Abstract

Emotion classification plays a critical role in advancing human-computer interaction by enabling systems to interpret human emotional cues accurately. Traditional models often rely on a single modality, either audio or visual, limiting their ability to discern subtle variations, such as posed versus genuine emotions. This unimodal approach results in lower accuracy for complex applications that demand nuanced emotional detection. To overcome these limitations, we developed a multimodal emotion classification model using CNN architectures to analyze both audio and visual data. Our approach distinguishes 14 emotions, encompassing seven real and seven posed categories. Separate CNN models were implemented for each modality: the audio model extracts features like Mel-spectrogram and MFCC, while the visual model processes 48x48 grayscale facial images through analysis of face, eye frame, and eyebrow expressions. By integrating these distinct feature sets, the model improves its ability to capture cross-modal emotional cues and distinguish subtle differences. Evaluation across multiple datasets demonstrated promising results, with our visual model achieving a training accuracy of 71.00% and the audio model reaching 75.41%, highlighting the value of multimodal inputs in enhancing emotion classification performance.

Keywords: Audio and Visual, Facial Recognition, Fake Emotion, Multimodal, Posed Emotion

1 Introduction

Emotion classification is a rapidly advancing field that seeks to decode human emotional expressions, both verbal and nonverbal, by leveraging statistical and machine learning techniques. With applications ranging from human-computer interaction to mental health diagnostics, accurately identifying emotions can significantly enhance the effectiveness of automated systems. The field has seen a notable shift from traditional statistical models, which often relied on feature extraction and simpler classifiers, to deep learning models that can automatically learn intricate features. Convolutional Neural Networks (CNNs) are especially well-suited for this task, given their ability to capture patterns in image and audio data. However, a major challenge remains in achieving robust generalization across different emotional expressions, especially when considering posed versus real emotions. Previous works in emotion classification have demonstrated impressive results with uni-modal models that rely solely on either audio or visual data. However, these models often struggle to capture the full context of an emotion, particularly when handling complex emotions or distinguishing between posed and genuine expressions. While some multi-modal approaches have been introduced, they frequently lack the granularity needed to

identify these nuanced distinctions, leading to reduced accuracy in real-world applications. Moreover, many studies have not incorporated datasets that include a wide range of both posed and real emotions, limiting the generalizability of their findings. This limitation highlights the need for a comprehensive, multi-modal approach that can differentiate between subtle emotional cues while maintaining robustness across diverse datasets. To address these challenges, a multi-modal emotion classification model is developed that incorporates both audio and visual data to distinguish between 14 distinct emotions, including both posed and genuine expressions. By processing audio data through Mel-spectrograms and MFCC features and visual data through facial expression images, this approach combines features across modalities to better capture the complexity of emotional cues. This design allows for the identification of emotions from diverse datasets, including those with posed expressions, which are critical for applications in fields like mental health and customer service. The integration of separate CNN architectures for each modality enhances the model's ability to accurately distinguish between real and posed emotions, providing a more nuanced understanding of each classification.

The structure of the model is designed to process audio and visual inputs separately, utilizing dedicated CNN architectures optimized for each modality. The audio data is transformed into Mel-spectrograms and MFCCs, which are fed into a CNN designed to capture time-frequency features. Meanwhile, the visual CNN processes 48x48 grayscale facial images, extracting spatial features indicative of different expressions. By combining outputs from these two parallel CNNs, the model performs final classification on the concatenated feature space, enabling it to jointly leverage audio and visual cues for enhanced emotional recognition. This multimodal architecture not only allows for greater accuracy but also supports the detection of subtle differences between real and posed emotions, addressing a key gap in previous emotion classification models.

2 Related Work

Recent studies have made significant strides in distinguishing between genuine and posed emotions, particularly through the analysis of facial expressions in videos. For instance, one study focused on determining the genuineness of facial expressions by classifying them as sincere or fake[1]. This research employed various convolutional and recurrent deep neural networks, incorporating facial landmarks and their trajectories. The best-performing model achieved an accuracy of 73% using a Recurrent Delta (RD) network with LM5 attributes related to facial landmarks. Another study aimed to classify authentic and fake anger expressions by implementing and comparing three methods: Long ShortTerm Memory (LSTM) networks, fuzzy logic classification, and fully connected neural networks [11]. The LSTM network outperformed the other approaches, achieving an accuracy of 80% on the testing dataset. Additionally, research investigating the physiological signals of observers provided insights into differentiating between genuine and fake smiles [12]. By recording physiological signals such as ECG, GSR, BVP, and eye activity from observers while watching smile videos, the study achieved high classification accuracy, surpassing human self-reports.

The best performance recorded was 98.8% using features derived from the observers' pupillary response (PR). Explorations into laughter and smile detection under various emotional conditions (e.g., genuine amusement versus negative emotions) have also been conducted. One such study applied transfer learning to a pre-trained laughter detection model, significantly improving accuracy to 83% on the held-out test set, compared to only 61% with the pretrained model alone [14]. Further investigations focused on recognizing both felt and unfelt facial expressions of emotion, with particular emphasis on comparing genuine and unfelt expressions [15]. This research employed a fine-tuned VGGFace deep network for facial expression recognition and an EMNet for static representations, achieving state-of-the-art results across various datasets, including 98.7% on CK+, 89.6% on OuluCASIA, and 70.2% on SASE-FE. A study examining the classification of real and fake emotions utilized various machine learning algorithms, including an Enhanced Boosted SVM (EBSVM) alongside deep learning techniques [1]. The EBSVM achieved the highest accuracy of 98.08% for classifying real and fake emotions . A study centers on reviewing existing datasets for Facial Emotion Recognition (FER) and the classification of real and fake expressions, with a specific emphasis on analyzing methodologies and techniques utilized in the literature [2]. The authors evaluated a wide range of techniques, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Deep Neural Networks (DNN). Findings revealed that several studies reported cumulative errors and suboptimal accuracies, underscoring the need for improved approaches in real versus fake emotion classification. Another paper presents a systematic review of various Spontaneous vs. Posed (SVP) facial expression detection methods [3]. The authors classified these methods into three main categories: temporal muscle activation-based methods, spatial pattern-based methods, and hybrid approaches. Results indicate that temporal muscle activation-based techniques, which measure activation in specific facial muscles, achieve high accuracy, occasionally surpassing 90%. Lastly, a study aims to provide insights into the automatic detection of genuineness in emotional facial expressions [4], employing interpretable machine learning models. The authors utilized small decision trees and a linear Support Vector Machine (SVM) model with LASSO regularization to enhance interpretability. Results show that the Decision Tree model achieved an average accuracy of 77%, while the SVM model attained an average accuracy of 72%.

