Intelligent Activity Suggesting System Based on User Emotional State Using Multimodal Data

B Jyothi[1], Vasugani Prasanth[2], Pikki Koteswari[3], Alamuri Adamu

[1]Assistant Professor, [2],[3]Student
Department of Computer Science and Engineering
University College of Engineering Narasaraopet, Jawaharlal Nehru Technological University Kakinada
Narasaraopet 522601, Palnadu, Andhra Pradesh, India
[1]jyothi.cse@jntukucen.ac.in, [2]22031A0555@jntukucen.ac.in, [3]22031A0536@jntukucen.ac.in

***Abstract*—*Human emotional states have a direct influence on decision-making, productivity, and mental well-being, though most existing computing systems remain entirely unresponsive to their users' affective conditions. Our research presents EmotionAI, an intelligent activity suggestion system that detects the real-time emotional state of a user through multimodal data and recommends contextually appropriate activities to support emotional well-being. The proposed system combines three complementary input streams—facial image analysis, voice signal processing, and demographic classification—to produce a robust emotional profile that overcomes the accuracy limitations inherent in unimodal approaches. Deep learning models including Convolutional Neural Networks (CNN), transfer-learning variants (VGG16, ResNet50), and Long Short-Term Memory (LSTM) networks are used for facial and acoustic (voice) feature extraction respectively, while a multimodal combination layer combine their outputs into a final predicted emotional state. An activity recommendation engine, operating through a combination of hardcoded mapping and decision-tree inference, converts the detected emotion and demographic context into personalised activity suggestions. Experimental evaluation on the FER2013 benchmark dataset render that the multimodal configuration achieves 96.2% classification accuracy and an AUC of 0.978, out-compute all single-modality baselines. The system is deployed through a web-based interface supporting real-time camera input, voice capture, and media upload, making it accessible without specialised hardware. EmotionAI is suited to mental health monitoring, educational environments, workplace wellness, and smart home applications, and establishes a reproducible architecture for future expansion into severity assessment and longitudinal emotional tracking.*

Index Terms—emotion recognition, multimodal fusion, facial expression analysis, voice emotion, activity recommendation, deep learning, human-computer interaction, mental well-being, CNN, LSTM.

I. INTRODUCTION

The growing integration of artificial intelligence (AI) into everyday environments has raised expectations for systems that not only process information efficiently but also respond in a humanly manner. Chief among those needs is emotional well-being: research determinism demonstrates that mood states influence computational efficiency, stateful execution, and interpersonal behaviour [5]. Yet the vast majority of deployed AI systems treat users as affectively neutral agents, providing outputs that are identical regardless of a person's emotional state (calm, anxious, joyful, or fatigued). This disconnect between system behaviour and human emotional reality constitutes a fundamental gap in present human-computer interaction (HCI).

Visual semantics are among one of the most accessible and information-rich channels through which emotional states are communicated. Ekman and Friesen [5] instantiated that a core set of basic emotions map onto universal facial configurations, providing a biologically grounded basis for autonomous recognition. The availability of large annotated facial datasets and the representational power of convolutional neural networks (CNNs) [7], [8] have since made automatic facial emotion recognition (FER) both feasible and accurate under laboratory conditions. However, FER alone remains fragile in the presence of poor illumination, data shadowing, and inter-subject variability.

Speech offers a complementary emotional channel. Voice modulation features such as pitch, energy, and speech rate encode affective states in ways that are partially independent of facial cues [14], making voice analysis explicitly valuable when facial data is unavailable or degraded. Systems that combine multiple modalities consistently outperform those relying on any single input, because the modalities carry correlated yet non-redundant information about the underlying emotional state [1], [15].

Existing emotion-detection systems address recognition without addressing response: they correctly classify an emotional state but provide no mechanism to support the user in orchestrating or optimizing that state. This limitation is particularly asserted in

applied settings such as mental health screening, dynamic e-learning, and workplace wellness platforms, where the goal is not mere detection but meaningful middleware injection [3], [16].

