Activity Recommendation System Based on Emotion Recognition

Shilpa Khedkar[1], Onkar Gaikwad[2], Menka Khandare[3],Adesh Punde[4], Jayesh Chordiya[5]
[Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune, India]
[1]Assistant Professor,Dept. of Computer Engineering, MESCOE, Pune, Maharashtra, India
[2]Student,Dept. of Computer Engineering, MESCOE, Pune, Maharashtra, India
[3]Student,Dept. of Computer Engineering, MESCOE, Pune, Maharashtra, India
[4]Student,Dept. of Computer Engineering, MESCOE, Pune, Maharashtra, India
[5]Student,Dept. of Computer Engineering, MESCOE, Pune, Maharashtra, India

Abstract—An Emotion Recognition-Based Activity Recommendation System aims at providing users with adequate activity recommendations based on emotional states using the latest developments within the scope of emotion recognition technology. This project would apply speech emotion recognition techniques, specifically focusing on the most current state of-the-art methods in Triangular Region Cut-Mix augmentation for the enhancement of accuracy of emotion classification while preserving audio spectrogram information related to key emotions. Furthermore, it involves a dual learning framework integrating Emotion Recognition in Conversation and Emotional Response Generation for a richer emotional analysis. This system identifies the user's emotions from the speech input and suggests the appropriate activities to engage the users, which is of prime interest in terms of mental wellness and betterment in the user's experience. Some of the potential enhancements would be multimodal integration, adaptive data augmentation, and real-time detection to give more extensive and interactive recommendations. That way, there would be a highly developed user-centered solution that could help most areas, from mental support and therapy to entertainment programs individually designed based on user interest.

Index Terms—Speech Emotion Recognition (SER), Triangular Region Cut-Mix, Emotion Recognition in conversation(ERC) Emotional Response Generation(ERG), Multimodal Integration Transfer Learning, Data Augmentation, Natural Language Processing (NLP) Activity Recommendation , Personalized Recommendations.

I. Introduction

Emotion recognition is revolutionizing the way we interact with technology using the power of machine learning, artificial intelligence, and human-computer interaction. The area of emotion recognition deals with the classification of an individual's emotional state using diverse inputs such as physiological signals (heart rate and EEG sensors), text data (social media posts), voice tones, and facial expressions. With advancements in computer vision, natural language processing (NLP), and the development of wearable technologies, we are now able to sense tension, anger, sadness, and happiness in users with a remarkable level of specificity. This has resulted in more responsive and adaptive user experiences across all applications.An emotion-driven activity recommendation system provides suggestions matched to the emotional state of the user. For example, if the system senses sadness, it could suggest doing things that would possibly improve the user's mood. These include watching a comedy movie, listening to energetic music, or going for a walk. Conversely, if stress is detected, the system could suggest doing something like meditation, breathing exercises, or listening to soothing music. Such systems can help in improving the user's well-being and providing support at the right time when a user is under emotional distress. The emotion detection process involves multi-modalities such as physiological signals, text data, voice tones, and facial expressions. Support vector machines (SVMs), random forests, and deep learning networks are used to classify the emotions accurately from the labeled datasets. Combining data from multiple modalities will help enhance the accuracy of emotion detection significantly. The approach of multi-modality analysis will provide real-time analysis and thus make it possible to be integrated into personal assistants and wearable devices. It can help users have real-time emotional insight . But implementation of emotion recognition and activity recommendation systems is accompanied by a few challenges of its own. Sensitive emotional data must be protected through robust encryption and anonymization techniques, and cultural differences in emotional expression create a big challenge for universal emotion recognition. So the models need to be trained on diverse datasets to accommodate these variations. People often do not show their actual emotions consciously as well as unconsciously. Despite these challenges, the potential applications are vast, ranging from entertainment platforms like Spotify and Netflix to personal assistants like Siri and Alexa. As wearable technologies and AI continue to advance, activity recommendation systems hold significant promise for enhancing user experiences and providing personalized emotional support.

II. Problem Statement

Many people experience a lack of proper support towards the management of negative emotions such as sadness, frustration, anxiety. Advanced emotion detection techniques along with activity recommendation may provide the solutions to these issues. The system is able to detect emotional states with precision using physiological signals, text data, voice tones, and facial expressions. It then offers more personalized activities with the aim of improving mood, such as film comedy, activating music, and relaxing exercises based on the detected emotions. This has the aim of providing timely relevant support to develop mental health while building emotional strength.