2.1 Summary of Research Gaps

- **Dataset Limitations**: Current datasets often suffer from being small and unbalanced, especially for emotions beyond happiness and sadness. This lack of diversity can lead to overfitting and reduce model generalizability.
  - The scarcity of audiovisual datasets with real and fake emotions limits multimodal approaches.
  - Many datasets lack dynamic sequences or spontaneous expressions, affecting real-world applicability.
- **Subjectivity and Contextual Influence**: Assessing the genuineness of emotions is inherently subjective, with cultural, personality, and mood factors influencing both expression and perception.

  - Most studies focus on isolated facial expressions, missing broader interactional contexts, impacting model robustness.
- **Methodological Constraints**: Studies often rely on specific muscle movements or a narrow set of emotions, limiting comprehensive understanding.
  - Self-reported data can introduce social desirability or recall bias.
  - Reliance on complex, uninterpretable models restricts insight into key classification features, affecting practical application.
- **Technical Challenges**: Subtle facial movements and physiological signals are challenging to capture accurately due to issues like low resolution, occlusion, lighting, and head pose variations.
  - There is a need for robust, user-friendly hardware for physiological data acquisition.
  - High computational resources are required for high-resolution video data, presenting additional obstacles.
- **Interpretability vs. Accuracy Trade-Off**: While high accuracy is often prioritized, interpretability remains essential for practical use and understanding the mechanisms behind genuine emotion detection.
  - Current models frequently favor accuracy at the expense of transparency, limiting practical deployment.
- **Ethical Concerns**: The potential misuse of genuineness detection technology raises ethical considerations around privacy, consent, and algorithmic bias.
  - Ensuring responsible development and application of these technologies is essential to prevent privacy infringements and biased detection results.
- **Future Directions for Improvement**: Addressing these challenges requires larger, more diverse datasets, greater inclusion of contextual and cultural factors, and a balance between accuracy and interpretability.
  - Advancements in technology and ethical guidelines are essential for translating research into meaningful, real-world applications.

2.2 Contribution of Research

- This research addresses the unique challenge of detecting genuine versus posed emotions through a **multi-modal approach**, integrating both audio and visual data—an area with limited integration in existing models.
- Utilizes a comprehensive set of **diverse audiovisual datasets** that include both real and fake emotions, ensuring model robustness across different emotional expressions.
- Develops a robust **machine learning model based on Convolutional Neural Networks (CNNs)** to classify emotions across 14 distinct classes, covering both genuine and posed expressions.
- Achieves high accuracy across both audio and visual modalities, showing **superior performance in distinguishing authentic from posed emotions** compared to traditional single-modality models.
- Implements detailed **data preprocessing, feature extraction, and rigorous training and validation**, enhancing the model's robustness and real-world applicability.

- Visual analysis is focused on the **face, eyeframe, and eyebrow regions**, while audio analysis incorporates **tonal features** to enrich emotion classification.

3 Datasets

3.1 Visual Emotion Datasets

Visual emotion datasets are essential for developing emotion recognition systems, providing foundational data for model training and evaluation. They fall into two main categories: posed and real emotion datasets, each with distinct applications in affective computing.

- **Posed Emotion Datasets:**
  - **Extended Cohn-Kanade Dataset (CK+):** Contains images of participants transitioning from neutral to peak emotional states, which enhances model training through consistent, exaggerated expressions.
  - **Japanese Female Facial Expression Dataset (JAFFE):** Provides posed expressions from Japanese female subjects, allowing for insights into cultural differences in emotional expression.
  - **Fake Facial Expression Recognition (FFER) Dataset:** Captures simulated emotions, useful for applications such as deception detection by distinguishing between genuine and posed expressions.
- **Real Emotion Datasets:**
  - **Facial Expression Recognition (FER) Dataset:** Captures spontaneous emotional expressions in diverse settings, reflecting everyday emotional interactions and enhancing model robustness.
  - **Real-world Affective Faces Database (RAF-DB):** Sources expressions from various public images, challenging models with real-world subtleties and diversity in emotional expressions.

3.2 Audio Emotion Datasets

Audio emotion datasets are critical for recognizing emotional expressions conveyed through speech, paralleling visual datasets with their categorization into posed and real emotion datasets.

- **Posed Emotion Datasets:**
  - **Toronto Emotional Speech Set (TESS):** Contains recordings of actors articulating sentences with specific emotional tones, helping models recognize explicit emotions.
  - **Persian Emotional Speech Dataset (ShEMO):** Offers simulated emotional expressions in Persian, highlighting the importance of posed datasets in culturally diverse contexts.
- **Real Emotion Datasets:**
  - **Vivae Dataset:** Captures genuine emotional reactions, enriching model training with authentic expressions from natural interactions.

– **Multimodal Affective Dataset (MAV):** Includes non-speech vocalizations, such as laughter and crying, allowing models to analyze emotions conveyed without words.
– **EmoDB:** The Berlin Emotional Database (EmoDB) is a German-language audio dataset featuring recordings of seven simulated emotions by professional actors, designed for emotion recognition research. It is widely used in speech and emotion recognition studies due to its high quality and phonetically rich content.

By integrating posed and real emotion datasets, models can be trained for accurate emotion recognition across various domains, such as mental health, customer service, and human-computer interaction. This dual approach improves both foundational understanding and real-life applicability, enhancing the accuracy and utility of emotion recognition systems.

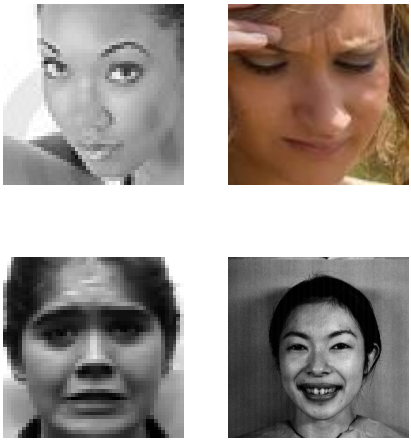

Fig. 1 Example images from different emotion datasets used in emotion recognition: (left top) FER neutral expression, (right top) RAF-DB sadness expression, (left bottom) CK+ posed fear expression, and (right bottom) JAFFE posed happy expression.