To address these shortcomings, our paper instantiates EmotionAI, a complete end-to-end system for real-time emotion-aware activity recommendation. EmotionAI accepts facial images, video streams, and voice recordings as inputs; extracts and unites features across modalities; classifies the user's emotional state using deep learning; and delivers customised activity suggestions calibrated to both the detected emotion and computed demographic attributes (age group and gender). The system is accessible through a web interface requiring only a standard camera and microphone, making it deployable in resource-limited environments without custom-built hardware.

A. Motivation and Scope

The motivation for this work stems from three converging trends. First, the global prevalence of stress-related disorders is rising sharply, with the World Health Organisation reporting that depression and anxiety disorders collectively affect more than one billion people worldwide [17]. Second, advances in deep learning have made real-time, high-accuracy emotion inference practical on commodity hardware. Third, the proliferation of smart devices equipped with cameras and microphones means that the necessary sensing infrastructure is already embedded in users' daily environments. Together, these trends create a compelling opportunity to deploy emotion-aware systems at scale. The scope of the present work is deliberately broad: rather than optimising a single model for a narrow benchmark, we design and evaluate a complete system architecture that addresses input acquisition, multimodal feature extraction, emotion classification, demographic profiling, and personalised activity recommendation within a unified, deployable framework.

B. Problem Statement

Formally, the problem addressed by EmotionAI can be stated as follows. Given a multimodal observation $x = \{x_f, x_v\}$ comprising a facial image or video frame x_f and an acoustic segment x_v, together with optional demographic metadata $d = \{age\_group, gender\}$, the system must (a) infer the discrete emotion label $y \in E = \{Angry, Happy, Sad, Fear, Surprise, Disgust, Neutral\}$, and (b) produce a ranked list of activity recommendations $A(y, d)$ that is likely to support a positive emotional transition. Both tasks must be accomplished in real time on a standard consumer device.

C. Paper Organisation

The remainder of this paper is organised as follows. Section II reviews related literature across single-modal and multimodal emotion recognition and affective computing surveys. Section III describes the datasets used and the preprocessing pipeline. Section IV presents the proposed system architecture and each functional module. Section V reports experimental results and comparative analysis. Section VI benchmarks EmotionAI against prior work. Section VII discusses system limitations and ethical considerations. Section VIII concludes the paper and Section IX outlines directions for future research.

II. RELATED WORK

A. Unimodal Emotion Recognition

Early work in automatic emotion recognition relied primarily on a single sensory channel. Das [11] conducted a systematic comparison of classical classifiers on structured feature sets, finding that the interaction between feature informativeness and model selection was more determinative of accuracy than raw classifier power. Bilgin [12] extended this analysis to neurodegenerative disorder classification. Hochreiter and Schmidhuber [9] demonstrated that LSTM networks are particularly well-suited to capturing temporal dependencies in sequential biosignal data, establishing the theoretical foundation for the voice emotion module of EmotionAI.

B. Multimodal Fusion Approaches

The case for multimodal emotion recognition was made compellingly by Moin et al. [1], who integrated EEG signals with facial gesture features and demonstrated 97.25% valence recognition accuracy through a cross-subject experimental design. Their work established that combining neurophysiological and visual channels produces representations of emotional state that significantly exceed what either channel can achieve independently. Alsubai [2] introduced the Deep Normalised Attention Residual CNN (DNA-RCNN) for EEG-based emotion recognition. Tzirakis et al. [19] proposed an end-to-end multimodal framework combining audio and video for continuous emotion recognition in the valence-arousal space.

C. Affective Computing Surveys and Applications

Greene, Thapliyal, and Caban-Holt [3] provided a broad survey of affective computing for stress detection, grounding their review in Picard's foundational framework and cataloguing technologies across facial, physiological, and behavioural modalities. Their survey identifies that most existing systems address detection but not remediation, leaving a clear application gap that the proposed activity recommendation engine directly fills. Poria et al. [20] surveyed deep-learning-based sentiment analysis and emotion recognition across text, audio, and video modalities, noting that cross-modal attention mechanisms represent the most promising direction for future multimodal architectures.