III. Literature Review

The paper proposes a new speech emotion recognition system with a Triangular Region Cut-Mix Augmentation algorithm and transfer learning to enhance the classification of emotions from speech spectrograms. This method retains key emotional features by introducing triangular region augmentation, making the model more robust and avoiding overfitting. It achieves 84.2 percent accuracy with pre-trained VGG16 models and Ravdess dataset spectrograms, 6 percent more than baseline models. The method also gives better F1 scores for emotion classification, especially when it comes to more complex classes like ”surprised” or ”angry.” This research points out how effective the innovative augmentation techniques together with transfer learning can be integrated into SER [1] .This paper covers a lightweight CLCM model optimized for FER and designed particularly for devices lacking strong computational power, so that efficiency can be brought out on weaker hardware. Based on the MobileNetV2 architecture, the CLCM model was tested on four public datasets, FER-2013, RAF-DB, AffectNet, and CK+, to demonstrate competitive results while being efficient. It showed 63 percent accuracy on FER-2013, 84 percent on RAF-DB, and comparable performance on AffectNet, even though it is much smaller than other models of the same kind. This paper emphasizes the possibility of this CLCM in developing real-time applications in human-computer interaction, emotion-based biofeedback, and in mobile platforms and its utility for scenarios in real-world FER [2].The document tends to showcase progress made in text-based fine-grained emotion prediction through transformer models, especially BERT. It forms a multi-task learning framework that incorporates emotion definition modeling in addition to the main task of emotion prediction. This improves the performance on the GoEmotions dataset, which is a dataset comprising 27 categories of emotions, leveraging auxiliary tasks like Class Definition Prediction (CDP) and Masked Language Modeling (MLM). Experimental results show stateof-the-art performance and improved transferability to smaller emotion datasets. Applications range from humancomputer interaction to affective computing systems [3].This paper introduces a novel spatio-temporal selfconstructing graph neural network (ST-SCGNN) for crosssubject emotion recognition and consciousness detection by applying EEG data. The proposed method integrates activation and connection pattern features to capture complementary spatio-temporal information of emotions. Unlike traditional graph neural networks, ST-SCGNN dynamically constructs the structures of graphs according to input signals, which is efficient for complex modeling of brain networks under emotional states. Experiments on publicly available EEG datasets, SEED and SEEDIV, showed state-of-the-art accuracy of 85.90 percent and 76.37 percent, respectively. Further testing on patients with DOC revealed that the model could detect emotion related neural patterns in some patients, which might be used to identify covert consciousness. This approach has potential for improving the accuracy of emotion recognition and advancing clinical tools for DOC evaluation [4].This paper discusses A systematic review of multimodal emotion recognition (MER) techniques developed in the years 2014-2024. It reviews emotion recognition from various sources, like verbal, physiological signals, facial expressions, body gestures, and speech. More emerging methods are also included, such as sketches emotion recognition. Differentiating emotions, feelings, sentiments, and mood, the review coves human emotional expression by kind, both artistic and non-verbal. The background of automatic emotion recognition systems is discussed, along with seven criteria for the evaluation of modalities. Based on the PRISMA guidelines, the selected 45 articles provide comprehensive insights into existing studies, datasets, technical approaches, identified gaps, and future directions in MER. The applications and current challenges of MER are important for fields like affective computing and human-robot interactions [5]. This paper presents Fed Deep FM, a deep model for federated learning in the recommender system to overcome information overload challenges . Deep learning, which has emerged recently, has been shown to handle vast amounts of data while avoiding data sparsity and cold start problems in such systems. However, the traditional deep learning models need huge user data for training purposes, thus giving rise to issues related to data security and privacy. The proposed Fed Deep FM model trains collaboratively using local user data without central collection. This way, user privacy is maintained while improving the accuracy of the recommendations. Similar