Table 1 Comparison of Emotion Counts in Different Datasets

Dataset	Data Type	Emotion Type	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
ShEMO	Audio	Posed	604	744	0	724	0	449	105
TESS	Audio	Posed	400	400	400	400	400	400	0
MAV	Audio	Genuine	20	10	10	10	20	10	10
EmoDB	Audio	Genuine	208	79	46	69	71	62	0
Vivae	Audio	Genuine	359	0	0	176	363	0	187
CK+	Image	Posed	135	54	177	75	207	84	249
JAFFE	Image	Posed	30	30	29	32	31	31	30
FER	Image	Genuine	0	6198	547	4281	8989	4002	0
RAF-DB	Image	Genuine	867	0	3204	355	5957	2460	1619

Table 2 Cumulative Comparison of Emotion Counts in Different Datasets

Data Type	Emotion Type	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Audio	Posed	1004	1144	400	724	400	849	105
Audio	Genuine	587	89	56	255	454	72	197
Image	Posed	165	84	206	107	238	115	279
Image	Genuine	867	6198	3751	4636	14946	6462	1619

4 Research Questions

This research investigates the efficacy of multimodal approaches in distinguishing between genuine and posed emotions by leveraging both visual and audio datasets. The central question is:

- How can integrating diverse audiovisual datasets, encompassing both real and simulated emotional expressions, enhance the accuracy and robustness of emotion recognition systems?
- This inquiry seeks to address the limitations of traditional unimodal models and explore the potential of combining features from various datasets to improve the classification of 14 distinct emotional expressions, including subtle differences between authentic and posed emotions in real-world applications.

5 Proposed Approach

In the multimodal emotion classification project, the primary goal is to identify and classify emotions based on multiple inputs, such as audio and visual data, including tonal analysis for audio and visual analysis of face, eyeframe, and eyebrow regions. This involves several key steps, each contributing to the overall effectiveness of the models employed in the classification process.

5.1 Data Collection

The first step involves gathering diverse datasets relevant to the emotions being classified. These include:

- **Posed Emotions:** Datasets where participants act out specific emotions.
- **Real Emotions:** Datasets capturing spontaneous emotional expressions.
- **Audio Datasets:** Contain audio recordings in `.wav` files representing various emotional states, analyzed through tonal analysis.
- **Visual Datasets:** Contain images in `.jpg` files, focusing on key facial areas like face, eyeframe, and eyebrow.

The names of the datasets are mentioned above.

5.2 Data Preprocessing

Once the datasets are collected, they undergo preprocessing to ensure uniformity and compatibility for model training. This stage typically includes several tasks:

- **Normalization:** Audio signals are normalized to a specific amplitude range, and images are scaled to a standard pixel value range (e.g., 0 to 1), stabilizing the training process.
- **Resizing:** Images are resized to a fixed dimension (e.g., 48x48 pixels) to ensure consistency across the dataset, especially for input into CNN models.
- **Augmentation:** Data augmentation techniques such as flipping, rotation, and scaling are applied to images and audio clips, increasing dataset diversity and improving model robustness.
- **Splitting:** The data is divided into training, validation, and test sets. The training set is for model training, the validation set is for hyperparameter tuning, and the test set is for evaluating model performance.

5.3 Feature Extraction

Feature extraction is an essential step, especially for audio and visual data. The key steps are:

- **Audio Feature Extraction:**
  - Raw audio signals are transformed into suitable formats for model training.
  - Techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram generation are used to capture tonal characteristics of audio signals.
- **Visual Feature Extraction:**
  - Features are derived directly from images using convolutional layers in a CNN.
  - CNNs capture spatial relationships and context, particularly in regions like the face, eye frame, and eyebrow.

5.4 Model Building

The next step involves constructing the models that will be trained on the prepared data. For multimodal emotion classification, two models are typically used:

Convolutional Neural Network (CNN) for Visual Data: In the multimodal emotion classification project, the visual component utilizes a Convolutional Neural Network (CNN) designed specifically for analyzing facial expressions by incorporating three key regions: the whole face, the eye region, and the eyebrow region. The model architecture is structured as follows:

1. **Face Input:** A dedicated input layer for the entire face image, which undergoes several convolutional and pooling operations to extract high-level features.
2. **Eye Input:** A separate input layer focusing on the eye region, where similar convolutional layers capture essential features specific to the eyes, providing crucial cues for emotion recognition.
3. **Eyebrow Input:** Another input layer for the eyebrow region, designed to highlight the expressive movements of the eyebrows, which play a significant role in conveying emotions.

The features extracted from these three branches are concatenated to form a comprehensive representation of the facial features. This combined representation is

then passed through dense layers for classification into 14 distinct emotion classes, employing techniques such as dropout for regularization. The model is compiled with the Adam optimizer and categorical cross-entropy loss function, enabling effective training for accurate emotion detection based on visual cues from the face, eyes, and eyebrows.

Convolutional Neural Network (CNN) for Audio Data: The audio CNN for emotion classification includes:

1. **Convolutional Layers:** Extract features from audio spectrograms to identify patterns.
2. **Max Pooling Layers:** Reduce dimensionality, retaining important features and preventing overfitting.
3. **Flattening Layer:** Converts the 2D output to a 1D array for dense layer processing.
4. **Dense Layers:** A ReLU activation layer learns relationships, followed by dropout for regularization.
5. **Output Layer:** A softmax layer classifies audio into 14 emotion classes.

This structure focuses on effective feature extraction for improved emotion recognition.

5.5 Model Compilation

Before training, each model is compiled with specific configurations. This includes selecting an optimizer, such as Adam, which is commonly used due to its adaptive learning rate capabilities. The loss function, typically categorical cross-entropy for multi-class classification tasks, measures how well the model's predictions match the true labels. Metrics like accuracy are also defined to evaluate model performance during training.

5.6 Training the Models

During training, both the audio and visual models undergo the following steps:

- **Data Feeding:** The preprocessed data and their corresponding labels are fed into the models.
- **Learning Process:** The models learn to map input features to emotional categories through iterative updates of their internal parameters.
- **Performance Monitoring:** The models' performance is monitored on a validation set to ensure generalization to unseen data.

5.7 Evaluation

After training, the models are evaluated using a separate test set to measure their effectiveness. Metrics such as accuracy, precision, recall, and F1 score are used to assess how well the models perform in classifying both posed and real emotions.

5.8 Deployment

Once the models are trained and evaluated, the following steps are taken for deployment:

- **Real-World Applications:** The models are deployed in emotion recognition systems for interactive platforms or sentiment analysis tools.
- **Real-Time Emotion Detection:** The system enables real-time emotion detection based on both audio and visual inputs.
- **Enhanced User Experience:** The deployment improves user experience by enabling emotion-aware interactions across various domains.