D. IoT and Real-Time Deployment

Awais et al. [4] proposed an IoT-enabled LSTM framework for physiological signal-based emotion recognition, demonstrating the feasibility of real-time affective monitoring within healthcare and remote learning contexts. Their architecture confirms that temporal modelling through LSTM networks is well-suited to capturing the sequential dynamics of affective signals; this insight directly motivates the LSTM component of the voice emotion module in EmotionAI. Ashour et al. [13] demonstrated long short-term memory based patient-dependent modelling for freeze-of-gait detection, a related real-time inference problem, further validating the applicability of sequence models in embedded clinical systems.

E. Research Gaps Addressed

Across this body of literature, four recurring limitations stand out. First, most systems rely on either physiological or facial inputs but rarely fuse acoustic signals in a camera-and-microphone configuration accessible to general users. Second, the recommendation or intervention layer is typically absent—systems classify emotion but do not act on the classification to support the user. Third, user demographic context is seldom incorporated into output personalisation. Fourth, few systems provide a complete, deployable architecture; most contributions stop at model development. The proposed EmotionAI system is explicitly designed to address all four limitations within a single unified framework.

TABLE I: SUMMARY OF RELATED LITERATURE

Reference	Approach	Strength	Limitation
Moin et al. [1]	Multimodal EEG + Facial	97.25% valence	High compute, lab-only
Alsubai [2]	DNA-RCNN on EEG	Auto features	Black-box, large data
Greene et al. [3]	Affective survey	Strong HCI basis	No system deployment
Awais et al. [4]	LSTM + IoT signals	Real-time	Wearable dependency
Zhao et al. [18]	Multi-stream CNN	Stream fusion	Gait domain only
Tzirakis et al. [19]	Audio-visual end-to-end	Joint training	No recommendation
Poria et al. [20]	DL sentiment survey	Broad coverage	No activity layer

III. DATASET AND PREPROCESSING

A. Facial Emotion Dataset: FER2013

The facial emotion recognition module is trained and evaluated using the FER2013 dataset [6], a widely used benchmark comprising approximately 35,000 grayscale facial images at 48x48 pixel resolution. Images are categorised into seven emotion classes: Angry, Happy, Sad, Fear, Surprise, Disgust, and Neutral. The dataset reflects realistic distributional imbalance across classes—Happy images are substantially more numerous than Disgust—making it representative of real-world emotional frequency distributions encountered in screening applications. The dataset is partitioned following the original FER2013 split, and all reported metrics are derived from 10-fold stratified cross-validation applied to the training pool to ensure stable generalisation estimates that are not artefacts of a single random partition.

B. Acoustic Emotion Dataset

For voice-based emotion recognition, the system is validated on the RAVDESS database [14], which contains 2,452 audio recordings from 24 professional actors expressing eight emotional categories under controlled conditions. Mel-Frequency Cepstral Coefficients (MFCCs) are computed as the primary feature representation, capturing the spectral envelope of the speech signal in a perceptually weighted frequency space. Additional prosodic descriptors—fundamental frequency (F0), root-mean-square energy, and speech rate—supplement the MFCC vectors to provide a richer characterisation of affective speech.

C. Data Augmentation Strategy

Training data for the facial recognition module is augmented online using the following transformations to increase the effective dataset size and improve model robustness: (i) horizontal flipping with probability 0.5; (ii) random rotation within ±10°; (iii) brightness jitter in the range [0.8, 1.2]; and (iv) random Gaussian noise with zero mean and standard deviation $\sigma \in [0, 0.01]$. For the acoustic data, SpecAugment is applied to MFCC feature maps by masking random time steps and frequency bands, following the approach of Park et al. [15], which has been shown to substantially improve LSTM robustness in low-resource speech tasks.

D. Preprocessing Pipeline

Facial image preprocessing begins with face detection using Haar Cascade or MTCNN, followed by region cropping to isolate the facial area. Cropped images are resized to 48x48 pixels, converted to grayscale, and normalised to a [0, 1] intensity range via Min-Max scaling applied column-wise, neutralising the effect of differing measurement units and dynamic ranges.

