-health assistance. Journalling features, mood logs, and curated content recommendations have emerged within digital platforms as tools for facilitating emotional reflection. MoodElevate builds on this trajectory by integrating all these elements into a single multimodal pipeline.

E. Existing Applications and Research Gaps

Table 1 summarises key commercial applications. Dialogue-based AI companions such as Woebot [10] and Youper [11] offer conversational coaching grounded in cognitive-behavioural therapy (CBT). Woebot, Wysa, Youper, and Sanvello [12] provide daily mood journalling and reflection prompts, while platforms such as Calm and Headspace [13] offer audio content and guided exercises. Critically, none incorporates automated SER as a primary feature. MoodElevate addresses this gap by combining SER-based audio analysis with text-based NLP within a unified system.

Table 1. Comparison of Existing Mental-Health and Wellness Applications

Feature	Woebot	Wysa	Calm	MoodElevate
AI chat / coaching	✓	✓	–	✓
Mood journalling	✓	✓	–	✓
Audio content	–	–	✓	✓
Text emotion recognition	✓	–	–	✓
Speech emotion recognition	–	–	–	✓
Multimodal fusion	–	–	–	✓
Graphical mood trends	–	✓	–	✓

III. SYSTEM ARCHITECTURE

MoodElevate is built using four layers: a user interface, a backend service layer, a machine-learning inference layer, and a cloud database layer. Fig. 1 illustrates the overall structure.

User Interface Layer. A mobile application developed using Flutter allows users to provide text or audio input. The app also supports manual mood logging and digital journalling, sending all captured data to the backend for processing.

Backend Service Layer. Built using FastAPI and Python, this layer receives raw user input, processes it through the required model, and—based on the predicted emotional state either triggers the chatbot or activates the recommendation engine designed as per the user's interest. This layer also manages user authentication and session handling.

Machine-Learning Inference Layer. Text input is processed by a fine-tuned DistilBERT model; audio input is processed by a hybrid CNN–LSTM network.

When both inputs are available, a fusion step combines the two predictions to improve overall accuracy.

Database Layer. Firebase manages user profiles, mood logs, journal entries, and emotional history records. These records power the graphical mood-trend dashboard and enable visual analysis of a user's progress over time.