In summary, the process of multimodal emotion classification involves a systematic approach, from data collection to model deployment, ensuring effective and accurate emotion recognition. The focus is on tonal analysis for audio and region-specific visual analysis of the face, eyeframe, and eyebrow.

6 Result Analysis and Discussions

Data Handling and Preprocessing

Libraries:

- NumPy: For numerical operations and handling arrays.
- Pandas: For data manipulation and analysis, particularly useful for organizing dataset labels and metadata.
- librosa: For audio processing, including feature extraction (e.g., Mel-spectrograms, MFCC, STFT).
- OpenCV/PIL: For image processing especially for resizing and normalizing images.

Feature Extraction

Audio Features

- Mel-spectrogram: Represents audio signals in the Mel frequency scale.
- MFCC: Mel-frequency cepstral coefficients, capturing the short-term power spectrum of sound.
- STFT: Short-Time Fourier Transform, analyzing the frequency content of audio signals over short time intervals, providing a time-frequency representation of the signal.

Visual Features Image resizing to 48x48 pixels for CNN input.

Model Architecture

Convolutional Neural Networks (CNNs):

- Two separate CNN architectures for processing audio and visual data.
- Keras: Part of TensorFlow, used to build and train the CNN models.
- Multi-input Model:
- Combine audio and visual CNN outputs using ‘Concatenate‘ to create a unified representation for final classification.

Table 3 Audio CNN

Layer(type)	Output Shape	Param
separable_conv2d (SeparableConv2D)	(None, 77, 150, 32)	73
max_pooling2d (MaxPooling2D)	(None, 38, 75, 32)	0
separable_conv2d_1 (SeparableConv2D)	(None, 38, 75, 64)	2400
max_pooling2d_1 (MaxPooling2D)	(None, 19, 37, 64)	0
separable_conv2d_2 (SeparableConv2D)	(None, 19, 150, 32)	8896
max_pooling2d_2 (MaxPooling2D)	(None, 19, 37, 128)	0
flatten (Flatten)	(None, 20736)	0
reshape (Reshape)	(None, 162, 1128)	0
lstm (LSTM)	(None, 128)	131584
dense (Dense)	(None, 128)	16512
dropout (Dropout)	(None, 128)	0
dense$_{(}Dense)$	(None, 14)	1806

Table 4 Visual CNN

Layer(type)	Output Shape	Param
separable_conv2d (SeparableConv2D)	(None, 77, 150, 32)	73
max_pooling2d (MaxPooling2D)	(None, 38, 75, 32)	0
separable_conv2d_1 (SeparableConv2D)	(None, 38, 75, 64)	2400
max_pooling2d_1 (MaxPooling2D)	(None, 19, 37, 64)	0
separable_conv2d_2 (SeparableConv2D)	(None, 19, 150, 32)	8896
max_pooling2d_2 (MaxPooling2D)	(None, 19, 37, 128)	0
flatten (Flatten)	(None, 20736)	0
reshape (Reshape)	(None, 162, 1128)	0
lstm (LSTM)	(None, 128)	131584
dense (Dense)	(None, 128)	16512
dropout (Dropout)	(None, 128)	0
dense$_{(}Dense)$	(None, 14)	1806

Training Techniques

- Loss Function: Categorical Crossentropy for multi-class classification.
- Optimizer: Adam optimizer for effective weight updates during training.
- Callbacks: Early stopping to avoid over-fitting.

Model checkpointing to save the best model during training.

Evaluation Metrics

- Accuracy: Primary metric to evaluate model performance.
- Confusion Matrix: To visualize performance across different classes.
- Precision, Recall, F1-score: To evaluate the quality of classifications, especially important in a multi-class setting.

Libraries for Evaluation and Visualization

- Matplotlib/Seaborn: For plotting loss curves, accuracy curves, and confusion matrices.
- Scikit-learn: For calculating evaluation metrics and confusion matrices.

Development Environment

Google Colab: For training models, utilizing free GPU resources, and managing datasets.

The overall pipeline involves data collection from various datasets, preprocessing using audio and image libraries, feature extraction using deep learning techniques, followed by building a multi-input CNN architecture. The model is trained using Keras/TensorFlow, monitored for performance using various metrics, and visualized with Matplotlib/Seaborn.

6.1 Results

In this research, the preliminary work was conducted on an HP Pavilion Aero laptop 13-be20xx equipped with a Ryzen 7 7735U processor.

In the context of the multimodal emotion classification model designed to identify 14 distinct emotions, the outputs of the model are organized to reflect the various classes that represent both posed and real emotions. The model classifies seven real emotions, which include Anger (0), where it identifies expressions or vocalizations indicative of genuine anger. The Neutral (1) class represents a lack of strong emotional expression, indicating a state where no particular emotion is being conveyed. Disgust (2) captures expressions or tones that convey feelings of distaste or revulsion, while Fear (3) identifies vocal signals or facial expressions associated with genuine fear. Additionally, the model recognizes Happy (4) expressions that showcase joy or positive emotions and Sadness (5), which reflects signs of sadness through facial cues or vocal intonations. Lastly, the model captures Surprise (6), identifying expressions or sounds that indicate astonishment or unexpected reactions.

The model also distinguishes between posed emotions, which are intentionally displayed rather than felt authentically. Posed Anger (7) represents instances where the anger expression is deliberately acted out, contrasting with genuine anger. Similarly, the Posed Neutral (8) class captures a neutral expression that is intentionally displayed, while Posed Disgust (9) reflects a deliberate expression of disgust. The model also identifies Posed Fear (10), which represents a posed expression of fear that differs from true fear. Moreover, Posed Happy (11) indicates a happy expression that is created purposefully, and Posed Sadness (12) captures a deliberately expressed sadness. Finally, Posed Surprise (13) reflects an intentional display of surprise, completing the comprehensive classification of the 14 distinct emotions within the model.

The actual model implementations and experiments were performed using Google

Colab, leveraging its enhanced computational support for deep learning tasks. The models developed include an Audio Convolutional Neural Network (CNN) and a Visual CNN, both aimed at classifying emotions based on multimodal inputs. The audio CNN achieved a training accuracy of 74.70%, while the visual CNN reached a training accuracy of 72.16%.

Visual

The classification report shows that the model achieved an overall accuracy of 72%, with precision, recall, and F1 scores varying across emotions, where 'Posed Disgust' scored the highest precision of 0.93, and 'Disgust' exhibited the lowest recall of 0.40.