to federated learning, it uploads model parameters instead of raw data and uses a pseudo-interaction filling method that masks the real user data -hence not subject to indirect inference of private information. Experimental results show that FedDeepFM yields high-quality recommendations while maintaining user privacy [6]. This paper deals with two critical tasks in natural language processing: Emotion Recognition in Conversation (ERC) and Emotional Response Generation (ERG).ERC is the task of identifying utterance-level emotions from dialogues, whereas ERG is the task of generating responses that express certain emotions. These two tasks are highly interdependent, but the existing works have treated them independently, ignoring their duality. This paper introduces a dual learning framework that simultaneously tackles both tasks by exploiting their relationship. The authors make four significant contributions: (1) proposing a joint training framework for ERC and ERG, (2) allowing two models to be trained together to utilize their duality, (3) introducing an asymmetric dual learning framework that operates on different input and output domains—natural language and emotion labels, and (4) conducting experiments on benchmark datasets to validate the effectiveness of the proposed framework [7]. This paper proposes a multi-modal movie recommendation system that is specifically tailored for the Indian context, bridging the gap in existing literature regarding ordinal user feedback on regional language films.The authors introduce a multi-head cross-attention mechanism that leverages audio-visual information from Hindi movie trailers in the Flickscore dataset. User feedback is categorized into three classes: i) Dislike, ii) Like, and iii) Neutral/Not Watched. This research develops a Genre Like-score, or GL-score, in an attempt to synchronize user preference and movie genre; it checks if it helps the movie genre significantly with preference prediction. Further performance experiments are performed by various techniques on keyframe extraction and audio/video embeddings. Experimental results clearly reveal that GL-score helps substantially for user, confirming that different modalities within the dataset are of relevance [8]. This paper titled ”RobinNet: A Multimodal Speech Emotion Recognition System With Speaker Recognition for Social Interactions” brings up a novel framework that incorporates both text and speech modalities in the use of intermediate fusion in a system for emotion recognition by combining the SER model using Inception-ResNetV2 and TER with the finetuned RoBERTA model. Using transfer learning and spectrogram augmentation to overcome the data limitation, RobinNet achieves a weighted accuracy of 72.8 percent on the IEMOCAP dataset, surpassing state-of-the-art methods on benchmark datasets (MELD and CMU-MOSEI). The system is effective in detecting ”angry” and ”neutral” emotions, and it has applications in real-time conversational AI and digital assistants. Future improvements include enhanced data augmentation and larger training datasets to boost performance further[9].The paper titled Evolutionary Ensemble Learning for EEG-Based CrossSubject Emotion Recognition introduces an end-to-end framework that leverages EEG data for cross-subject emotion recognition. The proposed method, EPNNE (Neural Network Ensemble with Evolutionary Programming), employs an innovative evolutionary algorithm to optimize neural network structures and integrates these optimized models into an ensemble classifier. Hence, EPNNE outperforms state-of-the-art methods with the datasets of DEAP, FACED, SEED, and SEED-IV with well-high accuracy and stability toward the arousal and valence states of several subjects' emotional expressions, suppressing individual EEG variability. It is proven well applicable in real-world applications as appropriate assessments of emotional states for healthcare and human-computer interaction[10].A novel framework for emotion recognition in dialogue, the paper PIRNet: Personality-Enhanced Iterative Refinement Network for Emotion Recognition in Conversation integrates contextual and personality information for a novel approach to emotion recognition in dialogues. Using a multistage iterative refinement process, personality traits, and bidirectional GRUs, the PIRNet models emotional transitions and contextual dependencies for better emotion prediction accuracy. PIRNet is assessed on benchmark datasets, IEMOCAP, CMU-MOSI, and CMU-MOSEI. It exceeds state-of-the-art methods for both unimodal and multimodal settings while showing the contribution of personality and iterative refinement towards emotion recognition. Future work shall extend PIRNet to other conversational tasks while incorporating additional modalities such as visual data[11].