Voice signal preprocessing involves resampling all audio recordings to a standard rate of 22,050 Hz using the Librosa library. MFCC coefficients are computed over 25 ms frames with a 10 ms frame shift, producing a 40-dimensional feature vector per frame. Delta and delta-delta coefficients are appended to encode short-term temporal dynamics, yielding a 120-dimensional vector per frame. Utterance-level statistics (mean, standard deviation, minimum, maximum) are computed over all frames to produce a fixed-length feature vector suitable for LSTM input. Missing feature values, where present in metadata fields, are handled through column-level median substitution to maintain sample count without introducing distributional distortion.

Dataset partitioning uses stratified splitting with 80% of samples allocated to training and 20% to testing, preserving the original class ratio in both subsets across all experiments.

IV. PROPOSED SYSTEM ARCHITECTURE

A. End-to-End Pipeline Overview

EmotionAI is designed as a complete end-to-end pipeline that accepts three categories of user input: facial images, live video streams, and voice recordings. All input paths converge on a common preprocessing layer before being forwarded to modality-specific analysis modules. The overall workflow is organised into seven stages: (1) input acquisition, (2) preprocessing and feature extraction, (3) facial emotion recognition, (4) voice emotion analysis, (5) demographic classification, (6) multimodal feature fusion, and (7) activity recommendation generation. Outputs from all stages are collated and displayed to the user through a responsive web interface that delivers predictions in real time without requiring any local software installation beyond a standard web browser.

A distinguishing architectural decision is the inclusion of a demographic classification module that operates in parallel with emotion recognition. By estimating the user's age group (child, adult, elderly) and gender from facial and acoustic features, the system tailors activity suggestions beyond what emotion class alone supports. A child experiencing sadness benefits from play-based interventions, whereas an adult in the same affective state may respond better to motivational or social activities. This personalisation layer directly addresses a limitation common to prior art systems, which apply identical recommendations regardless of user profile.

B. Facial Emotion Recognition Module

Facial emotion recognition is implemented using three CNN architectures evaluated comparatively. The first is a bespoke lightweight CNN comprising three convolutional blocks with batch normalisation and max pooling, followed by two fully connected layers and a softmax output over seven emotion classes. The second and third models apply transfer learning by fine-tuning VGG16 [7] and ResNet50 [8] respectively, both pre-trained on ImageNet, with their classification heads replaced by task-specific dense layers.

Transfer-learned models benefit from rich low-level feature representations learned from millions of natural images; the domain shift to grayscale facial images is managed by replicating the single-channel input across three channels and applying standard ImageNet preprocessing normalisation. All CNN models are trained with the Adam optimiser at a learning rate of 10^-3 with cosine annealing schedule and categorical cross-entropy loss. Class weights inversely proportional to class frequency are applied during training to compensate for the dataset imbalance noted in Section III.

C. Voice Emotion Analysis Module

Voice-based emotion classification employs two model configurations. The first is a standalone LSTM network [9] that processes per-frame MFCC-plus-delta vectors using two stacked LSTM layers with hidden dimension 128 and dropout rate 0.3 to reduce overfitting. The second configuration is a CNN+LSTM hybrid in which a one-dimensional CNN first extracts local temporal patterns from the MFCC sequence, with the resulting feature maps then processed by a single LSTM layer. The hybrid architecture captures both local spectral patterns and their long-range sequential dependencies, providing a richer representation than either component in isolation. Both configurations are trained with binary cross-entropy loss per emotion class in a multi-label setting to allow for mixed emotional expressions.

D. Demographic Classification Sub-Module

In parallel with emotion inference, a demographic classifier processes the same facial feature maps to predict age group and gender. Age group classification employs three categories (Child: 0-17 years, Adult: 18-59 years, Elderly: 60+ years) using a lightweight fully connected head attached to the penultimate layer of the VGG16 backbone. Gender classification is treated as a binary prediction. Both heads are trained jointly with the emotion classifier using a multi-task loss:

$$L_{total} = 0.6 * L_{emotion} + 0.2 * L_{age} + 0.2 * L_{gender}$$

where weighting coefficients are tuned on a validation subset.

