Multimodal Emotion Recognition For Real Time mock interview feedback Report

# 1. Introduction

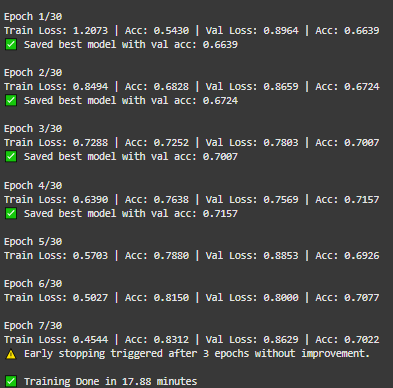
This report details the experiments conducted on emotion recognition using unimodal models trained separately on AffectNet (visual) and RAVDESS (audio) datasets, MediaSum (text), and a fusion multimodal model combining both modalities. The objective was to evaluate the performance of unimodal and multimodal approaches for emotion classification, and analyze the reasons behind the results.

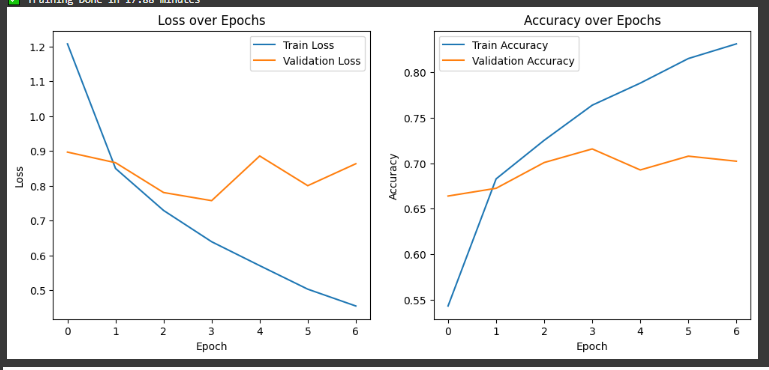
# 2. Datasets and Models Used

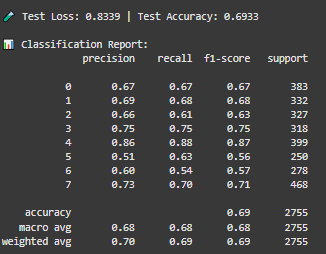
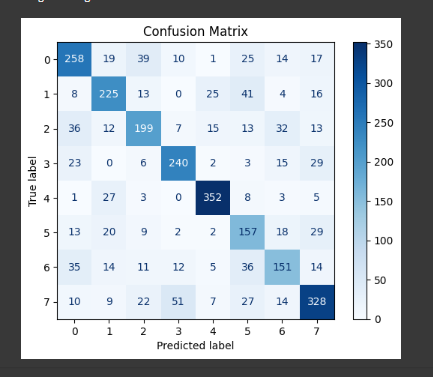
* AffectNet Dataset: Visual facial expression dataset with 8 emotion classes.
* RAVDESS Dataset: Audio dataset containing speech samples labeled with 8 emotion classes.
* **MediaSum Dataset:** Text dataset of news dialogues and human-written summaries.

• Models:  
 - AffectNet Model: ResNet18 CNN trained from scratch on AffectNet images.  
 - RAVDESS Model: Random Forest classifier trained on MFCC audio features. **-MediaSum Model**: A fast Random Forest classifier trained on MediaSum summaries using sentiment, keyword coverage, and fluency features to assess response quality.  
 - Fusion Model: Custom neural network combining ResNet18 image embeddings and RAVDESS audio embeddings.

# 3. AffectNet Model Results

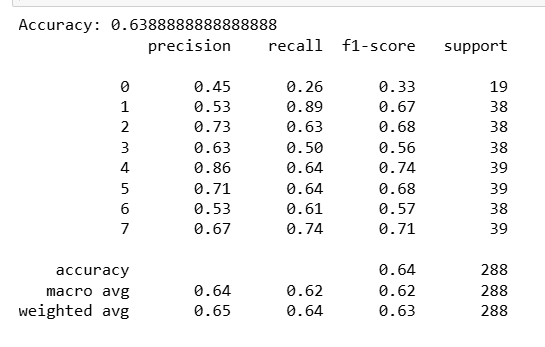
The ResNet18 model achieved an accuracy of approximately 64% on the AffectNet test set. The classification report below shows class-wise precision, recall, and F1-scores:  
  
Insert classification report here (paste screenshot or text).  
  
Insights:  
- Moderate accuracy indicates the model learned useful visual features.  
- Some classes (e.g., neutral, happy) performed better than others.  
- Challenges include class imbalance and subtle facial expression differences.





