Multimodal Emotion Classification Using Deep Learning

Priscilla Clark
Massachusetts Institute of Technology
poclark@mit.edu

Siddharth Mehta
Massachusetts Institute of Technology
sidmehta@mit.edu

Prashasti Agrawal
Massachusetts Institute of Technology
praa@mit.edu

Yoav Marziano
Massachusetts Institute of Technology
yoavma@mit.edu

Abstract

This study explores multimodal deep learning for emotion recognition, integrating audio and text features using pre-trained embeddings (Wav2Vec2, HuBERT, RoBERTa) and handcrafted acoustic features on the RAVDESS and TESS datasets. Initially, models trained on unscaled features suffered from overfitting in early-fusion architectures, but after feature scaling, early fusion matched or outperformed structured fusion models, with the best achieving 98.53% accuracy. A grouped train-test split revealed that prior results were slightly inflated (97.88% to 93.89%) due to text-based leakage, though audio embeddings remained the dominant signal. These findings emphasize the importance of feature scaling and strict data partitioning in multimodal AI, demonstrating that preprocessing decisions can be as impactful as model architecture choices for optimizing generalization and performance.

1 Introduction

Emotion recognition from speech is a rapidly growing field in artificial intelligence, with applications spanning virtual assistants, mental health monitoring, and customer service. Accurately detecting emotions from spoken language can significantly enhance human-computer interactions, enabling systems to respond with empathy and contextual awareness. However, emotion recognition is inherently challenging due to the complexity of human affect, variations in speech patterns, and the ambiguity of linguistic cues. Traditional approaches to emotion classification often rely on a single modality—either acoustic features derived from speech or textual content extracted from transcripts. While audio captures tone, pitch, and prosody, the text provides semantic context, and neither alone fully encapsulates emotional expression.

This project explores a multimodal deep-learning approach to emotion recognition, leveraging speech and text features to improve classification accuracy. The primary goal is to evaluate different fusion strategies for integrating these modalities and determine whether combining them enhances performance over unimodal models. We investigate six neural network architectures, ranging from simple audio-only baselines to complex multimodal fusion models that integrate pre-trained speech embeddings (Wav2Vec 2.0 and HuBERT), handcrafted audio features, and text embeddings (RoBERTa). These models are trained on a dataset of 8,480 labeled speech utterances, with features extracted using state-of-the-art techniques.

A key challenge addressed in this work is determining the optimal method for combining multimodal features. We explore early-fusion and late-fusion strategies, comparing the effectiveness of structured multi-branch architectures against large, high-capacity networks that merge all features. Additionally, we examine the role of regularization techniques (e.g., dropout) and feature scaling in preventing overfitting when training. Initially, we ran our experiments with unscaled features, which led to significant performance degradation in early-fusion models due to feature magnitude imbalances. However, after rerunning our models with properly scaled features, we found that early-fusion architectures, which initially struggled, could match or even exceed the performance of structured fusion models. Our experiments provide insights into which modalities contribute most to emotion classification, the impact of fusion strategies, and how preprocessing choices—particularly feature scaling—affect model generalization.

Through this study, we aim to advance the understanding of multimodal emotion recognition and contribute to developing more robust and generalizable emotion classification systems. The results demonstrate that properly structured and normalized multimodal models outperform unimodal baselines, with our best-performing network achieving 98.53% test accuracy. Our findings underscore the critical role of feature preprocessing in multimodal deep learning and offer valuable lessons for future research in speech-based emotion AI and fusion-based architectures.

2 Literature Review

Prior to beginning the assignment, we conducted research on existing studies related to multimodal emotion recognition, speech-based classification, and deep learning fusion techniques. Below, we categorize key papers into three focus areas: feature extraction and pretraining, multimodal fusion strategies, and domain-specific adaptation challenges.

2.1 Feature Extraction and Pretraining for Speech and Text

These studies focus on self-supervised learning and deep learning models that extract speech and text representations for emotion recognition.