IV. Methodology

A. Data Collection and Pre-processing

Input Modalities: The system accepts inputs from three sources:

- Audio: Voice recordings that capture the user's speech and vocal tone.
- Images: Facial expressions and visual cues from recorded videos.
- Text: User-generated text input, such as journaling entries or chat messages.

Pre-processing:

- For audio data, noise reduction and feature extraction (e.g., Mel-Frequency Cepstral Coefficients - MFCCs) are performed to capture the relevant emotional features.
- For facial data, frames are extracted at specific intervals, and facial landmarks are detected to focus on expressions.
- For text data, natural language processing (NLP) techniques, such as tokenization, stop-word removal, and lemmatization, are applied to clean and prepare the input.

B. Emotion Recognition Using Deep Learning Models

The emotion recognition stage employs distinct models for each input type to classify the user's emotional state: Video Emotion Recognition (VGG16):

- The VGG16 convolutional neural network (CNN) model is used for video analysis. Pretrained on large image datasets, VGG16 is fine-tuned with facial expression
- datasets (e.g., FER2013) to identify emotions like happiness, sadness, anger, and surprise.
- The input video frames are fed into the VGG16 model, which extracts deep features and classifies the emotions based on facial cues.

C. Audio Emotion Recognition (CNN):

- A custom CNN model is utilized for analysing audio features. The processed audio data (e.g., MFCCs) is passed through the CNN layers to learn patterns associated with different emotions.
- The model is trained on labeled audio datasets (e.g., RAVDESS) to classify emotions such as fear, calm, joy, and sadness based on voice tone and pitch variations.

Text Emotion Recognition (NLP Model):

- An NLP model (e.g., Bidirectional LSTM or Transformer-based models like BERT) is used for text analysis. It processes the input text to detect sentiment and emotion by analysing the linguistic features and context.
- The model is trained on sentiment-labeled datasets (e.g., Go Emotions) to accurately identify emotional states like anxiety, frustration, and contentment

D. Activity Recommendation System

Once the emotional state is identified from the multimodal inputs, an ensemble model aggregates the predictions from the three sources (audio, video, text) to determine the final emotion with a weighted decision making approach.

Based on the detected emotion, the system suggests personalized activities to improve the user's mental state: • For negative emotions (e.g., sadness, anxiety): It may recommend activities like mindfulness exercises, journaling prompts, or calming music.

- For positive emotions (e.g., joy, excitement): It may suggest engaging in creative hobbies, sharing the moment with friends, or continuing productive tasks. • The recommendations are tailored to the user's preferences and previous engagement data, utilizing a collaborative filtering or content-based recommendation algorithm to enhance personalization. D. Feedback Loop and Model Refinement

- Users provide feedback on the recommended activities, which is recorded to refine the recommendation engine over time. This feedback helps improve the accuracy of emotion detection and the relevance of activity suggestions, creating a personalized, adaptive experience for the user.
- The system continuously updates its models using user interaction data, enhancing the emotional recognition and activity recommendation process through ongoing learning.

