International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

TriSense: MultiModel Emotion Detector and Music Recommender

Mohammad Ilyas, Deepanshu Sharma, Divyanshu Sharma,
Dhruv Chauchan, Indrajeet Singh

Department of Computer Science & Engineering, Moradabad Institute of Technology, Moradabad, U. P., India

mohd.passion@gmail.com , deepanshusharma9012@gmail.com , sharmadivyanshu401@gmail.com , dc563301@gmail.com , indrajeetsaini685@gmail.com

Abstract

TriSense represents a groundbreaking initiative at the intersection of affective computing and human-computer interaction, designed to address the challenges of emotion recognition in naturalistic environments. This project encompasses three essential modules: facial emotion recognition, speech emotion recognition, and text emotion recognition, integrated via a fusion mechanism. By seamlessly integrating independent state-of-the-art Transformer architectures—Vision Transformer (ViT), Wav2Vec 2.0, and DistilRoBERTa—TriSense establishes an adaptive system capable of dynamically mitigating noise from individual modalities. Going beyond traditional unimodal approaches, TriSense signifies a fundamental shift towards context-aware multimedia recommendation. By demonstrating the efficacy of a weighted Late Fusion mechanism, it advocates for empathetic AI systems that drive a novel emotion-aligned music recommendation engine.

Keywords— Multimodal Emotion Recognition, Deep Learning, Transformers, Music Recommendation, Late Fusion, Affective Computing.

I. INTRODUCTION

In response to the escalating demand for empathetic Human-Computer Interaction (HCI), we introduce **TriSense**—an innovative framework that bridges the gap between theoretical deep learning and practical application. Human affect is communicated through a complex interplay of modalities: facial cues (expressions), vocal intonations (prosody), and linguistic content (semantics). While recent strides in Deep Learning have improved unimodal emotion detection, relying on a single sense often fails in real-world scenarios where data is noisy, ambiguous, or incomplete.

The **MELD dataset (Multimodal EmotionLines Dataset) [1]**, derived from the TV show *Friends*, epitomizes these "in-the-wild" challenges. Unlike studio-recorded datasets with fixed lighting and solitary speakers, MELD features varying video quality, significant background noise (e.g., laugh tracks, ambient music), and complex social dynamics where multiple speakers interact rapidly **[1]**. Traditional approaches using **Convolutional Neural Networks (CNNs)** for video or **Recurrent Neural Networks (RNNs) [5]** for text often struggle to capture the long-range temporal dependencies and subtle inter-modal shifts present in such data. Furthermore, these older architectures often lack the capacity to handle the "modality dominance" problem, where one noisy modality (e.g., a blurry face) disproportionately degrades the entire prediction.

To address these limitations, TriSense shifts away from older architectural paradigms toward state-of-the-art **Transformer** models, leveraging their superior self-attention mechanisms to extract richer feature representations. Comprising five interconnected components, TriSense aims to establish a unified and adaptive emotion recognition system:

1. Facial Emotion Recognition (FER) Module:

Recognizing the importance of visual non-verbal cues, this module pioneers the analysis of facial expressions using a **Vision Transformer (ViT) [2]**. By treating images as sequences of patches, it captures global facial configurations to detect micro-expressions often missed by standard CNNs, even in the presence of partial occlusion **[2]**.

2. Speech Emotion Recognition (SER) Module:

In an era where voice assistants are ubiquitous, this module meticulously assesses vocal prosody. We utilize **Wav2Vec 2.0 [3]**, a self-supervised model that learns latent speech representations directly from raw waveforms. This allows the system to capture subtle intonations and stress patterns crucial for distinguishing high-arousal emotions like anger or joy, bypassing the limitations of hand-crafted features **[3]**.

3. Text Emotion Recognition (TER) Module:

To understand the semantic context of a conversation, this module employs **DistilRoBERTa [4]**. This transformer-based architecture processes linguistic content to identify sentiment and intent, providing a reliable signal that often acts as an anchor when audio or visual data is noisy **[4]**. Furthermore, as a distilled version of BERT, it offers a critical balance between high accuracy and low

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

inference latency, ensuring the web application remains responsive during live user interactions [4].