- **Baevski et al. (2020)** [1]: Introduces Wav2Vec 2.0, a self-supervised learning framework that significantly improves speech representation learning. By masking latent speech representations and solving a contrastive task, the model learns robust embeddings from raw audio with minimal labeled data. This work informs our approach to speech feature extraction, as we leverage Wav2Vec 2.0 for generating rich speech embeddings before multimodal fusion.
- **Chuang et al. (2020)** [2]: Proposes SpeechBERT, a model that jointly learns representations from both audio and text, improving spoken language understanding. By adapting the BERT framework to integrate speech and text modalities, SpeechBERT demonstrates the advantages of early fusion, aligning well with our study's focus on improving multimodal emotion classification.
- **Sharma (2023)** [6]: Evaluates the interplay between speech and text embeddings in emotion recognition models. By comparing models trained on HuBERT speech embeddings and BERT text embeddings, this study shows that jointly trained models significantly outperform unimodal approaches. These findings support our investigation into how combining Wav2Vec 2.0 and RoBERTa embeddings improves multimodal fusion for emotion classification.

These studies guide our feature selection and pretraining strategy. We leverage Wav2Vec 2.0 and HuBERT for speech and RoBERTa/BERT for text, ensuring that our models start with rich contextual embeddings before fusion.

2.2 Multimodal Fusion Strategies for Emotion Recognition

These studies explore different ways to combine speech, text, and visual modalities to improve emotion classification.

- **Huang et al. (2024)** [4]: Introduces MM-NodeFormer, a Transformer-based multimodal fusion model for emotion recognition in conversation. Unlike traditional concatenation methods, MM-NodeFormer dynamically weighs the contributions of speech, text, and visual inputs based on their emotional richness. While this approach integrates three modalities, our work focuses on optimizing speech-text fusion, leveraging Transformer-based structured fusion strategies.

- **Lian et al. (2023)** [5]: Provides a comprehensive review of multimodal emotion recognition techniques, categorizing fusion methods into early, late, and hybrid strategies. The survey also reviews datasets such as IEMOCAP and MELD, which we utilize in our experiments. This work validates our decision to explore both early and late fusion models and helps frame our study within the broader landscape of multimodal fusion techniques.

These studies influence our fusion strategy selection. While MM-NodeFormer integrates text, speech, and visual cues, our work focuses on optimizing speech-text fusion using structured Transformer-based approaches. The survey by Lian et al. helps validate our decision to experiment with both early and late fusion techniques.

2.3 Domain-Specific Adaptation and Challenges in Emotion Recognition

These studies highlight challenges in generalizing emotion recognition models across different domains and datasets.

- **Ewertz et al. (2024)** [3]: Analyzes domain adaptation issues in speech-based emotion models, showing that traditional models trained on acted speech fail to generalize to spontaneous corporate speech. The study introduces FinVoc2Vec, a model tailored for corporate vocal tone classification, reinforcing the importance of domain adaptation. This aligns with our focus on improving multimodal models' generalizability by incorporating feature normalization and data augmentation to mitigate dataset biases.

This study reinforces our focus on feature preprocessing and generalization. Since emotion recognition models struggle with dataset variability, our approach incorporates feature normalization and suggests data augmentation as a next step to improve cross-domain robustness.

3 Approach

3.1 Dataset and Feature Extraction

We assembled a dataset of spoken utterances labeled with seven emotional categories (anger, fear, sadness, neutral, happiness, surprise, disgust). The dataset consists of 8,480 audio recordings drawn from two public emotion corpora (the RAVDESS and TESS datasets), each clip paired with its transcript (a spoken sentence). The spoken phrases in these datasets are semantically neutral (e.g., "Dogs are sitting by the door."), meaning most emotion is conveyed through vocal tone rather than word choice.