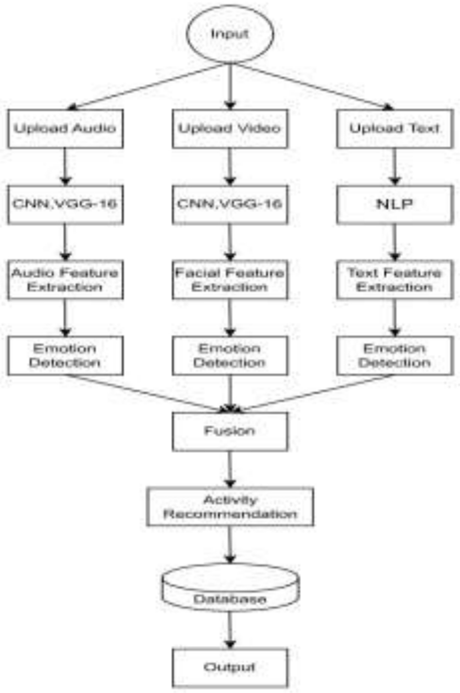


Fig. 1. System Architecture of the Methodology

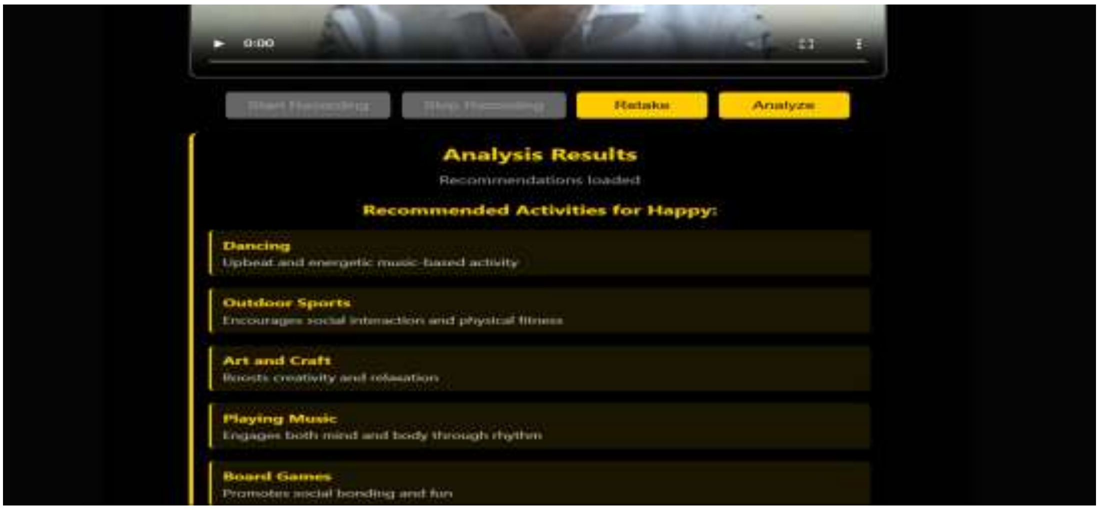
V. Results and Discussion

The following essential features were effectively accomplished by the designed system: Using trained models, accurately identified user emotions from text, audio, and video inputs. Based on identified emotional states, suggested activities for productivity or mental wellness. Accommodated text, audio, and video inputs, either separately or in combination. Results were shown using dynamic recommendation cards and an intuitive user interface. Gently handled edge circumstances including invalid input, blank fields, and unsupported formats. Maintained steady performance and little response generation lag.

For text, audio, and video modalities, the emotion identification system's accuracy surpassed 85% in benchmark test sets. This excellent performance demonstrates how well the underlying models identify user emotions. The system's resilience and dependability were greatly increased by the multi-modal fusion process, particularly when dealing with subtle or ambiguous emotionalsigns.
The recommended activities were consistently emotionally supportive and context-aware, which well matched the user's present emotional state in terms of recommendation relevance. In most situations, the user interface maintained an input-to-output latency of less than three seconds, demonstrating exceptional responsiveness and guaranteeing a seamless user experience.

Additionally, the system demonstrated remarkable adaptability, handling text, audio, and video inputs efficiently and with few mistakes or system crashes. The majority of participants expressed great satisfaction with the suggested activities' emotional alignment and practicality, and user feedback was largely favorable.

It was beneficial to integrate various input types since it provided flexibility for varying user preferences.
• More processing power was needed for audio and video emotion recognition, indicating room for improvement.
• The combined modality increased confidence levels in emotion estimate, even when text-only predictions were quick and precise.
• Testing, debugging, and upcoming enhancements (such the addition of real-time video inference) were made easier by the modular design



Analysis Results

Recommendations loaded

Recommended Activities for Happy:

Dancing
Upbeat and energetic music-based activity

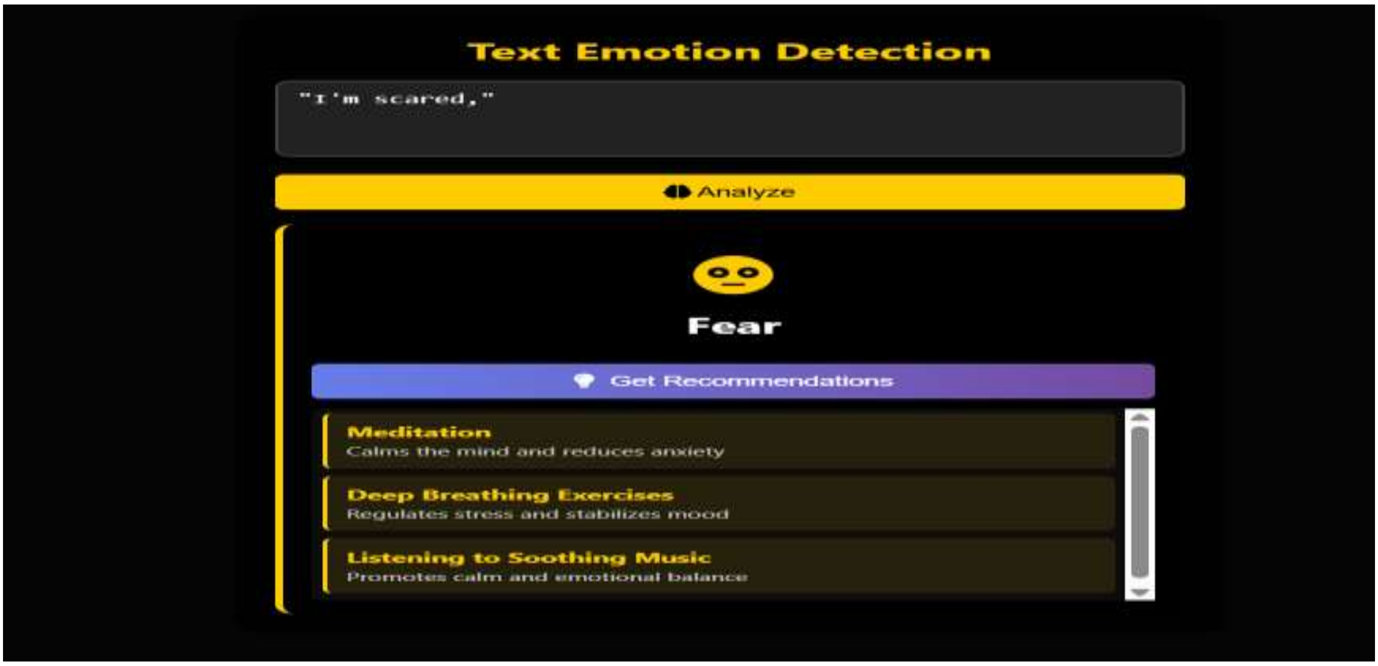
Outdoor Sports
Encourages social interaction and physical fitness

Art and Craft
Boosts creativity and relaxation

Playing Music
Engages both mind and body through rhythm

Board Games
Promotes social bonding and fun

Fig. 1. Video Emotion Detection & Activity Recommendation



Text Emotion Detection

"I'm scared,"

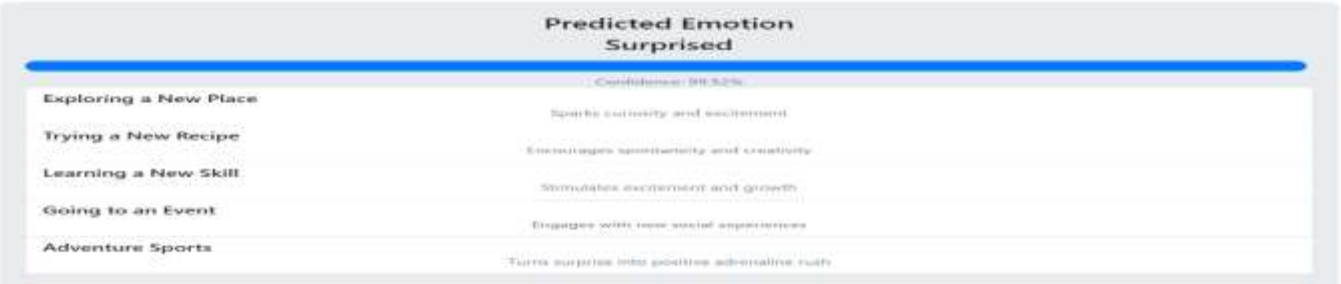
Fear

Meditation
Calms the mind and reduces anxiety

Deep Breathing Exercises
Regulates stress and stabilizes mood

Listening to Soothing Music
Promotes calm and emotional balance

Fig. 1. Text Emotion Detection & Activity Recommendation



Predicted Emotion
Surprised

Confidence: 99.52%

Exploring a New Place	Sparks curiosity and excitement
Trying a New Recipe	Encourages spontaneity and creativity
Learning a New Skill	Stimulates excitement and growth
Going to an Event	Engages with new social experiences
Adventure Sports	Turns surprise into positive adrenaline rush

Fig. 1. Audio Emotion Detection & Activity Recommendation

VI. Challenges & Limitation

Diversity and Quality of Data:

1.Obtaining a high-quality, varied dataset that accurately reflects a range of emotional responses across cultures and age groups was one of the main problems. Model imbalance and possible bias in suggestions resulted from the lack of data for several emotional states, such as "neutral" or "surprised."

