# AI Emotion Detection Through Facial and Voice Recognition

Gabriella Torres-Santiago
Department of Computer Science
University of South Carolina Upstate
Spartanburg, South Carolina, USA
December 2, 2024

torressg@email.uscupstate.edu

## ABSTRACT

Emotion detection is increasingly important with the rise of user-centered AI and machine learning. Accurate emotion detection has applications in areas like mental health, customer service, and human-computer interaction. This research focuses on detecting emotions through the analysis of facial expressions and vocal tones, comparing three AI models: one using facial expression analysis with the MELD dataset, one using MobileNetV2 and a custom dataset, and one using vocal emotion recognition from RAVDESS, CREMA-D, TESS, and SAVEE datasets. A hybrid system integrating these models achieved 82-99% accuracy, significantly improving vocal emotion recognition. The results indicate that combining these models enhances emotion detection accuracy and reliability, providing a more comprehensive approach for real-world scenarios.

## Keywords

Machine Learning, ML Artificial Intelligence, AI, Emotion Detection, Facial Detection, Voice Detection

Within the broader domain of artificial intelligence, emotion detection represents a growing subfield. For this study's purposes, emotion detection refers to the AI-driven analysis of human emotions through the interpretation of facial expressions and vocal cues. Emotion detection is important because it provides a bridge between human emotions and machine learning systems, enabling AI to interact with humans in a more empathetic and intuitive manner. Emotion detection systems have been successfully applied in various industries, and their importance continues to grow as developers refine the accuracy and effectiveness of these technologies.

My contribution to this field focuses on developing an AI-based emotion detection system that combines facial expression analysis with vocal tone assessment to enhance accuracy and context understanding. By integrating these two modalities, this research aims to overcome some of the challenges associated with interpreting complex emotions, such as distinguishing between similar emotions like frustration and sadness. Additionally, the project seeks to explore the ethical considerations of emotion detection, ensuring that the system respects user privacy and maintains transparency in its data collection methods. This study's findings could provide valuable insights into creating more robust and ethically sound AI models that can be applied in areas like mental health support and personalized customer interactions.

## 2. LITERATURE REVIEW

AI emotion detection systems hold significant promise but also face several challenges. Human emotions are complex, and cultural differences in emotional expression make it difficult to create universally accurate models. Developers must consider these factors when designing AI systems, particularly in industries such as healthcare and customer service, where emotional accuracy is critical. Additionally, the ethics of emotion detection are still under scrutiny, as the ability of machines to interpret private emotional states raises questions about privacy and consent.

Building on the contributions of Dalvi et al. [1], Saxena and Khanna [3], Siirtola [4], Teye et al. [5], and Heyday et al. [2], this research compares two AI models for emotion detection, focusing on their methods and outcomes in analyzing human emotions. The first model utilizes MELD (Multimodal Emotion Lines Dataset) AI, which specializes in detecting emotions through facial expressions. MELD AI offers a rich set of features for analyzing visual cues, providing detailed insights into facial movements and expressions.

This aligns with the focus of Dalvi et al., who highlight the importance of feature selection and advanced algorithms in improving the accuracy of facial emotion recognition. MELD AI addresses these considerations by leveraging deep learning techniques to adapt to diverse facial expressions across different demographics. However, similar to challenges noted by Saxena and Khanna, MELD AI can struggle with accurately recognizing emotions in cases where visual data is insufficient, such as when facial expressions are subtle or partially obscured.

The second model employs RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) for emotion detection through voice recognition. This model is designed to interpret emotional states by analyzing vocal cues, such as tone, pitch, and rhythm. RAVDESS enables the detection of emotional nuances that are often missed in visual analysis alone, addressing the limitations of purely visual systems as discussed by Saxena and Khanna. The integration of audio data also resonates with Siirtola's emphasis on the importance of human-AI collaboration and incremental learning, as the RAVDESS model adapts to user feedback to refine its analysis of vocal tones. This adaptability makes it particularly suitable for contexts like telehealth and virtual communication, where audio plays a key role in understanding user emotions.

