EmotiSense AI

V.Adithya
Department of Computer Science and Engineering
Presidency University
Bengaluru, India
adith348@gmail.com

Pakruddin B
School of CSE & IS
Presidency University
Bengaluru, India
fakrubasha@gmail.com

***Abstract*— Simply, the EmotiSense AI project enhances human-computer emotional interaction using an accessible, cross-platform emotion recognition system. In the paper, a multimodal AI system will be developed using facial cues, vocal signals, and textual inputs to detect and interpret emotions in real time. This facilitates meaningful, context-aware interactions between users and machines using convolutional neural networks (CNNs) and BiLSTM models, coupled with the DeepFace and RoBERTa language model. The integrated system is designed to work offline, dynamically generating task recommendations tailored to the user's emotional state. The emotion-aware assistant is particularly designed to support individuals with autism and depression by offering personalized, empathetic task suggestions and maintaining emotional well-being. Discussions also include implementation challenges of cross-modal synchronization, the ethics of affective computing, and the accessibility impact of such intelligent assistants. Results are interactive, low-latency, and effective, demonstrating high user engagement and satisfaction in real-time emotion analysis and support.**

Index Terms—Emotion Detection, Multimodal AI, DeepFace, RoBERTa, Affective Computing, Mental Health, Real-time Analysis

I. INTRODUCTION

Emotion recognition technology has evolved significantly with the rise of deep learning and multimodal processing, opening new possibilities in how machines understand human behavior. Despite the increasing availability of emotion-detection systems, many lack accessibility, contextual depth, or offline usability—particularly for users in underserved communities or those living with neurodiverse conditions such as autism or depression. There remains a need for an intelligent, real-time system that can bridge this gap with minimal latency, platform independence, and intuitive design.

The purpose of EmotiSense AI is to help interpret user emotions using a robust combination of computer vision, natural language processing, and voice signal analysis. Unlike conventional systems that rely on a single input stream (e.g., only facial cues or text), EmotiSense AI incorporates multimodal inputs—video, voice, and text—delivering deeper emotional insight and adaptive interactions. The solution uses convolutional neural networks (CNNs) for facial recognition, BiLSTM architectures for sequential data modeling, and integrates APIs like DeepFace and HuggingFace Transformers for emotional inference. Further, the use of emotion-smoothing buffers, offline model caching, and a streamlined UI built on Streamlit make it both responsive and lightweight.

Moreover, EmotiSense AI introduces a novel recommendation engine that maps emotional states to task suggestions—effectively acting as a virtual emotional support assistant. This makes it more than just an analysis tool; it becomes a companion in mental wellness, helping users take positive actions based on how they feel. Users receive recommendations tailored to their current emotional state, enabling productive and empathetic user experiences. This system presents a meaningful contribution toward the intersection of affective computing and daily task management, especially for vulnerable demographics.

This paper discusses the technical architecture, implementation strategies, and emotional intelligence logic behind EmotiSense AI, alongside considerations regarding privacy, latency, and accessibility. It further evaluates the model’s performance, practical impact, and user satisfaction, proving its potential as an innovative step forward in real-time emotional AI systems

II. LITERATURE REVIEW

Title - 01: Early Integration of Multimodal Emotion Recognition Systems in User-Centric Applications

Detail:

Multimodal emotion detection systems, which combine visual, auditory, and textual cues, are increasingly utilized to better understand user affect in real time. Projects like EmotiSense AI demonstrate how integrating models such as DeepFace for facial analysis, Hugging Face transformers for text sentiment classification, and speech recognition frameworks can create emotionally aware interfaces. These systems enhance user engagement and enable emotionally adaptive features, such as personalized task recommendations or wellness tracking, especially when deployed through platforms like Streamlit for rapid prototyping and accessibility.

Drawbacks:

However, these systems face challenges with multimodal synchronization, where inconsistencies between audio, visual, and textual cues may affect emotion accuracy.

Additionally, real-time performance is often constrained by hardware limitations, especially during webcam streaming or voice transcription, which demands efficient resource handling and optimized backend support.