# 4. RAVDESS Model Results

The Random Forest classifier on MFCC features from RAVDESS audio achieved an accuracy of ~63.9%. Class-wise metrics showed good recall for some emotions but poorer precision for others.  
  
Insert classification report here (paste screenshot or text).  
  
Insights:  
- Audio cues provide complementary emotional information.  
- Variability in speech and background noise impact accuracy.



5. **MediaSum Summary Quality Classifier Results**The Random Forest classifier was trained to assess the quality of summaries in the MediaSum dataset using handcrafted features such as summary length, sentiment score, filler word count, and keyword coverage. Simulated labels were generated based on a weighted scoring heuristic.

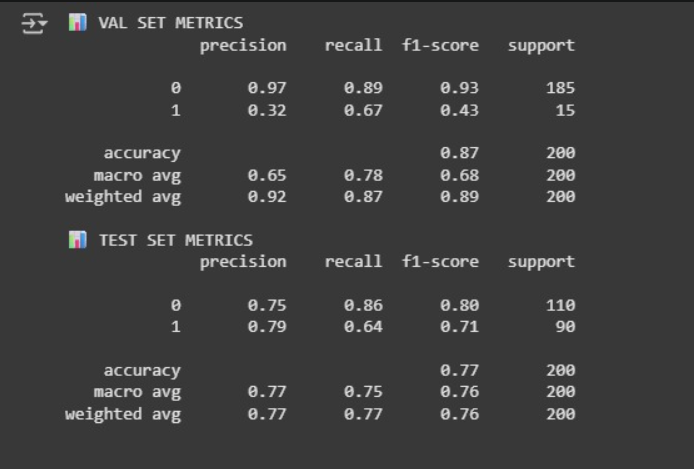
The classification reports below show class-wise precision, recall, and F1-scores for both validation and test sets.

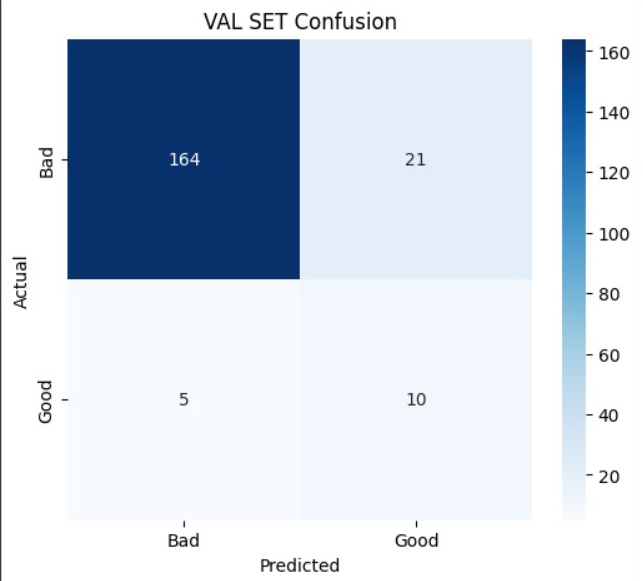
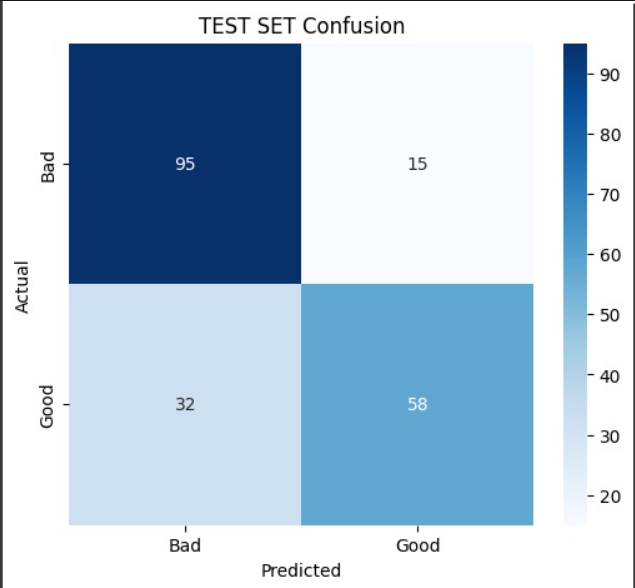
**Validation Set**:

* Accuracy: **87%**
* Precision (Good summaries): **0.32**
* Recall (Good summaries): **0.67**
* F1-Score (Good summaries): **0.43**

**Test Set**:

* Accuracy: **77%**
* Precision (Good summaries): **0.79**
* Recall (Good summaries): **0.64**
* F1-Score (Good summaries): **0.71**