Table 5 Classification Report for Visual CNN

Emotion	Precision	Recall	f1-score	Support
Anger	0.74	0.55	0.63	174
Neutral	0.64	0.71	0.67	1881
Disgust	0.85	0.40	0.55	109
Fear	0.50	0.41	0.45	927
Happy	0.80	0.84	0.82	2989
Sadness	0.72	0.54	0.62	492
Surprise	0.71	0.75	0.73	1124
Posed Anger	0.91	0.91	0.91	33
Posed Neutral	0.73	0.65	0.69	17
Posed Disgust	0.93	0.95	0.94	41
Posed Fear	0.86	0.90	0.88	21
Posed Happy	0.90	0.96	0.93	48
Posed Sadness	1.00	0.43	0.61	23
Posed Surprise	0.95	0.98	0.96	56
Accuracy			0.72	7935
Macro Avg	0.80	0.71	0.74	7935
Weighted Avg	0.71	0.72	0.71	7935

Graphical Analysis

The confusion matrix shows strong classification performance for happy, neutral, and surprise, with happy having the highest correct predictions (2517). Fear and sadness are often confused with neutral and surprise. For posed emotions, posed anger and posed surprise are well classified, while posed neutral and posed sadness have some confusion with neutral and fear. Overall, the model performs well on distinguishing emotions.

The ROC curve illustrates the model's performance across different threshold settings, plotting the true positive rate (TPR) against the false positive rate (FPR) for each emotion class.

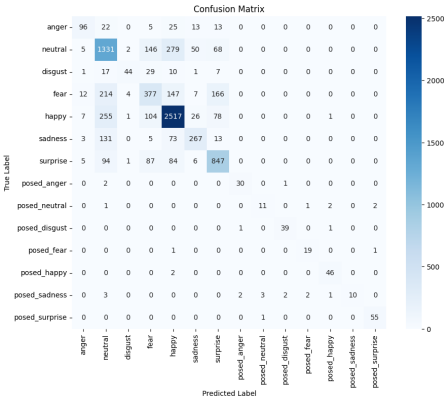

Fig. 2 Confusion matrix for Visual CNN

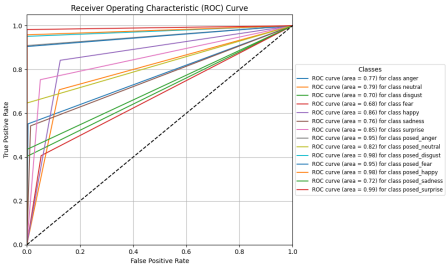

Fig. 3 ROC Curve for Visual Model

Each colored line represents the ROC curve for a specific emotion, with the area under the curve (AUC) values indicating how well the model distinguishes between classes:

- High AUC (0.95-0.99): Indicates excellent performance, especially for posed anger, posed fear, and posed surprise.
- Moderate AUC (0.68-0.86): Indicates varying performance for fear, sadness, and happy, where the model struggles to differentiate those emotions effectively.
- The diagonal dashed line represents a random classifier (AUC = 0.5). Curves above this line signify better-than-random classification abilities.

Overall, the model shows strong classification capabilities for most emotions, especially the posed emotions.

The Precision-Recall curve illustrates the performance of the visual model for each class by plotting precision against recall.

- High Performance: Classes like posed anger and posed surprise show high precision and recall, indicating effective emotion identification.
- Steeper Decline: The curves for fear and disgust reveal a trade-off between precision and recall, suggesting challenges in maintaining both metrics.
- Variability: Class neutral exhibits fluctuating performance, while sad and posed neutral demonstrate lower precision and recall at higher thresholds.

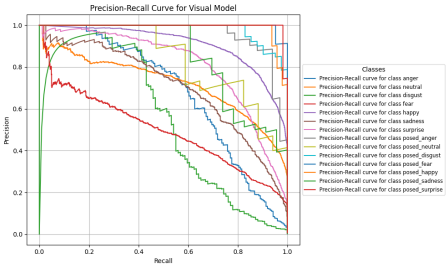

Fig. 4 Precision-Recall Curve for Visual Model

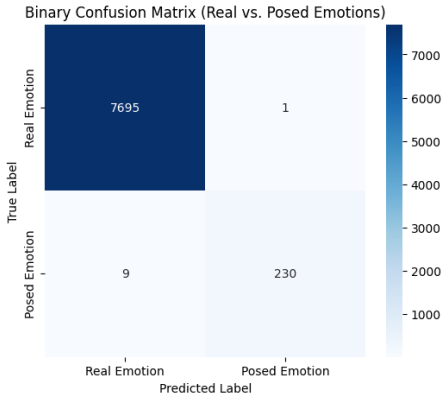

Fig. 5 Confusion Matrix for Binary Visual Model

Summary: These curves highlight the model's strengths and weaknesses in emotion classification, pinpointing classes that excel and those needing improvement.

Binary Classification

This binary confusion matrix illustrates the model's ability to distinguish between real and posed emotions.

- 1. High True Positive Rate for Real Emotions: Out of 7696 instances of real emotions, 7695 were correctly classified. This indicates the model is exceptionally accurate in identifying real emotions.
- Low False Positives: Only 1 instance of a real emotion was misclassified as posed, showing the model's strong reliability in avoiding false positives for real emotions.
- Good Performance on Posed Emotions: While there is some misclassification (9 instances), the model correctly identified 230 out of 239 posed emotions. This is a strong result given the subtle differences between real and posed emotions.

These results highlight the model's robustness, especially in detecting real emotions accurately, and show a solid performance in identifying posed emotions, with minimal misclassification.

Table 6 Classification Report for Binary Classification of Visual model

Emotion Type	Precision	Recall	F1-score	Support
Real Emotion	1.00	1.00	1.00	7696
Posed Emotion	1.00	0.96	0.98	239
Accuracy		1.00		
Macro Avg	1.00	0.98	0.99	7935
Weighted Avg	1.00	1.00	1.00	7935

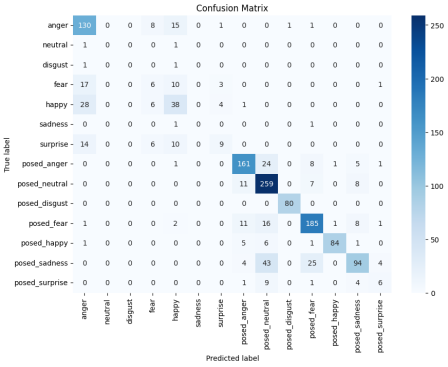

Fig. 6 Confusion matrix for Audio Model

Audio

The classification report shows that the model achieved an overall accuracy of 74.7%, with precision, recall, and F1 scores varying across emotions, where 'posed happiness' scored the highest precision of 0.90, and 'posed sadness' exhibited the lowest recall of 0.43.