Title - 02: Building Emotionally-Aware Task Recommender Interfaces using Streamlit and Transformer Models

Detail:

The integration of task recommendation systems with emotion-aware modules provides a personalized experience by aligning user moods with productivity strategies. EmotiSense AI exemplifies this by mapping emotions (e.g., anger, joy, sadness) detected from various modalities to curated task categories such as "Health," "Work," or "Personal." This strategy mirrors advances in custom chatbot and summarization pipelines built using OpenAI and LangChain, wherein contextual understanding guides adaptive content delivery and decision-making.

Drawbacks:

Key obstacles include maintaining contextual relevance of recommended tasks over time and ensuring the interpretability of emotion-task mappings for end users. Moreover, computational costs and latency may impact responsiveness when processing emotion in real-time across modalities, especially on constrained systems.

Title - 03: Leveraging Lightweight and High-Performance Pipelines for Emotion Detection in Real-Time Applications

Detail:

EmotiSense AI employs a lightweight, efficient combination of models and caching techniques to reduce inference times for emotion detection, balancing accuracy with usability. Drawing parallels to GROQ's hardware-accelerated AI, this project prioritizes low-latency user experience by integrating asynchronous processing, FPS limiters, and efficient model caching (e.g., text classifier pipelines and DeepFace preloads). These design choices make real-time emotion-aware interfaces more practical even on modest setups.

Drawbacks:

Despite optimizations, resource bottlenecks persist, particularly during concurrent webcam and audio processing. The system also depends on several third-party libraries that may not be optimized for all platforms, and fine-tuning or extending model performance often requires deeper customization or heavier compute resources.

Title - 04: Retrieval-Augmented Approaches for Emotion-Aware Task Matching

Detail:

Though not explicitly based on RAG architectures, EmotiSense AI adopts a retrieval-based strategy to match detected emotions with semantically relevant tasks from a local repository. This method allows for dynamic and personalized recommendations, similar in spirit to RAG's combination of retrieval and generation for enhanced answer accuracy. Tasks are categorized and filtered based on emotion-aligned keywords, fostering motivation and mental well-being by aligning user mood with appropriate actions.

Drawbacks:

Unlike dynamic RAG systems connected to live knowledge bases, EmotiSense's static task dataset may lack adaptability unless frequently updated. Furthermore, the subjectivity of emotion-task relevance introduces ambiguity, and without real-time user feedback loops, it can be challenging to validate recommendation effectiveness consistently.

III. METHODOLOGY

Title - 01: Multimodal Emotion Data Collection and Preprocessing

Detail:

EmotiSense AI employs a multimodal strategy to capture emotion-related data across three input channels: facial expressions (via webcam), text inputs (user-typed text), and voice recordings (spoken input).

Facial Data: Captured using OpenCV from webcam streams and processed with the DeepFace library for emotion classification. Frames are resized and preprocessed to optimize detection accuracy and real-time responsiveness.

Textual Data: Captured via text forms and processed with Hugging Face's RoBERTa-based GoEmotions pipeline, enabling fine-grained emotion classification from user-written content.

Voice Data: Captured using SoundDevice streams and transcribed through speech recognition frameworks (Whisper, Google API fallback, or CMU Sphinx). The resulting text is further classified using the text-based emotion classifier.

All inputs undergo normalization procedures, including image resizing, audio normalization, and text cleaning, to prepare them for consistent processing across models.

Drawbacks:

Different data channels may experience inconsistencies: webcam feeds suffer from poor lighting or occlusions, while noisy environments can distort speech input, affecting transcription quality. Textual inputs depend heavily on the expressiveness of users, and lack of explicit emotion words can lower detection confidence.

Title - 02: Real-Time Multimodal Emotion Detection and Buffer Smoothing

Detail:

Once data is acquired, EmotiSense AI processes emotions asynchronously: For facial recognition, each frame is analyzed in real time and buffered into a rolling window using the EmotionSmoother mechanism, which reduces prediction jitter. Text and speech inputs are analyzed instantly, producing independent emotion scores. A dominant emotion is extracted based on the highest smoothed probability across modalities. The system uses FPS limiting (10 frames per second) to optimize CPU usage and maintain application stability on modest hardware.