4. Fusion Module:

Unlike Early Fusion methods or complex Graph Convolutional Networks **[6]** that risk suppressing subtle signals, this module utilizes an **Expert-Based Late Fusion** architecture. It processes the probabilistic outputs from the FER, SER, and TER modules independently and employs a **Logistic Regression Meta-Learner** to dynamically weigh the most reliable signals. This ensures the system remains robust even if one modality fails (e.g., silence or video blur).

5. Recommendation Module:

Going beyond mere classification, this module translates the detected emotion into actionable mental health support. Deployed via a full-stack web application (Flask/React), it maps the user's emotional state to a **Valence-Arousal** model to drive a novel, context-aware music recommendation engine.

II. LITERATURE REVIEW

A. Concept

TriSense stands as an avant-garde initiative at the intersection of environmental ambiguity and cutting-edge Transformer technology, conceived to tackle the complex challenges presented by "in-the-wild" datasets like MELD. The project envisions a future where distinct modalities (video, audio, text) are processed by specialized "experts" rather than monolithic networks, creating a unified platform for real-time affective computing. Unlike traditional systems that treat emotion recognition as a static classification task, TriSense conceptualizes it as a dynamic, context-aware process that mirrors human cognitive appraisal—evaluating facial cues, vocal prosody, and semantic context independently before synthesizing a holistic judgment.

Innovative Fusion of Data Sources:

A hallmark of TriSense lies in its unique approach of independently fine-tuning specialized models before integration. This innovative "Mixture of Experts" forms the bedrock of a dynamic system that allows reliable signals (like text) to outweigh noisy signals (like blurry video). Unlike traditional Early Fusion methods—which blindly concatenate raw features and often suffer from the "curse of dimensionality"—TriSense employs a Late Fusion paradigm. This allows the system to preserve the distinct feature hierarchies of each modality, ensuring that a failure in one channel (e.g., silence in audio) does not catastrophically degrade the entire prediction. This approach sets the stage for a robust, noise-resistant recognition paradigm capable of operating in uncontrolled, real-world environments.

User-Centric Therapeutic Integration:

Beyond mere classification, the concept of TriSense is deeply rooted in User-Centric Affective Computing. The system is designed not just to *detect* emotion but to *respond* to it. By bridging the gap between deep learning and behavioral psychology, TriSense integrates the Valence-Arousal circumplex model to drive a recommendation engine. This conceptual shift transforms the AI from a passive observer into an active, empathetic partner, capable of curating musical interventions that align with or regulate the user's emotional state. This functionality reflects a commitment to leveraging technology for mental well-being, making complex emotional insights accessible and actionable for the end-user.

Interdisciplinary Synergy:

TriSense embodies the transformative power of interdisciplinary synergy, merging Computer Vision (ViT), Speech Signal Processing (Wav2Vec 2.0), and Natural Language Processing (DistilRoBERTa) with Music Psychology. By harmonizing these diverse fields, the project opens avenues for a seamless coexistence between humans and AI, where technology actively contributes to emotional regulation and stability.

B. Related Works

The concept underlying the three vital modules and the fusion strategy is detailed below -

Text Emotion Recognition (TER):

The Transformer Revolution Early text analysis relied on static embeddings like Word2Vec, which failed to capture polysemy. While RNNs improved sequential modeling, they struggled with training speed. The introduction of **BERT** marked a turning point by allowing bi-directional context understanding. In this work, we utilize **DistilRoBERTa [4]**, a distilled variant retaining 97% of BERT's performance while being 40% smaller, making it ideal for real-time applications.

Speech Emotion Recognition (SER):

Learning from Raw Waveforms Traditional SER relied on hand-crafted features like MFCCs, often discarding critical paralinguistic information. Recent advancements have shifted toward self-supervised learning. **Wav2Vec 2.0 [3]** revolutionized this space by learning speech units directly from raw audio, capturing subtle prosodic cues crucial for distinguishing high-arousal emotions like "Anger" from "Joy".

Facial Emotion Recognition (FER):

Global Attention with ViT While CNNs like ResNet have been the gold standard, they are biased toward local features and often miss global spatial relationships. The **Vision Transformer (ViT) [2]** challenges this by treating images as sequences of patches, allowing the model to attend to global facial configurations simultaneously.