**Key Differences and Considerations:**

The MELD AI-based model focuses on extracting emotional data from facial expressions, making it particularly effective in face-to-face interactions where visual cues are predominant. It is well-suited for applications such as video-based mental health assessments and user feedback analysis, where recognizing subtle facial changes is crucial. This is aligned with the findings of Dalvi et al., who emphasize the importance of accurate feature extraction in facial emotion detection. However, as highlighted by Teye et al., MELD AI may struggle in environments where cultural differences influence how emotions are visually expressed, making cross-cultural accuracy a challenge.

In contrast, the RAVDESS-based model captures a wider range of emotional variations through vocal analysis, making it ideal for scenarios where auditory information is a primary indicator of emotion. Its ability to interpret nuances in vocal expression aligns with the emphasis on multimodal data integration discussed by Saxena and Khanna, enhancing the accuracy of emotion recognition in settings like telehealth or virtual customer service. However, similar to the challenges described by Dalvi et al., this model may have limitations when dealing with users whose speech patterns or dialects vary widely, potentially impacting its performance.

| Model | Dataset | Type | Training Time (Epochs) |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Facial | MELD | CNN | 30 |
| Facial | MobileNet V2 | Transfer | 30 |
| Vocal | RAVDES | 1D CNN | 50 |

**Figure 1. Summary of Models Used**

**Ethical Considerations and Contextual Awareness:**

Both models integrate ethical considerations outlined by Heyday et al. [2], focusing on transparency in data processing and ensuring user privacy during emotion detection. The MELD AI model is designed to respect user data by anonymizing facial inputs, while the RAVDESS-based model incorporates consent protocols for audio recordings, addressing privacy concerns in vocal data analysis. Teye et al. [5] emphasizes the need for AI systems to be culturally aware and sensitive to diverse backgrounds, a challenge that both models aim to address through region-specific training and adaptation mechanisms.

Moreover, the incremental learning approach suggested by Siirtola [4] is applied in both models to enhance their performance over time. The MELD AI model leverages user feedback to refine its facial analysis capabilities, while the RAVDESS model adapts to variations in vocal tone based on user interactions, ensuring that the models remain responsive to real-world complexities. This adaptability, coupled with an emphasis on ethical management, positions both models as promising candidates for practical deployment in various emotion-sensitive applications.

# 3. METHODOLOGY

This study investigates the performance of AI models for emotion detection through facial expression analysis and vocal tone analysis. The goal is to evaluate the effectiveness of these models individually and explore the potential of combining them to improve overall emotion detection accuracy. The methodology consists of the following steps:

## 3.1 Overview of the Study

The study aims to develop a robust emotion detection system by analyzing facial expressions and vocal tones.

Three models were implemented:

Model 1: Facial expression analysis using the MELD dataset.

Model 2: Facial expression analysis using a custom dataset and MobileNetV2.

Model 3: Vocal emotion recognition using RAVDESS, CREMA-D, TESS, and SAVEE datasets.

The integration of these models ensures a multimodal approach, improving emotion detection, accuracy and reliability.

## 3.2 Datasets and Evaluation Metrics

*3.2.1 Datasets*

The MELD dataset is a multimodal dataset with annotated facial expressions. It includes emotions such as happiness, sadness, anger, surprise, and fear. To prepare the dataset for analysis, preprocessing steps were undertaken, including resizing images to 224x224 pixels and normalizing pixel values. Additionally, augmentation techniques, such as flipping, rotation, and cropping, were applied to increase data diversity. The MobileNetV2 Transfer Learning Model was used with the dataset, which comprised diverse emotions including neutral, happy, sad, angry, disgust, fear, and surprise. The images were standardized to a size of 224x224 pixels, and augmentation techniques were again applied to promote variety within the training data.

For vocal emotion recognition, multiple datasets were utilized. The RAVDESS dataset provides emotional speech labeled with eight distinct emotions. CREMA-D includes labeled audio recordings for emotions such as happy, sad, and neutral, while TESS is primarily focused on surprise

and other emotions. SAVEE also contains audio samples reflecting expressive emotions. Preprocessing for these datasets involved standardizing audio samples in terms of sample rate. Features were then extracted using Librosa, focusing on Mel-frequency cepstral coefficients (MFCCs), Chroma STFT, Zero Crossing Rate, and Mel Spectrogram to capture a wide range of emotional characteristics.