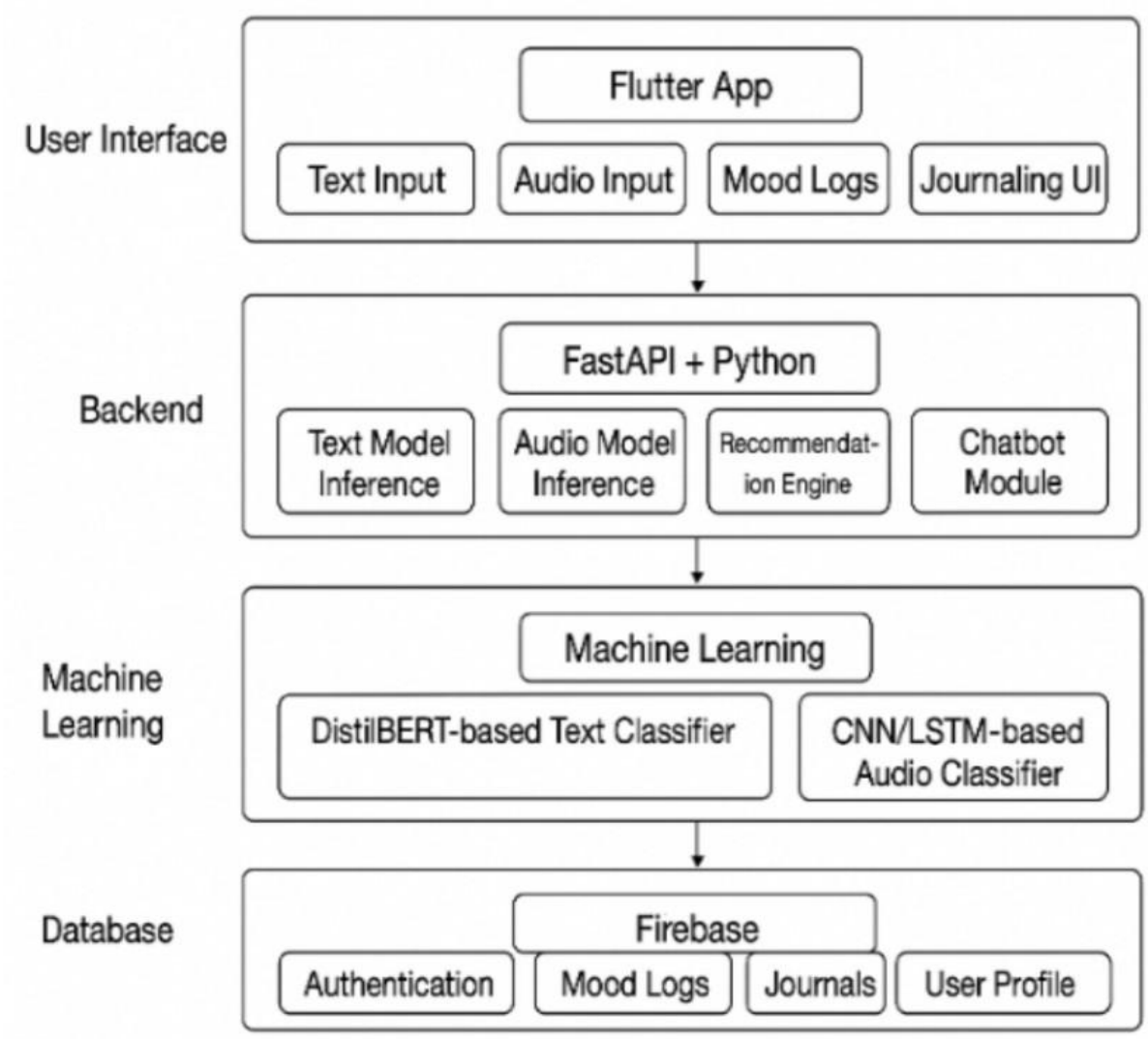

Figure 1. System architecture of MoodElevate.

IV. METHODOLOGY

The MoodElevate methodology integrates multimodal user input, ML-based emotion classification, and personalised response generation into a structured pipeline illustrated in Fig. 2.