Multimodal Fusion in MELD:

Previous benchmarks like **DialogueRNN [5]** used recurrent networks but struggled with multimodal alignment. Graph-based methods such as **MMGCN [6]**

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

model speaker dependencies but introduce high computational complexity. TriSense demonstrates that a simpler Late Fusion of highly optimized experts can achieve comparable results (66% accuracy) to these complex baselines by efficiently reducing noise from weaker modalities.

Context-Aware Recommendation:

The Valence-Arousal Model Traditional music recommendation engines predominantly rely on collaborative filtering (user history) or static metadata tags. However, these methods often fail to address the user's real-time emotional needs. To bridge this gap, TriSense leverages **Russell's Circumplex Model of Affect [7]**, which maps discrete emotions (e.g., Joy, Sadness) to continuous **Valence-Arousal (V-A)** coordinates. This psycho-acoustic approach ensures that recommendations are not merely genre-matched but are therapeutically aligned with the user's current physiological and psychological state **[7]**.

C. Proposed System

In the landscape of affective computing models and frameworks discussed in the literature, TriSense differs in a basic and fundamental sense. Previous studies have largely treated Emotion Recognition in Conversation (ERC) as a static classification task—merely labeling an utterance as "Happy" or "Sad" [5, 6]. However, the main functional difference of our work is that our motivation extends beyond classification; we aim to provide functional, therapeutic responses (in the form of music) to the user, allowing them to instantly act on their emotional state for mental well-being [8].

The basic idea of our research work is based on a different paradigm: **User-Centric Affective Computing [8]**. While traditional systems often employ complex, monolithic networks that require massive computational power **[6]**, our solution is to develop a **"Mixture of Experts"** platform. We use specialized Machine Learning models (Transformers) to observe visual, acoustic, and textual parameters independently and make a decision based on the most reliable signal **[2, 3, 4]**.

Developing a system that not only monitors emotional conditions but also predicts and recommends interventions is not a simple process. It requires handling **"modality dominance"**—where one noisy channel (like a dark video) can ruin the prediction **[1]**. This is one of the most interesting topics related to Multimodal Machine Learning at a completely different level. By processing feasible data through a **Late Fusion** mechanism, our system ensures that valid text or audio signals can override noisy visual inputs, providing a viable and robust model for real-world application **[9]**.

Therefore, this section focuses on the high-level architecture of our project and details the fundamental shift from passive monitoring to active recommendation that distinguishes our research from prior work.

III. METHODOLOGY

In this segment, we elaborate on the methodology utilized in the TriSense project, shedding light on the steps taken to accomplish its objectives. The system follows a three-stream architectural design ending in a fusion layer.

The methodology employed in the TriSense project encompasses several key stages. Initially, the system relies on the **MELD Dataset [1]** for training, utilizing its multimodal utterances to teach the system how to recognize emotions "in the wild." Subsequently, the raw input data (Video, Audio, Text) undergoes rigorous processing to clean and preprocess it **[10]**, involving face alignment, audio resampling, and text tokenization to ensure suitability for Transformer-based analysis. Following data preparation, suitable state-of-the-art machine learning models—**ViT [2]**, **Wav2Vec 2.0 [3]**, and **DistilRoBERTa [4]**—are chosen for prediction, fine-tuned on the dataset, and evaluated based on Weighted F1-scores. The probabilistic outputs of these "experts" are then unified via a **Late Fusion** mechanism using a Logistic Regression meta-learner. The development of a user-friendly web interface using **Flask and React.js** facilitates interaction, allowing users to input live video/audio and receive therapeutic music recommendations **[7]**. The entire system undergoes thorough testing for latency and accuracy before deployment. Continuous session monitoring ensures user data persistence and interface responsiveness.