*3.2.2 Evaluation Metrics*

Evaluation metrics were employed across all models to assess their performance comprehensively. Accuracy was used to measure the percentage of correct predictions, providing an overall sense of model effectiveness. Precision, recall, and F1-score were utilized to analyze performance at the class level, ensuring a detailed understanding of how well each emotion was classified. Additionally, confusion matrices were generated to visualize misclassifications and identify areas for improvement. Comparative analyses were conducted to evaluate the individual models' performance against the hybrid approach, highlighting the strengths and weaknesses of each.

## 3.3 Methods

*3.3.1 Facial Emotion Detection*

**MELD-Based Facial Expression Analysis:**

This model utilized a Convolutional Neural Network (CNN) architecture to extract features and classify emotions from the MELD dataset. The CNN consisted of convolutional layers for spatial feature extraction, dense layers with ReLU activation to capture complex relationships, and dropout layers to mitigate overfitting. The final output layer employed a softmax activation function to perform multiclass classification across the available emotion categories.

The training process involved using the Adam optimizer for efficient gradient updates and categorical cross entropy as the loss function to handle the multiclass nature of the problem. The MELD dataset was split into 70% training, 15% validation, and 15% testing subsets. Training was conducted over 30 epochs with early stopping to prevent overfitting and checkpointing to save the best-performing model. Evaluation metrics included accuracy to assess overall performance, F1-score to measure the balance between precision and recall, and a confusion matrix to analyze misclassifications across emotion categories.

**MobileNetV2-Based Facial Expression Analysis**

The second facial emotion detection model employed transfer learning with MobileNetV2, pre-trained on the ImageNet dataset. The architecture retained MobileNetV2's base layers for feature extraction while adding custom layers for emotion classification. These layers included dense layers with ReLU activation for non-linear transformations and a final softmax layer for classifying seven emotion categories: neutral, happy, sad, angry, disgust, fear, and surprise.

The model was fine-tuned using the Adam optimizer and sparse categorical cross entropy as the loss function. A batch size of 64 was used during training, which ran for 30 epochs. The dataset was preprocessed by resizing images to 224x224 pixels and applying normalization. Data augmentation techniques, such as flipping and rotation, were used to improve model robustness. The model underwent real-time testing to evaluate its ability to classify emotions accurately in dynamic environments, with metrics including accuracy, F1-score, and a confusion matrix for performance analysis.

*3.3.2 Vocal Emotion Recognition*

The vocal emotion recognition model was designed to classify emotions from audio recordings using CNN-based architecture. This model utilizes audio datasets from RAVDESS, CREMA-D, TESS, and SAVEE, each contributing a diverse range of emotional speech data. Preprocessing involved normalizing audio signals, extracting key features, and augmenting the data to improve model generalizability.

**Architecture**

The CNN-based architecture included multiple 1D convolutional layers to extract temporal features from the audio signals. These layers were followed by max-pooling layers to down sample features and reduce computational complexity. Fully connected dense layers were used to classify emotions, with dropout layers incorporated to prevent overfitting. The final output layer used softmax activation to classify audio samples into eight emotion categories: neutral, calm, happy, sad, angry, fear, disgust, and surprise.

**Training**

Training was conducted using the Adam optimizer, chosen for its adaptability in learning rate adjustments, and categorical cross entropy as the loss function. The dataset was split into 70% training, 15% validation, and 15% testing subsets. The model was trained over 50 epochs with a batch size of 64, and a ReduceLROnPlateau callback was implemented to adjust the learning rate dynamically if the loss plateaued. Data augmentation techniques such as noise addition, pitch shifting, time stretching, and time shifting were applied to improve model robustness.

**Evaluation**

To comprehensively evaluate the hybrid model's accuracy, the code was designed to make 25 predictions per video. The variety of videos used also contributed to understanding how the model interpreted different emotional expressions, allowing us to see how well it generalized across diverse scenarios. To further validate these results, the model was run three times for each video, resulting in a total of 75 predictions per video. The final accuracy was calculated based on the proportion of correct predictions out of 75. For example, if the model correctly predicted the emotion 70 times out of 75, the resulting accuracy for that video would be 93.33%. This repeated prediction approach provided a more stable estimate of the model's performance by averaging the outcomes across multiple runs. It also allowed for an assessment of the model's consistency in its predictions for the same video, rather than just producing random guesses with meaningless results. Additionally, analyzing the incorrect predictions provided insight into how the same emotional display could be interpreted differently, which highlighted how context and variation between videos could affect the model's accuracy, helping to identify potential areas for future improvement in the model's interpretation of nuanced emotions.