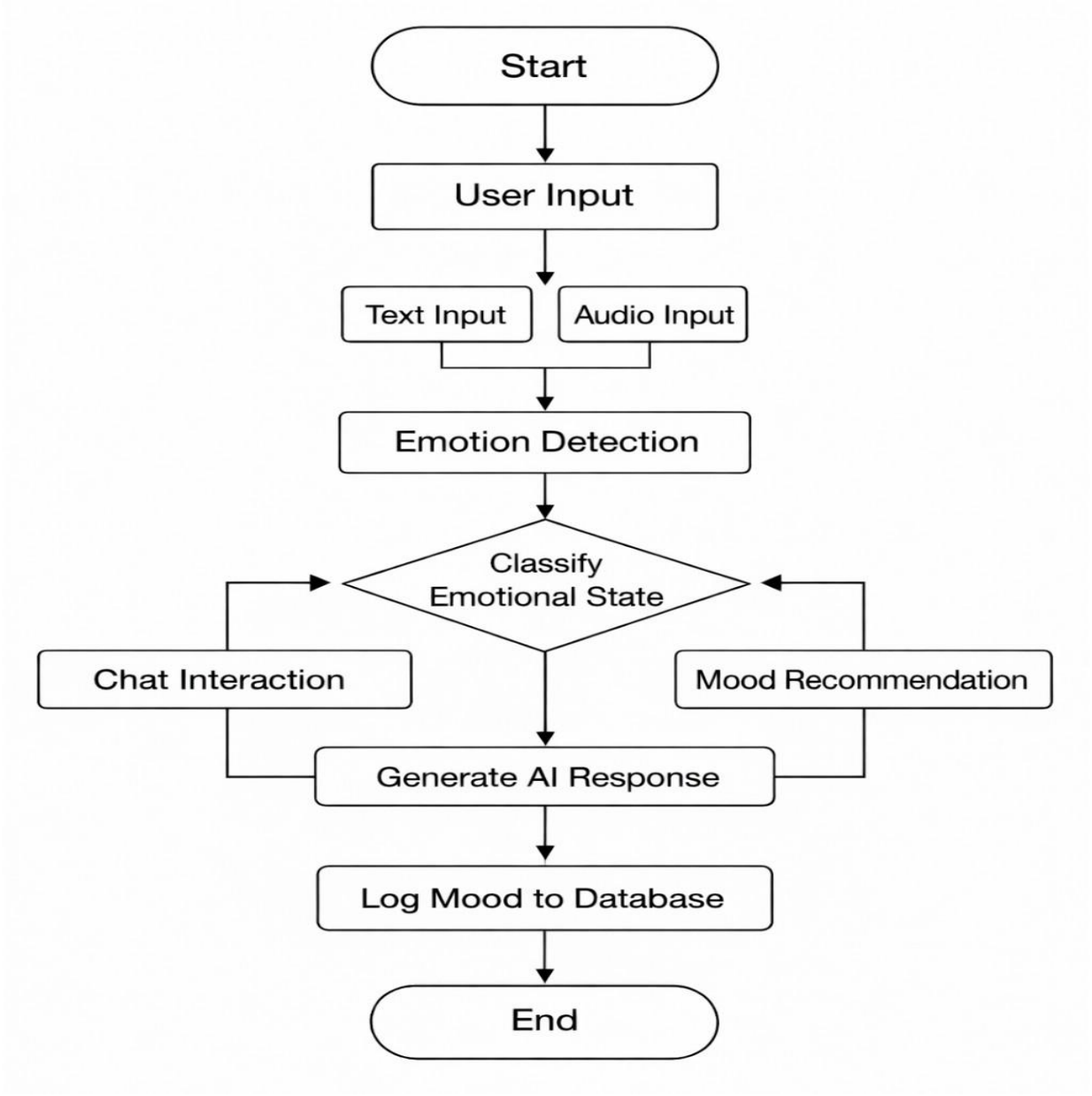

Figure 2. workflow of the emotion-detection and recommendation pipeline.

A. User Input Processing

The system accepts two input modalities. Text input is captured directly from the mobile keyboard interface. Audio input undergoes noise reduction and normalisation before feature extraction, where raw waveforms are converted to mel-frequency cepstral coefficient (MFCC) maps and 2-D mel-spectrograms for CNN-based spatial feature extraction.

B. Emotion Detection

Text-based emotion recognition. A DistilBERT model [9] is fine-tuned on the GoEmotions dataset [4] to classify input text into 27 fine-grained emotion categories. Input tokens are passed through the transformer encoder and a linear classification head, optimised with cross-entropy loss.

Audio-based emotion recognition. A hybrid CNN–LSTM architecture is trained on CREMA-D [5] and RAVDESS [6]. The CNN sub-network extracts spatial

features from mel-spectrogram frames, while the LSTM sub-network captures temporal dynamics. The combined result is fed to a softmax classifier to produce six main emotion classes: neutral, happy, sad, angry, fearful, and disgusted.

C. Emotion Fusion

When both modalities are simultaneously available, the system applies late (decision-level) fusion: softmax probability vectors from both models are averaged after label-space alignment, and the fused vector determines the final predicted emotion class. This approach is robust to partial modality failure.

D. Emotion Classification and Decision Logic

A rule-based default module determines the system response. Users with negative or distressed emotions are directed to the chatbot for empathetic, friend-like conversation support; users in a neutral or positive state are directed to the recommendation engine to further improve mood.

E. Response Generation and Recommendation Engine

The chatbot communicates in an empathetic tone using pre-defined template banks based on the detected emotion. The recommendation engine provides content as per user interests or by default rule-based logic: an anxious or stressed user receives calming music or guided meditation videos, while a low-mood user receives uplifting motivational content or articles. Content items are stored in a separate table indexed by emotion class.