Graphical Analysis

The confusion matrix reveals that the model performs well on certain emotions, particularly posed neutral (259 correctly classified) and posed disgust (80 correctly classified), with high precision in these areas. The model's strong performance in these classes indicates it effectively captures distinct and exaggerated expressions typical of posed emotions, particularly when there are clear, consistent features. These emotions may benefit from well-defined training examples, contributing to the

Table 7 Classification Report for Audio CNN

Emotion	Precision	Recall	f1-score	Support
Anger	0.70	0.78	0.74	156
Neutral	0.00	0.00	0.00	2
Disgust	0.00	0.00	0.00	2
Fear	0.24	0.27	0.25	37
Happy	0.55	0.42	0.47	77
Sadness	0.00	0.00	0.00	2
Surprise	0.53	0.54	0.53	39
Posed Anger	0.83	0.76	0.79	201
Posed Neutral	0.70	0.91	0.79	285
Posed Disgust	0.99	1.00	0.99	80
Posed Fear	0.77	0.81	0.79	225
Posed Happy	0.95	0.86	0.90	98
Posed Sadness	0.88	0.55	0.67	170
Posed Surprise	0.50	0.29	0.36	21
Accuracy			0.75	1395
Macro Avg	0.54	0.51	0.52	1395
Weighted Avg	0.75	0.75	0.74	1395

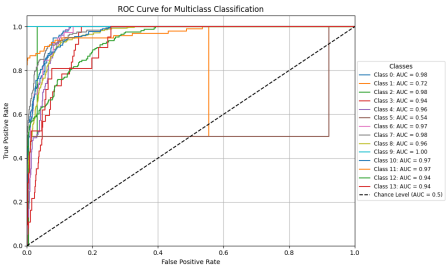

Fig. 7 ROC Curve for Audio Model

high classification accuracy in these categories.

This ROC curve illustrates the model's performance in a multiclass classification setup, showing how well it distinguishes between different classes. Each line represents the true positive rate (sensitivity) versus the false positive rate for one of the 14 emotion classes, with the corresponding area under the curve (AUC) score shown in the legend.

- High AUC Scores: Classes like 0 (0.98), 2 (0.98), 7 (0.98), 8 (0.96), 9 (1.00), 10 (0.97), and 11 (0.97) have high AUC values, indicating that the model is effective at distinguishing these classes from others. AUC scores close to 1.0 signify strong predictive performance.

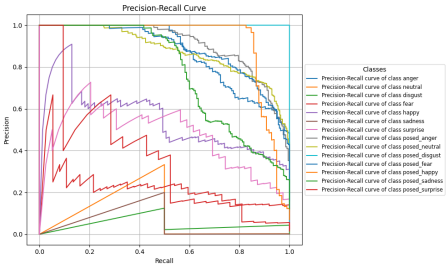

Fig. 8 Precision-Recall Curve for Audio Model

- Moderate to Low AUC Scores: Class 1 (0.72) and Class 5 (0.54) have lower AUC values, suggesting that the model struggles to separate these classes from others. These may be challenging to classify due to similar characteristics with other emotions or insufficient distinguishing features in the data.

Overall, the majority of classes show strong discrimination ability with high AUC values, indicating good model performance in classifying emotions.

This precision-recall curve reveals some strong points in the model's performance:

- High Precision and Recall for Certain Classes: Some classes, like neutral, anger, and sadness, have curves that remain relatively high, indicating the model can achieve both high precision and recall. This means that for these classes, the model is both accurate (high precision) and comprehensive in its detections (high recall).
- Clear Distinction Between Some Emotions: The curves for classes like neutral and anger remain distinct from other classes, suggesting the model effectively differentiates these emotions, even at different thresholds.
- Good Coverage Across All Classes: While there's variation in performance, the model provides a precision-recall curve for each class, indicating that it has learned to some degree across a wide range of emotions, including subtle ones like posed surprise and posed sadness.

These strong points show that the model has a good foundation, especially for commonly encountered emotions, and indicates areas where it may only need fine-tuning rather than major adjustments.

Binary Classification

This confusion matrix represents the model's performance in distinguishing between posed and real emotions.

- High True Positive Rates: The model correctly classified 307 instances of posed emotions and 1,074 instances of real emotions. This shows strong accuracy in recognizing both categories.
- Low Misclassification Rates: Only 8 posed emotions were misclassified as real, and only 6 real emotions were misclassified as posed. This indicates that the model is generally reliable, with minimal errors in distinguishing between the two types.

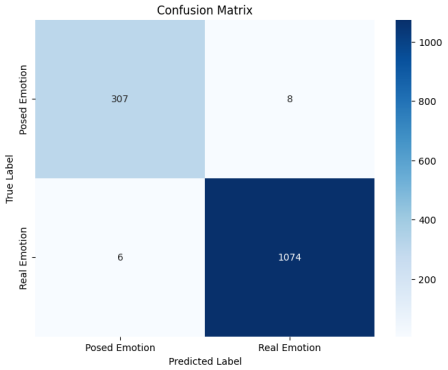

Fig. 9 Confusion Matrix for Binary Audio CNN

- Clear Separation Between Classes: The low number of misclassifications in both categories suggests the model effectively distinguishes between posed and real emotions, capturing the nuances between them.

Overall, this matrix highlights the model's strong performance and reliability in correctly classifying both posed and real emotions, with very few instances of confusion.