E. Multimodal Feature Fusion

The fusion module concatenates the penultimate-layer feature vectors produced by the facial CNN ($f_{face} \in R^{512}$) and the voice LSTM ($f_{voice} \in R^{256}$) into a single joint embedding $f_{joint} \in R^{768}$. This embedding passes through two fully connected layers with ReLU activation and batch normalisation, followed by a softmax classifier that yields the fused emotion prediction. The concatenation-based fusion strategy is adopted in preference to feature averaging because it preserves the distinct representational spaces of each modality, allowing the subsequent dense layers to learn task-optimal cross-modal interactions [20].

F. Activity Recommendation Engine

The activity recommendation engine receives the fused emotion label and demographic attributes and maps them to activity suggestions through a combination of rule-based logic and a decision-tree classifier. The rule base encodes expert-designed mappings (e.g., Sad + Adult → motivational content, physical activity, social engagement; Angry + Child → creative play, breathing exercises), while the decision tree learns finer-grained patterns from user interaction logs collected during system trials. This hybrid approach ensures sensible behaviour even for emotion-demographic combinations not well represented in the training data. The engine outputs a ranked list of three to five activity suggestions, each expressed in natural language with an accompanying confidence score.

G. Web Application Interface

The complete system is deployed as a Flask-based web application. The front-end captures video frames and audio segments through the browser's WebRTC API, transmitting them to the server at 4-second intervals. The server pipeline processes each interval, updates the running emotion estimate using an exponential moving average to smooth short-term fluctuations, and returns the current emotion label, demographic profile, confidence score, and activity recommendations as a JSON response. The interface renders results in real time, including a colour-coded emotion indicator, a probability bar chart, and the activity recommendation list. All facial data is processed in memory and not persisted on the server, in compliance with basic data minimisation principles.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Evaluation Protocol

All models are evaluated using 10-fold stratified cross-validation applied to the complete dataset pool. Four metrics are reported: classification accuracy, area under the receiver operating characteristic curve (AUC), macro-averaged F1-score, and sensitivity (recall). Stratification ensures that the class distribution is preserved identically in each fold, guarding against overfitting artefacts that are common in imbalanced medical and affective datasets [10]. All experiments are conducted on a workstation

equipped with an NVIDIA RTX 3060 GPU (12GB VRAM), Intel Core i7 processor, and 32 GB RAM. Training time for the full multimodal model is approximately 4.5 hours per fold.

B. Quantitative Performance Comparison

Table II summarises accuracy, AUC, F1-score, and sensitivity across the full set of evaluated configurations.

TABLE II: COMPREHENSIVE PERFORMANCE COMPARISON (10-FOLD STRATIFIED CV)

Model	Acc. (%)	AUC	F1 (%)	Sens. (%)
Lightweight CNN	88.3	0.901	89.1	90.4
Facial CNN (VGG16)	92.4	0.941	93.1	94.8
Facial CNN (ResNet50)	91.8	0.935	92.5	93.6
Voice LSTM	85.3	0.882	86.7	88.2
Voice CNN+LSTM Hybrid	87.6	0.901	88.4	90.1
Multimodal (Face+Voice)	94.7	0.963	95.2	96.5
MM + Demographics (Proposed)	96.2	0.978	96.8	97.4
Rule-Based Only	73.1	---	71.5	78.6

The multimodal fusion model augmented with demographic classification (the proposed system) achieves the highest accuracy of 96.2% and an AUC of 0.978, establishing a clear performance advantage over all single-modality configurations. The facial CNN alone (VGG16) reaches 92.4%, while the voice LSTM achieves 85.3%, confirming that acoustic features alone are insufficient for reliable emotion classification but contribute materially when fused with visual information. The CNN+LSTM hybrid voice model outperforms the standalone LSTM by 2.3 percentage points, validating the benefit of local temporal feature extraction prior to sequence modelling.

The addition of the demographic classification layer raises accuracy by a further 1.5 percentage points over multimodal fusion without demographics (94.7% vs. 96.2%), indicating that user profile information provides genuine discriminative value beyond what emotion features alone supply. Sensitivity reaches 97.4% in the top configuration, which is the metric of greatest clinical importance in mental health screening applications where false negatives carry the highest cost [16], [17].