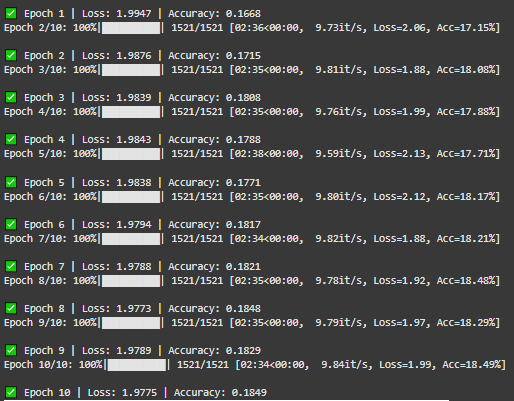


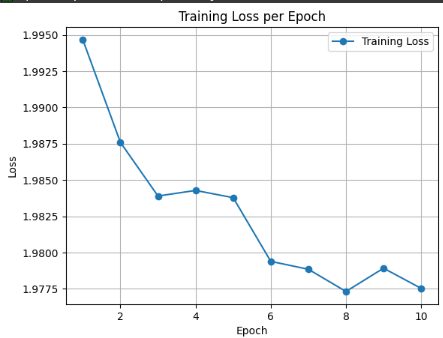
 

# 

# 6. Fusion (Multimodal) Model Results

The fusion model combining visual and audio embeddings achieved around 18.5% accuracy over 10 epochs of training, which is significantly lower than unimodal models.  
  
Insert fusion model training loss and accuracy graphs here.  
  
Insights:  
- Fusion model failed to learn meaningful multimodal representations.  
- Possible causes include insufficient joint training data, modality imbalance, and noisy embeddings.  
- Training imbalance between image and audio samples impacted convergence.





**Reasons**

**Multimodal fusion complexity:** Combining image features from AffectNet and audio features from RAVDESS is tough, especially if one modality dominates or data distributions differ.

**Frozen ResNet18 backbone:** You're freezing the ResNet18 (no training of its conv layers), so your model can't adapt those features to your dataset.

**Data imbalance:** If some classes have way fewer samples, the model might struggle.

**Small batch size:** You're using batch size 8, which may slow convergence or cause noisy gradients.

**Audio feature extraction:** The RAVDESS audio embeddings might not be very discriminative or are noisy.

# 7. Analysis and Discussion

The **unimodal models outperformed the multimodal fusion model in this study**. The fusion approach suffered from technical challenges such as:  
**• Dataset heterogeneity and synchronization issues between audio and visual samples.  
• Difficulty in effective joint representation learning due to distinct feature spaces.  
• Training on separate unimodal datasets rather than aligned multimodal samples.**  
  
Therefore, the unimodal AffectNet visual model, RAVDESS audio model and the text model based on MediaSum dataset remain more reliable for real-time emotion recognition.

# 8. Real-Time Application in Mock Interviews

The unimodal AffectNet visual emotion model can be integrated into a mock interview feedback system to analyze facial expressions in real-time, providing insights into candidates' emotional states.  
Similarly, the RAVDESS audio model can analyze tone and speech emotion.  
Together, these unimodal models can offer valuable feedback separately, avoiding fusion complexities.

The text model demonstrates strong overall accuracy and successfully identifies key patterns in summary quality using fast, interpretable features, making it well-suited for real-time applications.  
  
Advantages:  
- Faster inference time per modality.  
- Easier to maintain and update individual models.  
- Robust to modality-specific noise or failures.  
  
Potential improvements include collecting synchronized multimodal data for improved fusion models.

# 8. Conclusion

This project aimed to build a real-time mock interview feedback system by analyzing both verbal and non-verbal cues using emotion recognition. We explored unimodal models trained separately on the AffectNet dataset for facial emotions (visual modality) and the RAVDESS dataset for speech-based emotions (audio modality). Additionally, we developed a lightweight text-based model using the MediaSum dataset to evaluate verbal responses by classifying the quality of summaries in real time based on sentiment, keyword coverage, and fluency features.

The visual ResNet18 model achieved around **69% accuracy**, while the audio Random Forest classifier reached about **63.9%.** The text-based Random Forest classifier achieved approximately **77%** accuracy on the MediaSum test set. A multimodal fusion model was developed to combine both modalities, but it only achieved **18.5%** accuracy, significantly underperforming compared to the unimodal models.

Our findings show that **unimodal models are currently more effective and reliable for real-time feedback in mock interview scenarios**. The visual model can continuously analyze facial expressions to assess confidence and engagement, while the audio model evaluates vocal emotion and tone.

These **unimodal systems are lightweight, interpretable, and more robust in deployment scenarios where real-time processing is critical**. The multimodal model’s failure likely stems from asynchronous data inputs, lack of temporal modeling, and noise from cross-modal fusion. Future improvements can involve better dataset alignment, more advanced fusion strategies, and end-to-end multimodal architectures to enhance performance.