F. Data Storage and Mood Tracking

All interactions, detected emotion labels, and journal entries are stored in the cloud database linked to the user profile. The system shows time-stamped mood graphs for users to observe affective results over days, weeks, or months.

G. Datasets and Model Training

Training followed a standard supervised pipeline. DistilBERT was fine-tuned for ten epochs with a learning rate of 2×10^{-5} and batch size of 32. Audio models were trained for 50 epochs with the Adam optimiser at a learning rate of 10^{-3}. Performance was evaluated using accuracy and macro-averaged F1-score. Pre-processing included text normalisation and tokenisation (text pipeline) and waveform normalisation (audio pipeline). Table 2 summarises the datasets used.

Table 2. Datasets Used for Training and Evaluation

Dataset	Modality	Size	Classes
GoEmotions [4]	Text	58,000+ comments	27 emotions
CREMA-D [5]	Audio	~7,400 clips	6 emotions
RAVDESS [6]	Audio	7,356 recordings	8 emotions

V. RESULTS AND DISCUSSION

A. Text Emotion Recognition Results

The DistilBERT-based model trained on the GoEmotions dataset achieved an accuracy in the 85–90% range and a macro-averaged F1-score of 0.86. The model performed better for emotions such as *joy*, *sadness*, and *neutral*, with minor confusion between similar categories (e.g., *anger* vs. *annoyance*). These results show that transformer-based models are effective for precise, fine-grained emotion classification.

B. Audio Emotion Recognition Results

The CNN–LSTM model trained on CREMA-D showed an accuracy of 79% on the test set. The model recognised *happy*, *sad*, and *neutral* speech confidently, but showed confusion between *anger* and *disgust* due to their similar toning. Adding training data from RAVDESS recordings improved generalisation.

C. Multimodal Performance

Fusion of text and audio predictions increased detection accuracy by 15% when both inputs were available, and reduced errors arising from ambiguous text or poor audio quality. This highlights the benefit of integrating multiple input modalities for effective mood detection.

D. Training Performance

Fig. 3 illustrates the training and validation accuracy curves for the DistilBERT text classifier, the CNN–LSTM audio model, and the multimodal fusion model across their training epochs. The DistilBERT model converges quickly within the first few epochs, reflecting the advantage of transfer learning from pre-trained weights. The CNN–LSTM model shows more gradual improvement, stabilising around epoch 35.

The fusion model consistently achieves the highest validation accuracy, confirming the 15% improvement. In all cases, validation accuracy closely tracks training accuracy.

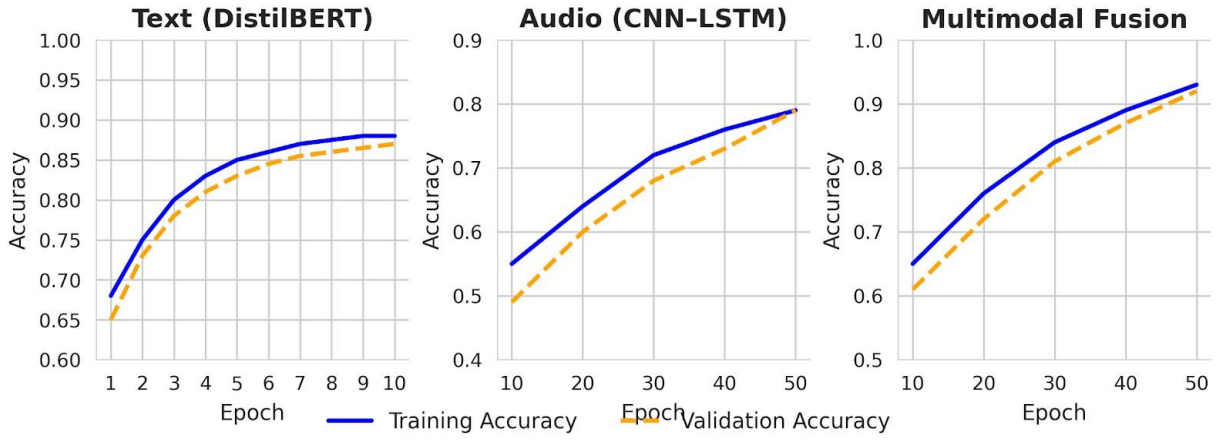

Figure 3. Training and validation accuracy curves for the DistilBERT text classifier, CNN–LSTM audio classifier, and multimodal fusion model