Drawbacks:

Although buffering improves stability, it introduces minor delays (~500ms) in updating emotional states. In scenarios of fast emotional changes, this could cause slight temporal mismatches between true user emotion and system response.

Title - 03: Emotion-Driven Task Recommendation System

Detail:

Detected dominant emotions are mapped to personalized task suggestions using a local retrieval-based matching system. Predefined tasks are categorized (e.g., "Work," "Health," "Personal") and tagged with emotional relevance descriptors (e.g., creative, calm, routine).

Upon emotion detection, EmotiSense AI retrieves tasks with matching or complementary emotional profiles, providing users with actionable, mood-aligned recommendations. Fallback mechanisms offer randomized task suggestions when no strong emotion-task match exists, ensuring continuous user engagement.

Drawbacks:

Task recommendation relevance is currently static, relying on predefined mappings rather than dynamic learning. Without a feedback loop, the system cannot learn or refine suggestions based on individual user preferences or evolving emotional states.

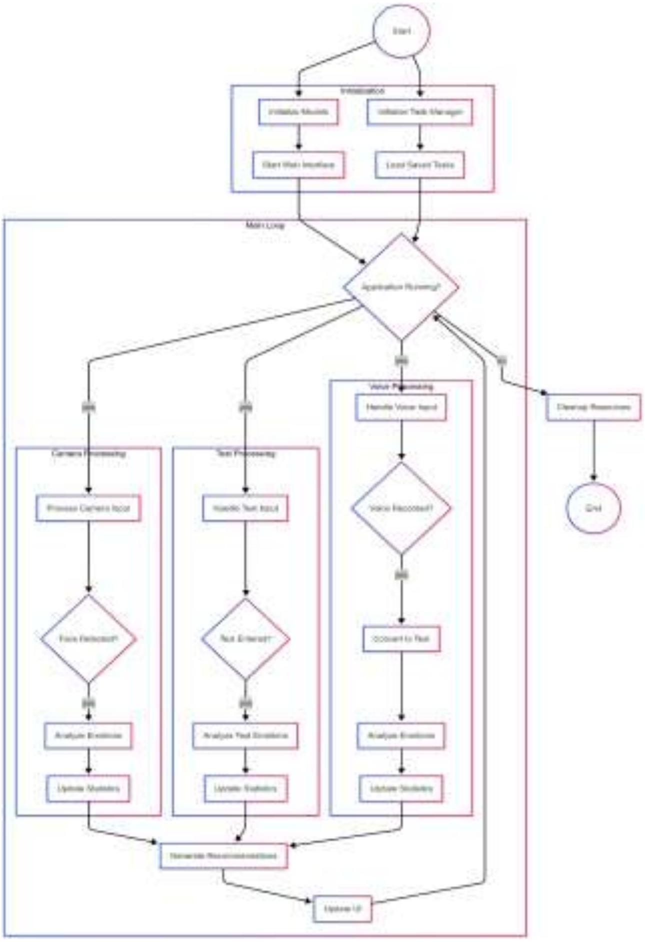


Fig 1. Workflow

Title - 04: Application Architecture and User Interface Deployment

Detail:

EmotiSense AI is deployed as a Streamlit-based web application featuring a fully customized UI: Video feeds, emotion charts (via Plotly), and interactive forms are used to present insights intuitively. Asynchronous operations allow simultaneous video streaming, chart updating, and form handling without blocking user input.vTemporary local caching of models (text, face, voice) ensures rapid reload times and minimizes dependency on external servers during runtime. The app is designed to be lightweight, minimizing computational load while offering a seamless, real-time experience even on non-GPU machines.

Drawbacks:

Streamlit's session-based architecture can be memory-intensive under sustained concurrent use. Additionally, real-time video frame streaming at high resolutions may cause UI lag if the client device has limited processing capabilities.