Table 8 Classification Report for Binary Classification of Audio model

Emotion Type	Precision	Recall	F1-score	Support
Real Emotion	0.98	0.97	0.98	315
Posed Emotion	0.99	0.99	0.99	1080
Accuracy		0.99		
Macro Avg	0.99	0.98	0.99	1395
Weighted Avg	0.99	0.99	0.99	1395

Research Study	Model/Method Used	Dataset/Emotion Type	Accuracy
Sincere vs. Fake Expressions with Facial Landmarks	RD Network with LM5	Facial Expressions (General)	73%
Authentic vs. Fake Anger Expressions	LSTM	Anger	80%
Differentiating Genuine and Fake Smiles via Physiological Signals	Features from Pupillary Response	Smile Detection	98.8%
Laughter and Smile Detection under Different Emotional Conditions	Transfer Learning	Laughter Detection	83%
Genuine vs. Unfelt Expressions with VGG-Face and EMNet	VGG-Face, EMNet	CK+ Dataset (General Expressions)	98.7%
Our model	Multimodal (Audio and Visual)	Visual: CK+, FER, RAF-DB, JAFFE / Audio: ShEMO, TESS, CREMA-D, Vivae, MAV	98.56% (Audio) / 99.65% (Visual)

Table 9 Comparative Analysis of Different Emotion Detection Models

Emotion	Prediction Probabilities
Posed Neutral	0.90
Posed Anger	0.05
Posed Sadness	0.03
Posed Happy	0.02
Posed Neutral	0.00

Table 10 XAI Analysis for Audio Model

Comparative Analysis

7 XAI Analysis

Audio

LIME (Local Interpretable Model-agnostic Explanations) was used to explain predictions made by a trained audio CNN model. First, a 'LimeTabularExplainer' is created, reshaping the training data to 2D and specifying feature and class names. A specific instance from the test data is selected, reshaped, and passed to a prediction function ('predict_fn'), which reshapes the input and predicts probabilities using the CNN model. LIME then explains the model's prediction by identifying the most important features.
The following results were derived for a randomly selected data:

Table 11 Feature Analysis

Feature Condition	Value
Feature_9881 > 0.00	0.0017842104435679335
Feature_10272 > 6.35	0.0014811276113776623
$-53.40 <$ Feature_6913 ≤ -43.65	0.0014396095467377622
$-57.76 <$ Feature_8744 ≤ -45.48	0.0014114545394854684
Feature_109 ≤ -28.41	0.0014108099298412882
$-50.19 <$ Feature_6463 ≤ -43.07	0.0014048525896023696
Feature_8485 ≤ -61.46	0.0013723561484326968
Feature_7821 ≤ -59.78	0.001357718466111189
$-18.77 <$ Feature_10551 ≤ -3.15	0.0013036393529666528
Feature_4275 ≤ -53.15	0.001149079296957994

Visual

The XAI analysis of the visual model aims to implement and numerically analyze Grad-CAM (Gradient-weighted Class Activation Mapping) heatmaps for a multi-input deep learning model. Initially, the 'make_gradcam_heatmap' function is defined to generate the heatmap by computing gradients and pooling them over the feature maps

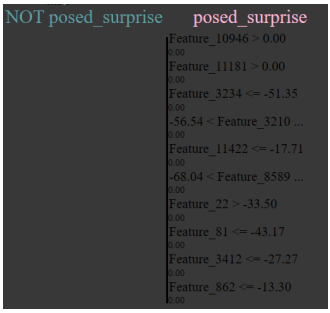

Fig. 10 Audio XAI Analysis - Posed Surprise

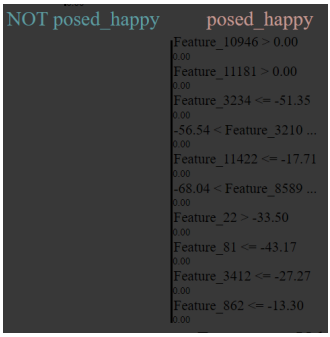

Fig. 11 Audio XAI Analysis - Posed Happy

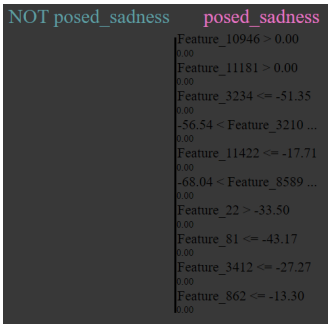

Fig. 12 Audio XAI Analysis - Posed Sadness

Fig. 13 Audio XAI Analysis - Posed Anger

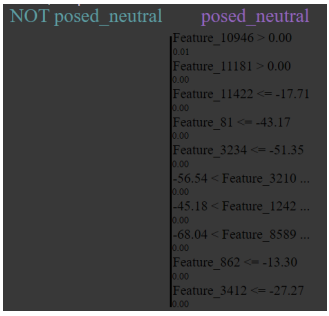

Fig. 14 Audio XAI Analysis - Posed Neutral

from the last convolutional layer. This function normalizes the heatmap values to highlight the important regions that contribute to the model's prediction. Subsequently, the 'display_gradcam' function visualizes the generated heatmap superimposed on the original image, providing a visual interpretation of the model's focus areas. For numerical analysis, the heatmap values are extracted, normalized, and flattened to facilitate statistical analysis. The resulting data is then used to calculate basic statistics such as mean, maximum, minimum, and standard deviation, and to visualize the distribution of heatmap values through a histogram. This dual approach of visual and

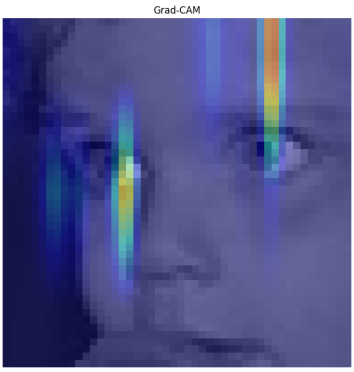

Fig. 15 Visual XAI Analysis - Grad-CAM

numerical analysis enhances the interpretability of the model by quantifying and visualizing the impact of different image regions on the prediction. The following results were obtained:

Numerical Heatmap Values:

$$\begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.15 & 1.00 & 0.26 & 0.05 & 0.00 \\ 0.00 & 0.00 & 0.19 & 0.06 & 0.37 & 0.00 & 0.91 & 0.73 & 0.04 & 0.27 \\ 0.07 & 0.79 & 0.30 & 0.08 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.28 & 0.02 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

The results

Table 12 Statistics of Heatmap Values

Statistic	Value
Mean Heatmap Value	0.07730904966592789
Max Heatmap Value	1.0
Min Heatmap Value	0.0
Standard Deviation	0.2081332951784134

8 Ablation Study

Ablation studies reveal feature extraction improves both audio and visual model performance, enhancing emotion classification accuracy, especially for posed emotions.

For the audio model, we tested different feature extraction methods, including Mel-frequency cepstral coefficients (MFCCs), Mel spectrograms, and Short-Time Fourier Transform (STFT), each capturing distinct aspects of vocal expression. By systematically adjusting these features' dimensions and excluding certain techniques, we found

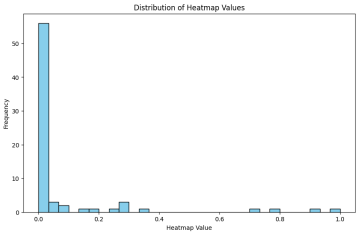

Fig. 16 Visual XAI Analysis - Graphical Visualisation

that the combined use of spectral and temporal features significantly boosted the model's performance, highlighting their importance in identifying posed emotions.