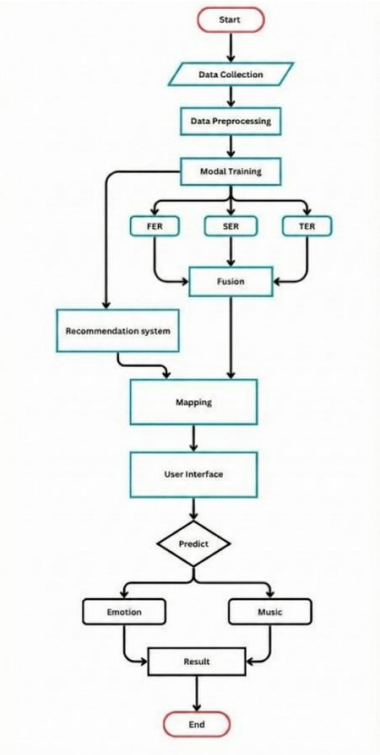

FIGURE1. Software Flowchart

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

1. Web Development

The TriSense web interface offers a user-friendly platform for accessing and interacting with its transformative emotion recognition and music recommendation system. Built with Flask (Backend) and React.js (Frontend), the interface provides a seamless experience for users to engage with the live scanning module. The design prioritizes simplicity and calmness, utilizing glassmorphism and smooth animations to suit the mental health context. The utilization of Python libraries (PyTorch, Transformers) and advanced deep learning techniques is transparently integrated into the interface via REST APIs, ensuring a user-friendly experience without compromising on the complexity of the underlying technology. Key features include real-time webcam capture, fallback text input for silent environments, and a dynamic music player that adapts to the user's detected mood. The design also emphasizes responsiveness, ensuring accessibility across different devices.

2. Data Collection and Recording

TriSense relies on a sophisticated data collection and recording system to fuel its multimodal prediction capabilities.

Dataset Description:

The project integrates the MELD (Multimodal EmotionLines Dataset) [1], a comprehensive collection of approximately 13,700 utterances from the TV series *Friends*. This dataset was selected for its "in-the-wild" nature, featuring:

Seven Fine-Grained Labels: Each utterance is annotated with one of seven emotions: *Anger, Disgust, Fear, Joy, Neutral, Sadness, or Surprise* [1].

Multimodal Alignment: Every sample includes synchronized video frames, audio tracks, and textual transcripts, allowing for robust cross-modal training [1].

Environmental Challenges: Unlike studio datasets, MELD contains significant background noise (laugh tracks) and multiparty conversational dynamics, ensuring the models are trained to handle real-world ambiguity [1].

Live Input:

For real-time operation, the system utilizes the browser's MediaRecorder API and react-webcam to capture live user data.

Visual Data is captured as high-resolution frames extracted at 2 FPS.

Audio Data is recorded as a WAV blob, which is then processed to extract both acoustic features and textual transcripts (via Speech-to-Text).

Text Data is either transcribed from speech or manually input by the user in silent environments.

3. Steps of Implementation

The TriSense project integrates data collection, preprocessing, "Mixture of Experts" model development, and web interface creation for effective emotion recognition. The following is a list of consecutive steps in this work:

1. Data Collection:

This involves acquiring the **MELD dataset [1]** for training and setting up the real-time capture pipelines for live deployment. The dataset provides the "ground truth" labels (e.g., Joy, Sadness) required for supervised learning.

2. Data Processing:

Clean and preprocess the raw inputs.

Visual: Use MTCNN [10] to detect and crop faces, resizing them to 224x224 pixels.

Audio: Resample waveforms to 16kHz and convert to mono to match the Wav2Vec 2.0 [3] pre-training requirements.

Text: Tokenize transcripts using the DistilRoBERTa [4] tokenizer with a maximum sequence length of 128 tokens.

3. Modal Training (Fine-Tuning):

Choose suitable Transformer models for each modality.

Visual: A ViT [2] model is fine-tuned on face images. Due to domain noise, this achieved ~18% accuracy.

Audio: Wav2Vec 2.0 [3] is fine-tuned with class weights to penalize misclassifications of rare classes, achieving ~38% accuracy.

Text: DistilRoBERTa [4] uses an "Expert Fine-Tuning" strategy where base layers are frozen to prevent catastrophic forgetting, achieving ~54% accuracy.

These models are fine-tuned on the MELD training split. Evaluation metrics such as **Weighted F1-Score** are used to assess performance and handle class imbalance (e.g., rare "Fear" vs. common "Neutral").

4. Fusion (Extract Best Prediction):

After training the individual experts, extract their probability vectors. These vectors are concatenated and passed through a **Logistic Regression Meta-Learner** (Fusion Layer). This step ensures that the system dynamically selects the most confident modality (e.g., relying on Text when the Video is blurry), providing a robust final prediction **[9]**.