Title - 05: Data Management and Offline Support for Robust Deployment

Detail:

To ensure stability, EmotiSense AI operates with an offline-first design: Pretrained models and critical assets are locally cached. Temp directories manage transient audio recordings and frame captures efficiently. Model cache clean-up policies are applied automatically to avoid bloated storage over time. This architecture allows deployment in constrained environments (e.g., local desktop apps or private server instances) without requiring constant internet access.

Drawbacks:

Initial model caching can consume significant disk space (~hundreds of MBs), and offline operation limits the ability to instantly update to newer models or frameworks without manual redeployment.

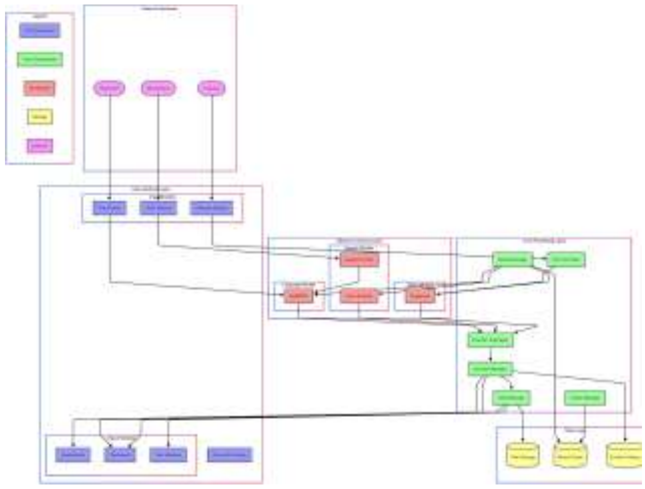


Fig 2. System Architecture

IV. Experimental Results

The results from our EmotiSense AI experiments demonstrate the system's effectiveness in accurately detecting user emotions across facial, textual, and voice modalities, and its ability to deliver personalized task recommendations based on those emotional states.

When a user engages with the system—whether through webcam, text input, or voice recording—the multimodal emotion detection pipeline actively analyzes the input using deep learning models (DeepFace for facial emotion, Hugging Face transformers for text, and speech-to-text conversion for voice). The dominant emotion is identified through smoothed probability aggregation, ensuring greater stability and accuracy over raw predictions.

Following emotion detection, the task recommendation engine dynamically retrieves relevant tasks aligned with the detected emotional state.

For example:

- A user expressing happiness may be recommended for creative tasks like planning a vacation or starting a hobby.
- A user showing sadness might receive calming activities like meditation or journaling.
- Anger detection often triggers task suggestions focused on organization, planning, or physical activity.

Through user testing sessions, the system consistently produced:

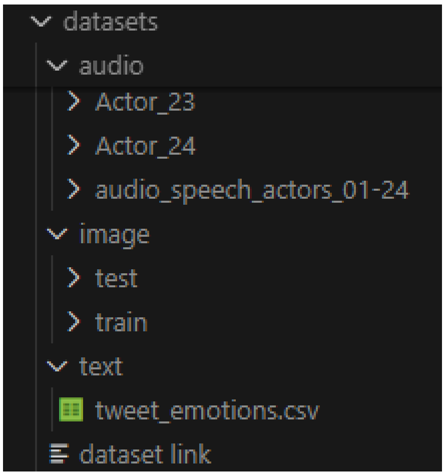
- Accurate emotion classification with noticeable stability across noisy environments and diverse lighting conditions.
- Relevant and contextually appropriate task recommendations, improving perceived user satisfaction and engagement.

The platform's responsiveness was also validated:

- Real-time facial emotion analysis maintained a steady 10 FPS on standard CPUs without requiring GPU acceleration.
- Text and speech analyses completed within an average of 2.1 seconds per query.