## 3.4 Implementation

*3.4.1 Model Integration*

The three models—two for facial emotion detection and one for vocal emotion recognition—were integrated into a unified system using a weighted average fusion methodology. The outputs of the individual models were combined by assigning weights based on their respective validation performances. This approach allowed the system to prioritize the most reliable predictions for each modality, leveraging the strengths of each model to improve overall accuracy and robustness.

The integrated model was evaluated on a reserved test dataset, distinct from those used for training and validation, to ensure unbiased performance analysis. A comparative analysis was conducted between the hybrid model and the individual models, highlighting the advantages of multimodal integration in terms of accuracy, precision, recall, and F1-score. This analysis demonstrated the hybrid system's ability to mitigate the limitations of single-modal approaches and improve overall emotion detection reliability.

*3.4.2 Deployment*

This real-time processing capability was implemented using Visual Studio Code, where the system was deployed to analyze pre-recorded videos stored in a designated folder. The code reads each video from the folder, processes the visual and audio

cues, and predicts the emotions displayed. This approach allows for efficient emotion detection from a batch of video inputs, simulating dynamic, real-world scenarios.

The deployment effectively demonstrated the system's capability to interpret emotions in varied settings, where having immediate feedback on emotional states is essential. Practical applications of the system include mental health monitoring, where emotion detection can aid in recognizing signs of distress or mood disorders in patients based on facial and vocal cues, as well as emotion-aware virtual assistants, which can adapt their interactions according to the user's emotional state to provide a more personalized and supportive experience.

## 3.5 Ethical Considerations

Ethical principles were prioritized throughout the project to ensure responsible and fair use of technology. All datasets were anonymized to protect personal information, and diversity in the datasets was emphasized to minimize biases and improve generalizability across different demographics. Additionally, the system's design ensures transparency in data usage and prediction processes, fostering trust and accountability.

# 4. IMPLEMENTATION

The implementation of this research focuses on two primary AI models for emotion detection: one utilizing MELD AI for analyzing facial expressions and the other employing RAVDESS for assessing vocal tones. The Python programming language is used in conjunction with deep learning libraries such as TensorFlow and Keras, with the process divided into three main stages: model development, data preprocessing, and model integration.

The first facial emotion detection model, based on the MELD dataset, utilized CNN architecture for feature extraction and emotion classification. Training was conducted using the Adam optimizer, with categorical cross entropy as the loss function. The model was trained for 30 epochs, with early stopping to prevent overfitting and checkpointing to save the best-performing model. The dataset was split into 70% training, 15% validation, and 15% testing subsets.

The MobileNetV2 Transfer Learning Model employed transfer learning with MobileNetV2, pre-trained on the ImageNet dataset. This model retained MobileNetV2's base layers for feature extraction and added custom dense layers for classifying seven emotions. Sparse categorical cross entropy was used as the loss function, and the model was trained with a batch size of 64 over 30 epochs. Data augmentation further improved its ability to generalize across diverse scenarios.

The vocal emotion recognition model used a 1D convolutional neural network (CNN) to analyze temporal features from audio signals. Its architecture included convolutional layers for feature extraction, max-pooling layers for dimensionality reduction, and fully connected layers for classification. Training was performed using the Adam optimizer and categorical cross entropy as the loss function, with 50 epochs and a ReduceLROnPlateau callback to adjust the learning rate dynamically when performance plateaued.

Each model was evaluated on reserved test datasets using metrics such as accuracy, F1-score, and confusion matrices. These metrics provided a comprehensive understanding of the models' performance and areas for improvement.

The three models were integrated into a hybrid system using a weighted average fusion methodology. Each model's output probabilities were weighted based on its validation performance, allowing the integrated system to prioritize the most reliable predictions for specific modalities. The hybrid model was tested on a reserved test dataset, demonstrating significant improvements in accuracy and robustness compared to the individual models. Comparative analysis confirmed that integrating facial and vocal modalities enhanced the system's ability to detect emotions accurately and reliably.

The final system was deployed as a real-time application capable of processing both image and audio inputs to provide emotion predictions. A graphical user interface (GUI), developed using Tkinter, enabled live visualization of results. The GUI displayed predicted emotions along with their confidence scores, making the system interactive and user-friendly.