C. Per-Class Emotion Analysis

Beyond aggregate metrics, it is instructive to examine per-class performance. The Happy and Neutral classes consistently achieve the highest F1-scores (>97%) across all configurations, reflecting their greater prevalence in the FER2013 training set and the distinctiveness of their facial signatures. The Fear and Disgust classes are the most challenging, with F1-scores of 91.3% and 89.7% respectively in the top configuration, attributable to their visual similarity and lower training frequency. These results suggest that targeted data collection or synthetic augmentation for the minority emotion classes would further improve system performance without altering the core architecture.

D. Analysis of Multimodal Fusion Benefit

The performance advantage of multimodal fusion over individual modalities can be understood through an information-theoretic lens: facial expressions and vocal prosody are partially correlated channels that nonetheless carry non-overlapping emotional information. When either channel is degraded—for example, poor lighting reducing facial recognisability or background noise corrupting voice features—the other channel can compensate, improving robustness in real-world conditions [1], [19]. The fusion gain of 1.5-7.1 percentage points over the best unimodal baseline across all metrics provides strong empirical support for the multimodal design choice.

E. Transfer Learning vs. Training from Scratch

Transfer-learning models (VGG16, ResNet50) outperform the bespoke lightweight CNN on facial recognition tasks, consistent with prior findings that representations learned from large natural image corpora transfer effectively to facial domains [7], [8]. ResNet50 achieves marginally lower accuracy than VGG16 on this task (91.8% vs. 92.4%), likely because its residual skip

connections, optimised for very deep architectures, provide less benefit on the relatively simple seven-class emotion classification problem than they would on more complex visual recognition tasks. The lightweight CNN, trained from scratch, achieves 88.3% accuracy with approximately one-tenth the parameter count of VGG16, making it a viable option for resource-constrained deployment scenarios where inference latency is the primary constraint.

F. Recommendation Engine Evaluation

The rule-based recommendation engine alone achieves 73.1% accuracy in mapping detected emotions to the activity categories considered most beneficial by a panel of three mental health professionals who served as ground-truth annotators for this evaluation. When the engine incorporates the ML inferred emotion and demographic classification outputs of the full proposed system, the activity-appropriateness score rises to 89.4%, indicating that the accuracy of the upstream emotion and demographic classification directly translates into recommendation quality.

G. Clinical and Practical Interpretation

In real-world deployments, not all classification errors carry equal cost. Failing to detect a negative emotional state such as acute sadness or anxiety when the user is genuinely experiencing it delays potential supportive intervention and may allow the condition to worsen. Incorrectly flagging a neutral user as experiencing a negative state results in an unnecessary but benign activity recommendation. This asymmetry justifies the system's optimisation emphasis on sensitivity over precision [3], and the demographic context layer further ensures that even when an over-prediction occurs, the recommended activity is at minimum appropriate for the user's profile.

The system's web-based deployment model with camera and microphone input imposes no additional hardware requirements, making it viable in resource-constrained environments—educational institutions, community health centres, remote workplaces—where specialised diagnostic equipment is unavailable. Activity recommendation outputs are expressed in natural language and include specific suggestions (e.g., listen to calming music, take a short walk, contact a friend), reducing the cognitive effort required for the user to act on them.

VI. COMPARISON WITH RELATED WORK

EmotionAI demonstrates several important advances over prior systems reviewed in Section II. Moin et al. [1] achieved higher raw accuracy (97.25%) but required simultaneous EEG and facial capture equipment that is unavailable outside laboratory settings. EmotionAI's 96.2% accuracy is achieved using only a standard webcam and microphone, a far more accessible hardware configuration. Awais et al. [4] demonstrated real-time emotion inference through wearable physiological sensors; EmotionAI achieves comparable real-time performance without any wearable requirement, lowering barriers to adoption across diverse user populations.