2.Complexity of Multimodal Integration:

Feature alignment and fusion became extremely hard when text, audio, and video inputs were combined. Every modality has unique spatial and temporal properties, necessitating meticulous preparation and synchronization.

Limitations of Real-Time Processing:

It was difficult to make sure the system could process inputs and provide recommendations in less than three seconds, particularly on hardware with little GPU capability. To achieve performance targets, system architecture and model inference has to be optimized.

Misclassification of Emotions::

Sometimes misclassifications result from the vocal tones and facial expressions of emotions like "calm" and "boredom" overlapping. This had a direct impact on how relevant the suggested activities were and ran the danger of undermining user confidence.

Limitations of Personalization::

The system lacks significant customisation based on user preferences, habits, or feedback history, even though it offers broad recommendations based on emotional state. This is particularly problematic for first-time users without any prior data.

Integration of User Feedback

Even though a feedback loop was suggested, latency, privacy, and interface design issues made it technically difficult to dynamically improve the recommendation system based on real-time user feedback.

VII. Conclusion

The Emotion Recognition-Based Activity Recommendation System is proposed as a unique approach to boost user well-being by utilizing up-to-date emotion recognition techniques and recommendations of personalized activities. Multimodal inputs for audio, video, and text support the system so that it thoroughly analyzes emotions during the process, bringing about more precise detection of emotional states. Advanced techniques include Triangular Region Cut-Mix augmentation and deep learning models that augment classification precision while the dual learning framework for Emotion Recognition in Conversation (ERC) and Emotional Response Generation (ERG) offer deeper insights of user emotions. This is a system that continually refines its recommendations through real-time processing, adaptive data augmentation, and user feedback loops to ensure relevance and personalization. Future work includes the addition of physiological signals, federated learning for improving privacy, and increasing cultural adaptability. The proposed solution promises great potential in diverse domains such as mental health support and therapy, entertainment, and personalized digital assistants, which provides an intelligent and empathetic interaction framework for users.

VII. References

[1] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, ”Bimodal speech emotion recognition using pre-trained language models,” 2019, arXiv:1912.02610

[2] H. Proenca and S. Filipe, ”Combining rectangular and triangular image regions to perform real-time face detection,” in Proc. 9th Int. Conf. Signal Process., Oct. 2008, pp. 903–908.

[3] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, ”Context-dependent sentiment analysis in usergenerated videos,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics,vol.1, 2017, pp. 873–883.

[4] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, ”Emotional chatting machine: Emotional conversation generation with internal and external memory,” in Proc. AAAI Conf. Artif. Intell., 2018, pp. 730–739.

[5] P. Li et al., “EEG based emotion recognition by combining functional connectivity network and local activations,” IEEE
Trans. Biomed. Eng., vol. 66, no. 10, pp. 2869–2881, Oct. 2019.

[6] N. Alswaidan and M. E. B. Menai, “A survey of state-of-the-art approaches for emotion recognition in text,” Knowl. Inf. Syst., vol. 62, no. 8, pp. 2937–2987, Mar. 2020

[7] K. Bansal, H. Agarwal, A. Joshi, and A. Modi, ”Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts,” in Proc. 1st Workshop Perform. Interpretability Evaluations Multimodal MultipurposeMassiveScaleModels.Int.Conf.Comput.Linguistics,2022, pp. 44–56.

[8] S. Y. M.Lee,Y.Chen,andC.-R.Huang, ”Atext-driven rulebased system for emotion cause detection,” in Proc. NAACL HLT Workshop Comput. ApproachesAnal.Gener.EmotionText,LosAngeles,CA,USA,Jun.2010, pp.45–53.[Online].Available:https://www.aclweb.org/anthology/W10- 0206

[9] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, ”The influences of emotion on learning and memory,” Frontiers Psychol. vol. 8, pp. 1–22, Aug. 2017. [Online]. Available: https://www.frontiersin. org/articles/10.3389/fpsyg.2017.01454

[10] P. Resnick and H. R. Varian, ”Recommender systems,” Commun. ACM, vol. 40, no. 3, pp. 56–58, 1997.”.