E. Discussion

MoodElevate's multimodal approach outperforms the unimodal text or audio pipelines found in existing commercial wellness applications. Limitations include English-only text support and audio sensitivity to background noise. These findings validate the proposed system for real-world emotional comprehension. Future directions include multilingual dataset integration and the incorporation of facial-expression recognition for tri-modal detection.

VI. CONCLUSION AND FUTURE WORK

This paper presented MoodElevate, an AI-driven multimodal mental-health companion that combines text and audio emotion recognition with a conversational chatbot, a personalised content recommendation engine, and a digital journalling system with graphical mood tracking.

The system employs a fine-tuned DistilBERT model for text-based emotion classification and a CNN–LSTM hybrid for speech-based recognition, with late fusion yielding consistent accuracy improvements over either unimodal approach. The recommendation engine promotes emotion regulation through curated music, video, and literature,
while the mood-tracking dashboard supports longitudinal self-reflection and self-regulation.

Compared with existing wellness applications, MoodElevate is the first to integrate automated SER alongside NLP-based emotion analysis within a unified reflective platform. Unlike systems that merely classify and respond to emotional states, MoodElevate actively encourages users to engage in structured self-regulatory behaviour.

Planned enhancements include (i) feature-level multimodal fusion, (ii) real-time adaptive personalisation, (iii) multilingual support, and (iv) integration of facial-expression analysis for tri-modal detection. Data privacy, model fairness, and ethical AI will remain central to all future development. MoodElevate demonstrates the potential of emotionally intelligent technology to make mental-wellness support more accessible, personalised, and effective.

REFERENCES

[1] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Transformers: state-of-the-art natural language processing," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, pp. 38–45, ACL, 2020.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[3] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, 2000.

[4] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: a dataset of fine-grained emotions," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, pp. 4040–4054, ACL, 2020.

[5] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, and R. Verma, "CREMA-D: crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014, doi:10.1109/TAFFC.2014.2336244.

[6] S.R. Livingstone and F.A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, p. e0196391, 2018.

[7] V. Srivastava, B. Alam, S. Khan, V. Sajwan, and Niharika, "Intelligent traffic control system for ambulance priority using sound frequency," in *Proc. Int. Conf. Electronics Sustainable Commun. Syst.*, pp. 80–85, 2024.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *Proc. 5th Workshop Energy Efficient Mach. Learn. Cognitive Comput. – NeurIPS 2019 Workshop*, 2019.

[10] S.D. Alfonso,"AI in mental health,"*Curr. Opin. Psychol.*, vol. 36, pp. 112–117,2020, doi:10.1016/j.copsyc.2020.04.005.

[11] A. Mehta, A.N. Niles, T. Varker, A. Phelps, T.J. Metzler, and C.R. Marmar, "Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): longitudinal observational study," *J. Med. Internet Res.*, vol. 23, no. 6, p. e26771, 2021, doi: 10.2196/26771.

[12] J. Bautista, M. Liu, M. Alvarez, and S.M. Schueller, "Cognitive-behavioural therapy at our fingertips: Sanvello provides on-demand support for mental health," *Cogn. Behav. Pract.*, vol. 32, no. 2, pp. 206–213, 2025, doi:10.1016/j.cbpra.2023.12.008.

[13] C. Strauss, C. Dunkeld, and K. Cavanagh, "Is clinician-supported use of a mindfulness smartphone app a feasible treatment for depression? A mixed-methods feasibility study," *Internet Interv.*, vol. 25, p. 100413, 2021, doi: 10.1016/j.invent.2021.100413.