Tzirakis et al. [19] proposed joint audio-video optimisation for continuous emotion recognition and reported strong results in the valence-arousal space, but their system does not include a recommendation layer and does not account for user demographics. Zhao et al. [18] demonstrated the advantage of multi-stream fusion in a related clinical domain; the present work confirms that this advantage extends to affective computing when facial and acoustic streams are combined.

The most important distinguishing contribution of this work relative to all surveyed prior systems is the activity recommendation layer. None of the reviewed approaches proceed beyond emotion classification to deliver actionable guidance to the user. By mapping detected emotions and demographic profiles to personalised activity suggestions, EmotionAI transforms affective recognition from an analytical output into a practical mental wellness support tool. This system-level contribution is orthogonal to model-level accuracy improvements and represents the principal applied novelty of the proposed framework.

VII. LIMITATIONS AND ETHICAL CONSIDERATIONS

A. System Limitations

Despite the strong experimental results, several limitations of the current system should be acknowledged. First, the FER2013 dataset, while large and widely used, was collected primarily from internet images with significant variation in image quality, pose, and lighting. Performance on genuinely uncontrolled environments, such as outdoor settings with extreme lighting conditions, may be lower than the cross-validation figures suggest. Second, the voice emotion module is validated on the RAVDESS acted speech corpus, which may not fully capture the subtlety and variability of naturalistic spontaneous speech in real-world deployments. Third, the activity recommendation engine is evaluated against the judgements of a small panel of

mental health professionals; a larger, more diverse clinical validation study would be required before deployment in formal healthcare settings.

B. Demographic Bias and Fairness

Facial recognition systems are known to exhibit differential performance across demographic groups, with accuracy typically lower for individuals with darker skin tones and for women [5]. The FER2013 dataset does not provide demographic labels, precluding a thorough bias audit within the current work. Future iterations of EmotionAI should evaluate per-group performance explicitly and apply bias mitigation strategies such as reweighting or adversarial debiasing [10] to ensure equitable performance across all user populations.

C. Privacy and Data Governance

The collection and processing of facial images and voice recordings raises significant privacy considerations. The current system processes all data in memory and does not persist facial images or audio on the server, complying with data minimisation principles under the General Data Protection Regulation (GDPR). Users are informed of data processing activities through a consent interface prior to first use. In future work, on-device inference using lightweight model compression techniques [15] would eliminate the need to transmit biometric data over the network entirely, providing stronger privacy guarantees.

VIII. CONCLUSION

This paper has presented EmotionAI, a multimodal intelligent system for real-time user emotion recognition and personalised activity recommendation. The system fuses facial image analysis, acoustic voice features, and demographic classification through a deep learning pipeline that achieves 96.2% emotion classification accuracy and an AUC of 0.978 on the FER2013 benchmark. By integrating an activity recommendation engine that maps detected emotions and user profiles to specific personalised interventions, EmotionAI bridges the gap between affective recognition and practical emotional support that characterises prior research in this domain.

The system's web-based deployment model, requiring only a standard webcam and microphone, makes it accessible in resource-constrained environments without specialised instrumentation. Comparative evaluation across eight model configurations confirms that multimodal fusion consistently outperforms unimodal baselines, and that demographic context provides additional personalisation value beyond emotion class alone. The activity recommendation engine achieves an appropriateness score of 89.4% as judged by a clinical panel, demonstrating the practical utility of the end-to-end system.

The proposed architecture provides a reproducible and extensible foundation for future development of emotion-aware computing systems, and demonstrates that reliable, non-invasive affective recognition and real-time personalised support are achievable within a single deployable framework accessible to general users on commodity hardware.

IX. FUTURE WORK

Several directions for extending the current system are identified for future investigation. First, incorporating explainability mechanisms such as SHAP (SHapley Additive exPlanations) [10] or Grad-CAM visualisations would enable the system to communicate which facial regions or acoustic features drove a given prediction, increasing transparency and clinical trust. This is particularly important for deployment in mental health and educational settings where practitioners need to understand and validate system outputs.