Real-time testing was conducted using live webcam feeds for facial emotion detection and real-time audio streams for vocal emotion recognition. The system demonstrated its applicability in scenarios requiring immediate feedback, such as mental health monitoring and emotion-aware virtual assistants. These applications showcased the potential of the integrated system in practical settings, where accurate and timely emotion detection is critical.

The integrated system outperformed individual models in terms of accuracy and reliability, validating the hypothesis that multimodal integration enhances emotion detection. Key findings included improved generalization across diverse datasets and dynamic scenarios, effective classification of complex emotions in real-time, and robust performance under varied conditions such as lighting and background noise**.**

# 5. EXPERIMENTAL SETUP

For this study, the experimental setup involved the use of Google Colab for testing code and Visual Studio Code (VS Code) for code editing. The integration of these tools facilitated both development and model training in a streamlined workflow. Google Colab provided a cloud-based environment with GPU support, essential for handling the computationally intensive training of deep learning models. VS Code served as a versatile platform for managing and editing the project's codebase, making it easier to maintain organized code and track changes throughout the development process.

The study used two primary datasets: one for facial images and another for audio recordings. Facial images were sourced from the Kaggle database, which included diverse examples of various emotions like happiness, sadness, anger, surprise, and fear, all converted into greyscale for consistent input to the AI models. In addition to this, a personal dataset of greyscale images was created, comprising self-images to ensure the model's adaptability to real-world variations in expressions. Audio recordings were similarly sourced from a Kaggle database that provided labeled emotional speech data. This dataset included samples in different accents and languages, providing a broad range of emotional expressions through vocal tones. To enhance this dataset's relevance, additional audio samples of personal speech were included, which served as a way to assess the model's performance on real-world audio variations.

The images were preprocessed by resizing them to a uniform dimension and applying normalization. Data augmentation techniques like rotation, flipping, and cropping were used to increase the diversity of training samples, which improved the model's generalization ability. For audio data, preprocessing involved segmenting audio recordings into smaller time frames to capture variations in vocal tone. Mel-frequency cepstral coefficients (MFCCs) were extracted from these segments as features for the models, enabling the audio data to be converted into a format suitable for input into neural networks.

The experimental phase involved training two separate AI models: one for analyzing facial expressions and another for assessing vocal tones. The facial analysis model utilized Convolutional Neural Networks (CNNs) to detect emotions based on facial features. This model was trained using both the Kaggle images and personal images, ensuring that it could generalize across different faces and lighting conditions. Meanwhile, the audio analysis model employed a Recurrent Neural Network (RNN) to process sequences of MFCCs extracted from audio recordings, identifying the speaker's emotional state.

Training and validation data were split in a 70-15-15 ratio, where 70% of the data was used for training, 15% for validation, and the remaining 15%

for testing. Model performance was evaluated through metrics like accuracy, precision, recall, and F1-score. These metrics helped in assessing the ability of each model to accurately classify emotions and provided insight into areas where improvements were needed. After evaluating the individual models, their outputs were combined using a weighted average fusion method, where the weights were based on each model's performance for particular emotions. This hybrid approach aimed to leverage the strengths of both models, providing a more accurate and reliable emotion detection system.

All model training processes were executed in Google Colab using TensorFlow and PyTorch libraries, while VS Code was used for refining scripts and implementing hyperparameter tuning strategies. This setup allowed for iterative testing and debugging, ensuring that the models were optimized for real-world applications like virtual therapy sessions and customer service interactions. The integration of personal datasets with the publicly available ones enabled the models to better adapt to diverse expressions and vocal variations, thereby improving their performance on unseen data.

Overall, this experimental setup provided a robust environment for developing and testing AI models for emotion detection, combining the ease of cloud-based training with the flexibility of local code management. The combination of facial and vocal analysis, along with real-world testing, offered valuable insights into the effectiveness of multi-modal approaches for accurately detecting human emotions.

# 6. RESULT ANALYSIS

After thorough evaluation of the available models, the MELD-based facial analysis model and the vocal emotion recognition model emerged as the most promising candidates for integration into the hybrid system. This decision was influenced by their respective architectures, capabilities, and complementary nature, making them suitable for a combined approach that leverages both facial and vocal data. An interesting observation was that the MobileNetV2 model and the RAVDESS model did not mesh very well together. Alot of changes and updates were needed to make it so the models worked alongside each other, while also keeping important aspects of the MELD.