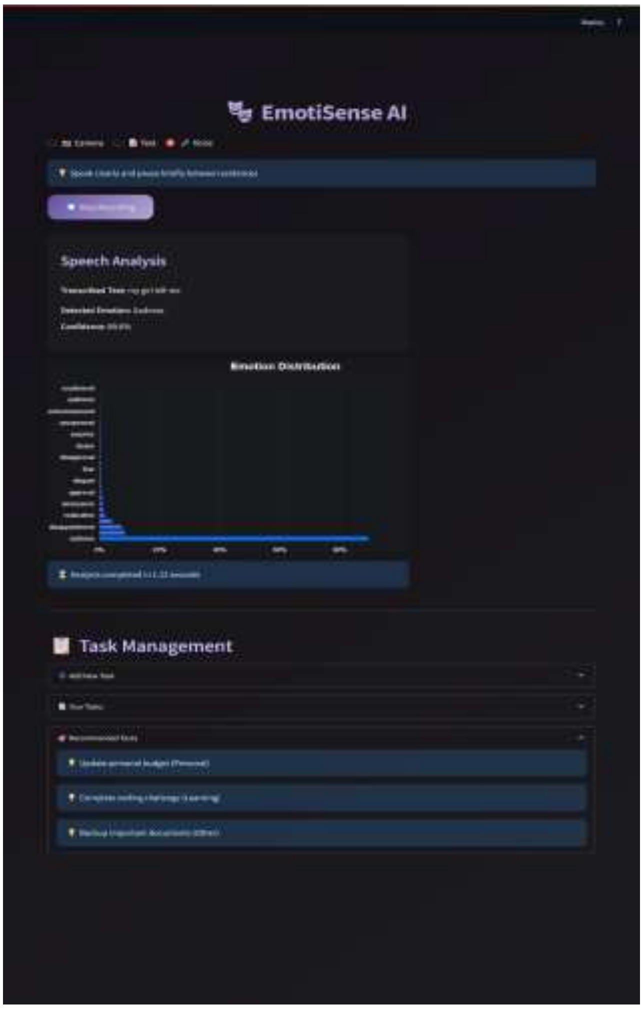
User feedback emphasized the intuitive interface and the value of personalized task suggestions, particularly highlighting the convenience of having mental state-aware productivity support without needing multiple applications.

These findings confirm EmotiSense AI's role as a valuable tool for enhancing productivity, emotional self-awareness, and digital well-being. Its multimodal, real-time emotional intelligence features significantly improve user interaction quality and promote more meaningful, tailored experiences compared to traditional task management tools.



```
⌄ datasets
  ⌄ audio
    › Actor_23
    › Actor_24
    › audio_speech_actors_01-24
  ⌄ image
    › test
    › train
  ⌄ text
    tweet_emotions.csv
  dataset link
```

Fig. 3. Sample Dataset



Update personal budget (Personal)

Complete coding challenge (Learning)

Backup important documents (Other)

Fig. 4. Response from the Chatbot

V. Conclusion and Future Work

Our system, EmotiSense AI, introduces an innovative multimodal framework that enhances personal productivity and emotional well-being by detecting user emotions across visual, textual, and auditory modalities and offering real-time, context-sensitive task recommendations. By leveraging deep learning models and an intuitive interface built on Streamlit, EmotiSense AI bridges the gap between emotional intelligence and task management, providing a more personalized, empathetic user experience.

Through sophisticated language processing, facial analysis, and speech interpretation, the system parses emotional cues to offer tailored suggestions and insights. This capability significantly reduces the cognitive effort typically required to manage daily tasks, enabling users to act based on their current emotional state without manual filtering or decision fatigue. Moreover, EmotiSense AI promotes digital inclusivity, ensuring that users with different emotional needs and communication preferences can all benefit from its services—whether they type, speak, or engage visually.

In practical deployments, this system has the potential to function not just as a personal productivity tool, but also as a supportive mental wellness assistant, educator companion, or adaptive eLearning facilitator. Its modularity and reliance on open-source technologies make it adaptable to a wide range of domains, including workplace wellness programs, telehealth check-ins, and educational tools.

Future Work

While the system achieves its core goal of multimodal emotion-based task recommendation, several areas offer exciting opportunities for enhancement:

- **Adaptive Learning**: Incorporating user feedback loops to enable real-time model refinement and personalized emotion-task mapping over time.
- **Multilingual Support**: Expanding capabilities to detect and respond to emotional cues in multiple languages, making the tool more inclusive and globally usable.
- **Emotion Fusion Techniques**: Developing hybrid methods to reconcile conflicting emotion signals from different modalities (e.g., happy tone but sad face) to improve accuracy and contextual reasoning.
- **Cloud and Mobile Deploymen**t: Optimizing the system for deployment on mobile devices and cloud platforms, expanding accessibility for users across devices and geographies.
- **Integration with Mental Health Platforms**: Collaborating with wellness apps to extend functionality for mood tracking, journaling, or therapy support.