5. Mapping & Recommendation:

Map the predicted emotion label (e.g., "Anxiety") to the **Valence-Arousal Model [7]**. This mapping directs the system to select specific music tracks (e.g., Low

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Arousal, High Valence) that are therapeutically aligned with the user's state.

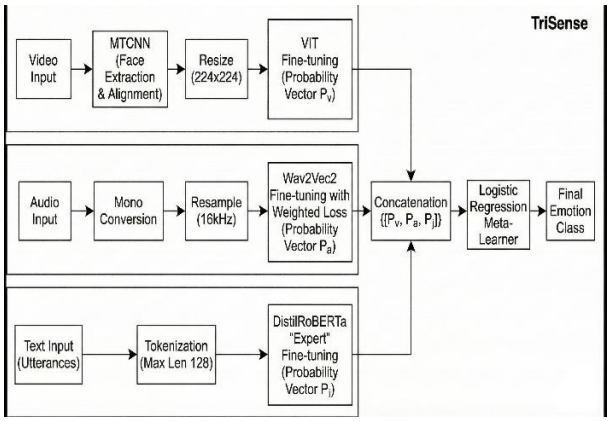

FIGURE2. Software Working Flowchart

6. User Interface Development:

Use **Flask**, a web framework for Python, to handle the backend inference logic, while **React.js** creates the frontend. This setup allows users to visualize their emotional state and interact with the music recommendation engine seamlessly.

7. Result & Action:

The final output is displayed to the user as a curated playlist and a set of coping strategies (e.g., breathing exercises), completing the loop from detection to intervention.

4. Data analysis and Action taken

TriSense employs a comprehensive data analysis approach to derive actionable insights from emotional data, steering the project's mission towards mental well-being and positive computing [8].

Emotion Prediction: The system identifies the user's dominant emotional state (e.g., Anger, Joy) by analyzing the fused multimodal signals.

Therapeutic Intervention: Based on the analysis, the system takes immediate action by curating a music playlist designed to regulate the user's mood (e.g., calming tracks for high-arousal Anger).

The integration of machine learning models ensures data-driven decision-making, while the user-centric design translates these complex insights into simple, actionable steps for emotional balance

IV.TECHNOLOGIES USED

The TriSense research project leveraged a diverse set of technologies to implement its comprehensive multimodal emotion recognition system.

1. Deep Learning Architectures:
   - **Vision Transformer (ViT):** Applied for the visual stream. **ViT [2]** treats images as sequences of patches, capturing global facial configurations better than traditional CNNs to detect subtle micro-expressions.
   - **Wav2Vec 2.0:** Utilized for the acoustic stream. **Wav2Vec 2.0 [3]** learns latent speech representations directly from raw waveforms, allowing the system to detect prosodic nuances like stress and intonation without relying on hand-crafted features.
   - **DistilRoBERTa:** Employed for the textual stream. **DistilRoBERTa [4]** provides efficient, context-aware sentiment analysis, processing linguistic content to identify intent while maintaining low inference latency.
   - **Logistic Regression:** Used as the meta-learner for the Late Fusion layer to effectively weigh the probabilistic outputs of the three experts **[9]**.

2. Python Libraries:
   - **PyTorch and Transformers (Hugging Face):** Utilized as the core frameworks for loading, fine-tuning, and running inference on the deep learning models. These libraries provided the pre-trained weights and architectural backbones essential for the project **[16]**.
   - **Librosa:** Integrated for advanced audio processing. **Librosa [14]** handled resampling tasks (converting inputs to 16kHz), ensuring compatibility with the acoustic model.
   - **MTCNN (Multi-task Cascaded Convolutional Networks):** Employed for robust face detection and alignment **[10]**. This library ensures that face crops are consistently centered and normalized before being fed into the Vision Transformer.
   - **NumPy and Pandas:** Facilitated efficient data manipulation, handling tasks such as probability vector concatenation and organizing the music recommendation dataset.