The individual facial recognition model consistently achieved 100% accuracy when used in isolation, whether the face was exaggerated or more neutral. This highlighted the model's effectiveness in capturing clear visual emotional cues for various situations, even being able to accurately guess the emotion from three-quarter profile of some faces. On the other hand, the vocal emotion recognition model performed significantly worse, with an accuracy of only 16%. This demonstrated the difficulty of interpreting emotional states through vocal intonations alone, which can be influenced by factors like individual speaking styles, background noise, and cultural differences. These challenges made it difficult for the vocal model to accurately detect nuanced or subtle emotions.

The hybrid model, which integrated components from both the MELD-based facial analysis model and the vocal emotion recognition model, achieved an overall accuracy of between 82% and 99%. The integration provided significant improvements to the vocal model's accuracy, particularly by adding reliability when analyzing videos that contained both visual and auditory cues. The hybrid approach enhanced the ability to capture emotional states more accurately in videos, especially when vocal information alone was insufficient. By combining vocal cues with visual data, the hybrid model was better able to infer emotional states in dynamic environments where audio alone would have been too ambiguous.

| Video | Run 1 | Run 2 | Run 3 | Average |
|---|---|---|---|---|
| **1** | 24/25 | 25/25 | 25/25 | 74/75<br>98.667% |
| **2** | 23/25 | 24/25 | 23/25 | 70/75<br>93.333% |

| | | | | |
|---|---|---|---|---|
| **3** | 20/25 | 21/25 | 21/25 | 62/75<br>82.667% |

**Figure 2. Results of the Hybrid Model**

However, it is important to note that while the hybrid model significantly improved the accuracy of the vocal emotion recognition aspect, it inadvertently reduced the facial recognition accuracy. The facial analysis model, which achieved perfect accuracy on its own, saw a reduction in performance within the hybrid system due to the influence of vocal data, which could occasionally introduce noise or ambiguity. This trade-off highlights an important observation: while the hybrid approach provided a more balanced and nuanced understanding of emotions in videos, it compromised the precision of facial analysis that was otherwise highly effective when used alone. Surprisingly, the speech model did not bring down the facial models' accuracy too low, as the lowest accuracy displayed was 82.667%.

The hybrid model tended to perform well in recognizing clearly expressed emotions but struggled with more subtle or ambiguous emotional cues. This outcome suggests that, while integrating both modalities was beneficial for adding depth and context to the vocal analysis, it also introduced variability that impacted the consistency of facial emotion detection. The hybrid approach is therefore more suitable for scenarios where both audio and video are present, allowing the system to compensate for weaker vocal cues with strong visual signals and vice versa.

The repeated prediction mechanism, in which the model made 25 predictions per video, further enhanced the reliability of the emotion detection process. By running the model three times for each video, resulting in a total of 75 predictions, the system provided a more stable estimate of performance, reducing the impact of any individual frame or audio segment that might not accurately represent the overall emotional state. This approach also helped to assess the consistency of the model's predictions across multiple runs, providing additional insight into its reliability and areas where further refinement is needed.


**Figure 3. Sample of Speech Model Results**


**Figure 4. Sample of Face Model Results**

The analysis of incorrect predictions also revealed interesting findings. It became evident that the hybrid model's reliance on both facial and vocal data could lead to misinterpretations, particularly when one modality conflicted with the other. For example, if the vocal tone suggested a different emotion than the facial expression, the hybrid model sometimes struggled to resolve this ambiguity accurately. Additionally, the variety of videos used in the evaluation contributed to understanding how well the model generalized across diverse scenarios, revealing that context and environmental differences played a crucial role in how emotions were interpreted.

Despite its strengths, the hybrid model's reliance on exaggerated facial expressions and the limitations in vocal analysis indicate areas for improvement. Future iterations should focus on enhancing the model's sensitivity to subtle emotional cues by incorporating more advanced feature extraction techniques, adding contextual factors such as body language, and refining the temporal analysis to better understand the progression of emotions over

time. This could help the model become more adaptable to real-world scenarios, where individuals may not always display overt or intense emotions, or where they would even exhibit complex emotions like sarcasm or dramatic overstatement.