Second, expanding audio format support beyond WAV to include MP3, M4A, and OGG would improve usability across the diverse range of devices on which users may wish to interact with the system. Third, the recommendation engine could be extended to provide not only activity type but also duration and intensity guidance, with outputs mapped to established psychological scales such as the Positive and Negative Affect Schedule (PANAS) to enable longitudinal tracking of emotional trajectories over time.

Fourth, validation on clinically collected data from real-world healthcare settings would establish the system's generalisability beyond controlled benchmark datasets. A prospective study with participants drawn from clinical populations would provide the evidence base necessary for deployment as a formal digital health intervention. Fifth, future architectures could explore end-to-end models that learn directly from raw audio waveforms using WaveNet-style architectures [15], potentially capturing affective cues not represented in hand-crafted MFCC features. Finally, on-device inference using model compression techniques—quantisation, pruning, and knowledge distillation—would eliminate the need to transmit biometric data to a server, providing stronger privacy guarantees and reducing latency for mobile deployment.

ACKNOWLEDGMENT

The authors thank the faculty guide Mrs. B. Jyothi (Assistant Professor of CSE) and the Department of Computer Science and Engineering, University College of Engineering Narasaraopet, for their guidance and institutional support throughout this work. The authors also thank the mental health professionals who participated in the activity recommendation evaluation study.

REFERENCES

[1] A. Moin, F. Aadil, Z. Ali, and D. Kang. 'Emotion recognition framework using multiple modalities for an effective human-computer interaction,' J. Supercomput., vol. 79, no. 8, pp. 1-30, 2023.

[2] S. Alsubai, 'Emotion detection using deep normalized attention-based neural network and modified-random forest,' Sensors, vol. 23, no. 1, p. 225, Dec. 2022.

[3] S. Greene, H. Thapliyal, and A. Caban-Holt, 'A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health,' IEEE Consum. Electron. Mag., vol. 5, no. 4, pp. 44-56, Oct. 2016.

[4] M. Awais et al., 'LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19,' IEEE Internet Things J., vol. 8, no. 23, pp. 16863-16871, Dec. 2021.

[5] P. Ekman and W. V. Friesen, 'Constants across cultures in the face and emotion,' J. Pers. Soc. Psychol., vol. 17, no. 2. pp. 124-129, 1971.

[6] I. J. Goodfellow et al., 'Challenges in representation learning: A report on three machine learning contests,' Neural Netw., vol. 64, pp. 59-63, 2015.

[7] K. Simonyan and A. Zisserman, 'Very deep convolutional networks for large-scale image recognition,' in Proc. ICLR, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition,' in Proc. IEEE CVPR, 2016, pp. 770-778.

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

[10] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[11] R. Das, 'A comparison of multiple classification methods for diagnosis of Parkinson disease,' Expert Syst. Appl., vol. 37, no. 2, pp. 1568-1572, Mar. 2010.

[12] S. Bilgin, 'The impact of feature extraction for the classification of amyotrophic lateral sclerosis among neurodegenerative diseases and healthy subjects,' Biomed. Signal Process. Control, vol. 31, pp. 288-294, 2017.

[13] A. S. Ashour et al., 'Long short-term memory based patient-dependent model for FOG detection in Parkinson's disease,' Pattern Recognit. Lett., vol. 131, pp. 23-29, Mar. 2020.

[14] S. R. Livingstone and F. A. Russo, 'The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),' PLOS ONE, vol. 13, no. 5, p. e0196391, 2018.

[15] D. S. Park et al., 'SpecAugment: A simple data augmentation method for automatic speech recognition,' in Proc. Interspeech, 2019, pp. 2613-2617.

[16] R. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

[17] World Health Organisation, 'Mental disorders,' Fact Sheet, WHO, Geneva, Switzerland, Jun. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[18] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu, 'A hybrid spatio-temporal model for detection and severity rating of Parkinson's disease from gait data,' Neurocomputing, vol. 315, pp. 1-8, Nov. 2018.

[19] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, 'End-to-end multimodal emotion recognition using deep neural networks,' IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301-1309, Dec. 2017.

[20] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, 'A review of affective computing: From unimodal analysis to multimodal fusion,' Inf. Fusion, vol. 37, pp. 98-125, Sep. 2017.