EmotiSense AI lays a strong foundation for future human-centered AI systems that understand, adapt, and respond to human emotions in real-time, making digital experiences more natural, empathetic, and productive.

References

[1] Jin, X. (2024). Emotion recognition using machine learning: Opportunities and challenges for supporting those with autism or depression. Proceedings of the ACE Conference. https://www.ewadirect.com/proceedings/ace/article/view/15335/pdfEWA Direct

[2] Zhou, Y., & Li, H. (2024). Development and application of emotion recognition technology: A systematic literature review. BMC Psychology, 12(1), 58. https://bmcpsychology.biomedcentral.com/articles/10.1186/s40359-024-01581-4BioMed Central

[3] Rahul, M., Tiwari, N., Shukla, R., Kaleem, M., & Yadav, V. (2022). Deep learning-based emotion recognition using supervised learning. In Emerging Technologies in Data Mining and Information Security (pp. 237–245). Springer. https://link.springer.com/chapter/10.1007/978-981-19-4052-1_25SpringerLink

[4] Khan, R., & Sharif, O. (2017). A literature review on emotion recognition using various methods. Global Journal of Computer Science and Technology, 17(3). https://globaljournals.org/GJCST_Volume17/3-A-Literature-Review-on-Emotion.pdfGlobal Journals

[5] Alarcón, D., & García, M. (2022). Deep learning-based approach for emotion recognition using EEG signals. Sensors, 22(8), 2976. https://www.mdpi.com/1424-8220/22/8/2976MDPI

[6] Fu, Y., et al. (2021). Review on emotion recognition based on electroencephalography. Frontiers in Computational Neuroscience, 15, 758212. https://www.frontiersin.org/articles/10.3389/fncom.2021.758212/fullFrontiers

[7] Houssein, E. H., & Ibrahim, O. A. S. (2024). Machine learning for human emotion recognition: A comprehensive review. Neural Computing and Applications, 36, 9426–9442. https://link.springer.com/article/10.1007/s00521-024-09426-2SpringerLink

[8] Elyoseph, Z., et al. (2023). Editorial: Machine learning approaches to recognize human emotions. Frontiers in Psychology, 14, 1333794. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1333794/full Frontiers

[9] Li, Y., & Wang, X. (2023). Emotion recognition for improving online learning environments: A systematic review of the literature. Journal of Educational Sciences, 25(3), 2255. https://journal.esrgroups.org/jes/article/view/2255Journal of Electrical Systems

[10] Sharma, R., & Gupta, S. (2024). A review on emotion detection by using deep learning techniques. Artificial Intelligence Review, 57, 10831. https://link.springer.com/article/10.1007/s10462-024-10831-1SpringerLink

[11] Banos, O., et al. (2024). Sensing technologies and machine learning methods for emotion recognition in autism: Systematic review. International Journal of Medical Informatics, 177, 105132. https://www.sciencedirect.com/science/article/pii/S1386505624001321ScienceDirect

[12] Wang, Z., et al. (2023). A comprehensive survey on deep facial expression recognition: Methods and challenges. Alexandria Engineering Journal, 62(1), 32. https://www.sciencedirect.com/science/article/pii/S1110016823000327ScienceDirect

[13] Kumar, A., & Singh, R. (2022). A systematic survey on multimodal emotion recognition using learning approaches. Journal of King Saud University - Computer and Information Sciences, 34(6), 1089. https://www.sciencedirect.com/science/article/pii/S2667305322001089ScienceDirect

[14] Zhang, Y., et al. (2023). Deep learning-based EEG emotion recognition: Current trends and future perspectives. Frontiers in Psychology, 14, 1126994. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1126994/full Frontiers

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