# 7. CONCLUSION

This study highlights the challenges of using AI for emotion detection, emphasizing the complexity of human emotions influenced by cultural, contextual, and individual factors. The facial detection model demonstrated high accuracy for exaggerated emotions, whereas the vocal model achieved only 16% accuracy, highlighting the challenges of interpreting vocal cues effectively.

Integrating facial and vocal emotion detection into a hybrid model represents a significant advancement in the field of AI-based emotion detection. Each modality—facial expression analysis and vocal emotion recognition—has its own strengths, but also notable weaknesses. By combining both modalities, the hybrid model leverages the complementary advantages of visual and auditory data, resulting in a more holistic understanding of emotional states, which is critical for capturing the full complexity of human emotions. This integrated approach addresses some of the key challenges in interpreting emotions accurately, especially in situations where one modality alone is insufficient or ambiguous. Combining facial and vocal data into a hybrid model showed promise, achieving an accuracy of 82% to 99%. This integration helped mitigate some limitations of individual models, providing a more comprehensive understanding of emotional states. However, the hybrid model struggled to detect more subtle and neutral emotions, indicating the need for improvements in integrating modalities for nuanced emotional detection. Future work should focus on refining sensitivity to subtle cues, incorporating contextual factors like body language, and exploring advanced techniques such as attention mechanisms to improve adaptability and accuracy in real-world scenarios.

The significance of this research lies in its potential to enhance the accuracy and reliability of emotion detection systems across real-world applications. The ability to combine multiple data types allows for richer emotional context, which is particularly beneficial in settings like mental health monitoring, customer service, and human-computer interaction—areas where understanding emotions plays a crucial role in delivering personalized and effective experiences. By improving the robustness of emotion detection through multimodal integration, this research contributes to the development of empathetic, adaptable, and user-centered AI systems that can interact more naturally and ethically with humans.

Cultural differences and ethical considerations are crucial in developing emotion detection systems. Variations in emotional expression across cultures can lead to inaccuracies, making diverse and region-specific training data essential. Ethical considerations, including privacy, transparency, and user consent, are vital to ensure the responsible use of these technologies and foster user trust.

Overall, integrating facial and vocal data into a hybrid model offers significant potential for enhancing AI-based emotion detection. Future research should aim to make these systems more reliable, culturally adaptable, and ethically sound. By addressing these challenges, emotion detection technologies can become more effective, ultimately improving applications in areas like healthcare, customer service, and human-computer interaction, where understanding emotions is crucial.

# 8. REFERENCES


[1] Dalvi, Chirag, et al. “A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions | IEEE Journals & Magazine | IEEE Xplore.” IEEE Xplore, IEEE, 30 Nov. 2021

[2] declare-lab. “GitHub - Declare-Lab/MELD: MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation.” GitHub, GitHub, 10 Oct. 2020

[3] Heyday, Teresa, et al. “Ethical Management of Human-AI Interaction: Theory Development Review - ScienceDirect.” ScienceDirect.Com | Science, Health and Medical Journals, Full Text Articles and Books., ScienceDirect, 2023

[4] Luna-Jiménez, Cristina, et al. “Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning.” MDPI, Multidisciplinary Digital Publishing Institute, 18 Nov. 2022

[5] Puri, Tanvi, et al. “Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network.” Wiley Online Library, Intelligent Solutions in E-Health and Medical Communication Services, 27 Feb. 2022

[6] Sandler, Mark, et al. “[1801.04381v4] MobileNetV2: Inverted Residuals and Linear Bottlenecks.” ArXiv.Org, Cornell University, 21 Mar. 2019

[7] Saxena, Anvita, and Ashish Khanna. “Emotion Recognition and Detection Methods: A Comprehensive Survey | Institute of Electronics and Computer.” Home | Institute of Electronics and Computer, Institute of Electronics and Computer, 7 Feb. 2020

[8] Siirtola, Pekka. “Sensors | Free Full-Text | Incremental Learning to Personalize Human Activity Recognition Models: The Importance of Human AI Collaboration.” MDPI, Multidisciplinary Digital Publishing Institute, 25 Nov. 2019

[9] Teye, Martha, et al. “Evaluation of Conversational Agents: Understanding Culture, Context and Environment in Emotion Detection | IEEE Journals & Magazine | IEEE Xplore.” IEEE Xplore, IEEE, 22 Feb. 2022