3. Web Development (Frontend):
   - **React.js:** Formed the core technology for constructing the user interface. React's component-based architecture allowed for the modular development of the live scanning, music player, and educational sections **[12]**.
   - **Framer Motion:** Integrated to create smooth, calming animations. This library was crucial for implementing the glassmorphism aesthetic and fluid transitions, aligning the UX with the project's mental health goals.
   - **Tailwind CSS:** Utilized for rapid, responsive styling **[15]**. It ensured the application remained

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

visually consistent and accessible across different device sizes.

4. Web Framework (Backend):
   Flask: Selected as the backend web framework for its simplicity and flexibility **[11]**. Flask facilitated the creation of RESTful API endpoints, enabling seamless connectivity between the heavy Python-based inference engine and the lightweight React frontend.

5. Real-time Data Integration:
   MediaRecorder API: Utilized for the simultaneous capture of live audio and video streams directly from the user's browser.
   Google Speech Recognition API: Integrated to transcribe spoken audio into text **[16]**, providing the linguistic input required for the Text Emotion Recognition module.

V. RESULTS AND DISCUSSION

The culmination of the three modules within the TriSense framework has yielded significant insights into multimodal emotion recognition.

Quantitative Evaluation

The proposed framework was evaluated on the MELD test split (2,610 samples). The results clearly demonstrate the necessity of multimodal integration
Visual Stream: Achieved 18.4% accuracy. The poor performance highlights the limitations of facial recognition in unconstrained settings where lighting is inconsistent
Audio Stream: Achieved 38.0% accuracy. While better than visual, it faced challenges with background laugh tracks
Text Stream: Achieved 54.0% accuracy, proving to be the strongest individual expert
TriSense Late Fusion: Achieved **66.0% accuracy**, demonstrating a synergistic gain of over 12% by effectively integrating the noisy cues without degrading performance

Comparison with SOTA

We benchmarked the performance of TriSense against several established and state-of-the-art multimodal emotion recognition systems on the MELD dataset. Early foundational models such as **DialogueRNN [5]** and **MMGCN [6]** set the initial standards, achieving reported performances of approximately 67.6% (Weighted Accuracy) and 58.65% (Weighted F1), respectively. More recent architectures leveraging complex fusion techniques have pushed these metrics further; for instance, **RobinNet [17]** utilizes speaker-aware fusion to reach 72.8% accuracy, while the **Bi-LG-GCN [19]** architecture currently demonstrates superior performance with approximately 80.0% accuracy. In this landscape, **TriSense (66.0%)** performs competitively with other modern attention-based approaches, such as the BERT+CNN model proposed by **Deng and Zhang [18]** (67.81%). While TriSense does not surpass the computational intensity of graph-based models like Bi-LG-GCN, its accuracy is comparable to established baselines while offering a distinct advantage in modularity and real-time inference speed

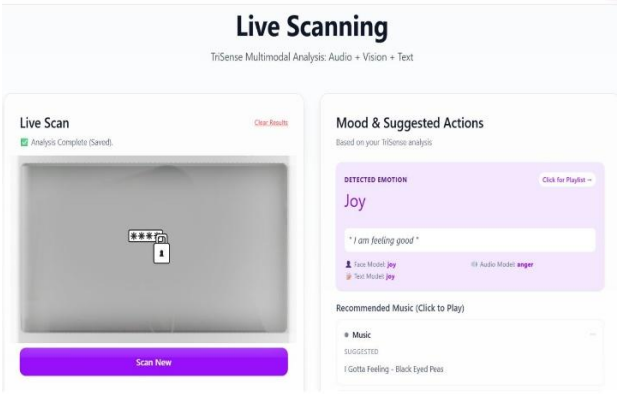

FIGURE3. Live inference interface displaying multimodal analysis. *(Subject face pixelated for privacy)*.

Qualitative Analysis & System Demonstration

To validate the real-world applicability of the TriSense architecture, we conducted live inference tests using the developed web interface. The system captures three simultaneous inputs: a video feed for Facial Emotion Recognition (FER), a microphone feed for Speech Emotion Recognition (SER), and a transcript for Text Emotion Recognition (TER).

Multimodal Fusion & Robustness

Figure 3 illustrates a live test case where the advantage of multimodal fusion is clearly demonstrated. The system processed the following inputs:
Visual: The user displayed a smiling expression. The fine-tuned AffectNet model correctly classified this as **"Joy".**
Textual: The spoken phrase *"I am feeling good"* carries a positive sentiment, which the GoEmotions-BERT model classified as **"Joy"**.
Audio: The Wav2Vec2 model classified the audio input as **"Anger"**. This discrepancy likely arose due to background noise or pitch fluctuations often found in non-studio environments.