Table 13 Comparative Classification Report for Audio Model with and without Feature Extraction

Emotion	Precision		Recall		F1-Score		Support
	w/o FE	w/ FE	w/o FE	w/ FE	w/o FE	w/ FE	
Anger	0.00	0.62	0.00	0.73	0.00	0.67	197
Neutral	0.00	0.40	0.00	0.67	0.00	0.50	18
Disgust	0.00	1.00	0.00	0.18	0.00	0.31	11
Fear	0.00	0.54	0.00	0.14	0.00	0.22	51
Happy	0.00	0.42	0.00	0.32	0.00	0.36	91
Sadness	0.00	0.64	0.00	0.64	0.00	0.64	14
Surprise	0.00	0.27	0.00	0.31	0.00	0.29	39
Posed Anger	0.00	0.85	0.00	0.78	0.00	0.81	201
Posed Neutral	0.19	0.71	1.00	0.87	0.32	0.78	286
Posed Disgust	0.00	0.99	0.00	0.97	0.00	0.98	80
Posed Fear	0.00	0.79	0.00	0.80	0.00	0.79	225
Posed Happy	0.00	0.94	0.00	0.83	0.00	0.88	98
Posed Sadness	0.00	0.75	0.00	0.69	0.00	0.72	170
Posed Surprise	0.00	0.50	0.00	0.24	0.00	0.32	21

Table 13 compares the audio classification model's performance with and without feature extraction (FE) across different emotions. Using FE significantly improves precision, recall, and F1-scores for most emotions, especially for complex posed emotions like "Posed Neutral" and "Posed Disgust," underscoring FE's value in enhancing classification accuracy.

For the visual model, another ablation study focused on analyzing specific facial regions, particularly the eyebrow and eyeframe areas. This approach allowed the model to capture subtle cues of emotional authenticity, such as micro-expressions around the eyes. The results showed that incorporating targeted region analysis led to increased accuracy in distinguishing between real and posed emotions, proving

Table 14 Accuracies of Audio Model with and without Feature Extraction

	Accuracy
Without Feature Extraction	0.19
With Feature Extraction	0.72

the effectiveness of detailed, region-specific visual analysis. Together, these studies underscore the value of both targeted feature extraction in audio and region-focused preprocessing in visual data for robust posed emotion classification.

Table 15 Comparative Classification Report for Visual Model with and without Feature Extraction of Eye region and Eyeframe

Emotion	Precision		Recall		F1-Score		Support
	w/o FE	w/ FE	w/o FE	w/ FE	w/o FE	w/ FE	
Anger	0.62	0.74	0.60	0.55	0.61	0.63	174
Neutral	0.69	0.64	0.60	0.71	0.64	0.67	1881
Disgust	0.47	0.85	0.60	0.40	0.53	0.55	109
Fear	0.48	0.50	0.43	0.41	0.45	0.45	927
Happy	0.80	0.80	0.85	0.84	0.82	0.82	2989
Sadness	0.58	0.72	0.73	0.54	0.64	0.62	492
Surprise	0.70	0.71	0.71	0.75	0.71	0.73	1124
Posed Anger	0.84	0.91	0.94	0.91	0.89	0.91	33
Posed Neutral	0.89	0.73	0.73	0.65	0.80	0.69	17
Posed Disgust	0.95	0.93	0.80	0.95	0.87	0.94	41
Posed Fear	0.81	0.86	0.89	0.90	0.85	0.88	21
Posed Happy	0.95	0.90	0.95	0.96	0.95	0.93	48
Posed Sadness	0.53	1.00	0.85	0.43	0.65	0.61	23
Posed Surprise	0.88	0.95	0.91	0.98	0.90	0.97	56

Table 15 shows the impact of eye region and eyeframe feature extraction (FE) on the visual model's classification performance across emotions. Using FE generally improves precision, recall, and F1-scores for both real and posed emotions, with notable gains in challenging categories like "Posed Anger" and "Posed Disgust." This demonstrates the importance of focused visual features in enhancing model accuracy.

9 Threats to Validity

One concern is data bias. If the datasets used for training and testing the model are not representative of the broader population or lack diversity in terms of age, gender, ethnicity, or cultural background, the model may fail to accurately classify emotions for individuals outside of the demographic that the data was derived from. Another potential threat to validity arises from contextual variability. The model may not account for situational contexts in which emotions are expressed. Emotions can be influenced by environmental factors, social settings, or specific interactions, which may not be reflected in the training data.

Table 16 Accuracies of Visual Model with and without Feature Extraction

	Accuracy
Without Feature Extraction	0.58
With Feature Extraction	0.72

10 Conclusion and Future Work

In conclusion, this study presents a multimodal approach to emotion classification, capable of distinguishing between 14 different emotions, encompassing both genuine and posed expressions. Leveraging separate CNN architectures for processing audio and visual data, the model achieved substantial training accuracies, demonstrating the effectiveness of multimodal integration. This design enabled the model to capture complex emotional indicators in vocalizations and facial expressions, underscoring the value of combining these data types for a comprehensive understanding of emotional cues. Notably, the model performed well in differentiating real emotions, highlighting its potential for applications in areas where accurate emotion recognition is essential. The distinction between posed and genuine emotions is a significant contribution of this model, as it aligns with practical scenarios where detecting subtle differences is crucial, such as in mental health, entertainment, and human-computer interaction contexts. The model's current accuracy indicates its effectiveness but also emphasizes areas for improvement, especially in classifying less distinct emotions and managing varied expressions across individuals and settings. This approach, therefore, provides a strong foundation for continued exploration into emotion classification, with a focus on enhancing both the depth and adaptability of such models. Future work will involve the use of a more comprehensive and designated dataset that better represents demographic, cultural, and contextual diversity. Additionally, refining the model architecture to incorporate advanced techniques such as attention mechanisms or transformers could improve its sensitivity to nuanced emotional cues. Expanding the dataset and employing more sophisticated architectures will enable a more robust model that can generalize across diverse settings and differentiate between subtle emotional variations. This enhanced approach will allow for a more accurate, contextually aware emotion classification system that is adaptable to real-world applications.

References

[1] Florio, G.F., Buemi, M.E., Acevedo, D., Negri, P.: Attribute classification for the analysis of genuineness of facial expressions. In: 11th International Conference of Pattern Recognition Systems (ICPRS 2021), vol. 2021, pp. 109–114 (2021). https://doi.org/10.1049/icp.2021.1467