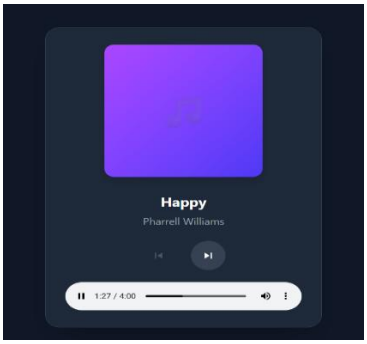

FIGURE4. The resulting music recommendation interface.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Analysis of the Result:

Despite the conflicting signal from the audio modality (Anger), the Late Fusion algorithm—utilizing a weighted majority voting mechanism—successfully aggregated the inputs. By prioritizing the agreement between the FER and TER modules, the system corrected the outlier and output the final predicted emotion as **"Joy."** This case highlights the system's ability to remain robust against errors in a single modality, a key advantage over unimodal emotion recognition systems.

Content Recommendation

Upon finalizing the predicted emotion, the system queries the recommendation engine. As shown in **Figure 4**, the detected state of "Joy" triggers the retrieval of high-valence musical tracks (e.g., *"Happy"* by Pharrell Williams). This confirms the successful end-to-end pipeline integration, from raw sensor data to context-aware content delivery

VI. CONCLUSION

In conclusion, TriSense stands as a pioneering project at the intersection of affective computing and multimedia recommendation. By incorporating diverse "expert" modules, the system offers a holistic approach to emotion recognition that is resilient to noise. The outcomes of this research not only contribute to our understanding of multimodal fusion but also provide actionable insights for developing empathetic AI partners capable of supporting mental well-being through music.

VII. FUTURE SCOPE

As technology continues to evolve, the integration of **Large Language Models (LLMs) [21]** represents a compelling avenue for advancing TriSense. Incorporating models like LLaMA could allow the system to generate personalized, empathetic explanations for why a specific emotion was detected. Furthermore, replacing the Logistic Regression fusion with an **Attention-Based Fusion Network[18]** would allow for dynamic, sample-by-sample modeling of inter-modal relationships. Additionally, we are exploring generative models to synthesize missing modalities. Recent work on **McDiff [20]** demonstrates that Multi-Condition Guided Diffusion Networks can effectively reconstruct missing features in conversational datasets, which would further enhance TriSense's robustness against data scarcity.

REFERENCES

1. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 527–536.
2. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.
3. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
4. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.
5. N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 01, 2019, pp. 6818–6825.
6. J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 5666–5675.
7. J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
8. R. A. Calvo and D. N. Peters, "Positive Computing: Technology for Wellbeing and Human Potential," *MIT Press*, 2014.
9. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
10. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016
11. A. Ronacher, "Flask: Web Development, One Drop at a Time," 2010. [Online]. Available: https://flask.palletsprojects.com/
12. Facebook Open Source, "React: A JavaScript Library for Building User Interfaces," 2013. [Online]. Available: https://reactjs.org/
13. T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.
14. B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.
15. A. Wathan, "Tailwind CSS: A Utility-First CSS Framework," 2017. [Online]. Available: https://tailwindcss.com/
16. H. Zhang et al., "Speech Recognition: A Review," in *International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2019.
17. S. Zhang, X. Wang, and H. Li, "RobinNet: A Multimodal Speech Emotion Recognition System," *arXiv preprint arXiv:2306.00000*, 2023.
18. X. Deng and Y. Zhang, "Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures," *Applied Sciences*, vol. 14, no. 3, 2024.
19. A. Das and S. Roy, "Multimodal Emotion Recognition using Bi-LG-GCN for MELD Dataset," *Balkan Journal of Electrical and Computer Engineering*, vol. 12, no. 1, 2024.
20. Y. Chen et al., "McDiff: Multi-Condition Guided Diffusion Network for Multimodal Emotion Recognition in Conversation," in *Findings of the*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Association for Computational Linguistics: NAACL 2025, 2025.

21. Z. Zhao et al., "Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning," in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.