3.1.1 Audio Preprocessing

Each audio clip was loaded at a 16 kHz sampling rate and trimmed/padded to a 3-second duration for uniformity. We extracted a rich set of traditional acoustic features using Librosa. This included 40-dimensional MFCCs (Mel-frequency cepstral coefficients) averaged over time, a 12-dimensional chroma vector (averaged), spectral statistics (centroid and rolloff frequencies), and the average zero-crossing rate. We also extracted prosodic features: using Librosa's pyin we estimated the pitch contour, took the mean pitch (to capture fundamental frequency), and computed an approximation of formant frequencies via LPC. Additionally, we measured voice signal quality features such as HNR (harmonics-to-noise ratio, approximated via RMS energy), jitter (variation in pitch), shimmer (variation in amplitude), and an estimated speech rate. These hand-crafted features provide interpretable cues about tone and vocal delivery. Each feature vector (MFCCs, chroma, etc.) was stored for each sample.

3.1.2 Audio Embeddings

We also leveraged pre-trained deep speech models to extract high-level audio representations. We used Facebook's Wav2Vec 2.0 and HuBERT (both pre-trained on large speech corpora) to obtain 768-dimensional embeddings of each audio clip. Specifically, the raw waveform was fed into the pre-trained models' encoder. We mean-pooled the last hidden state over time to produce a fixed-length embedding (1×768) for Wav2Vec and similarly for HuBERT. These embeddings encode rich prosodic and phonetic information learned from data, complementing the handcrafted features. We expected them to capture subtle cues (tone, inflection) that might not be reflected in simple statistics.

3.1.3 Transcription and Text Embeddings

Since the datasets did not provide exact transcripts for each audio, we generated them using OpenAI's Whisper speech-to-text model. We applied Whisper (small model) for each audio file to produce an English transcript. Given the controlled nature of the recordings, Whisper reliably produced the spoken sentence (e.g., "Dogs are sitting by the door.". We then obtained a text embedding for each transcript using a pre-trained RoBERTa language model. We tokenized the text and passed it through RoBERTa-base, extracting the 768-dimensional CLS-token embedding from the final layer. This embedding serves as a condensed representation of the content of the sentence. (Notably, because the spoken content is often neutral, the text embeddings mainly encode lexical information rather than emotion; however, they could capture linguistic features like syntax or any subtle word choice differences if present.)

3.1.4 Feature Aggregation

After extracting all features for each sample, we aggregated them into a structured dataset. Each audio clip is now represented by a collection of features: a 768-dim wav2vec vector, a 768-dim HuBERT vector, a 768-dim RoBERTa text vector, a 40-dim MFCC average, 12-dim chroma, and several scalar features (pitch, etc.). These were saved into a DataFrame and exported to CSV for reuse. The extracted features CSV will need to be uploaded at the start of part 2 in the Colab, as running the feature extraction for all 8,000+ records is very time-consuming. Part 1 of the Colab shows 1,000 records as a sample for feature extraction, but all 8,000+ records were extracted in another session and used in part 2 as described below.

In Part 2, we loaded this feature dataset and performed some cleaning: filling any missing values (e.g. if pitch was unvoiced for an entire clip, resulting in NaN) with column means (although this wasn't needed as no NaN features were present), and converting the stored list-form features back into numeric arrays. We then prepared input matrices for modeling. We also created fused feature sets by concatenating features: an "audio-only" vector combining wav2vec+hubert+all audio features (dimension 1596) and an "all features" vector combining audio and text (dimension 2364). These would be used for certain model variants described below.

3.1.5 Train/Test Split

Finally, we split the data into training and testing sets to evaluate model performance. We performed an 80/20 stratified split, ensuring each emotion class was proportionally represented in the training and test sets. After splitting, we applied features scaling independently on the training set and test to prevent any data leakage. This scaling step significantly improved model performance, particularly for large, high-capacity networks that previously suffered from overfitting when trained on raw, unscaled features. The training set was used to fit the models, and the held-out test set was used to evaluate the accuracy of unseen data. A grouped test-train split should have been used, and its repercussions are discussed in lessons learned.

3.2 Model Architectures

We experimented with six neural network models, each exploring a different strategy for multimodal feature fusion. All models were implemented in TensorFlow Keras. The emotion classification task was treated as a 7-way classification (one output neuron per emotion with softmax). Categorical cross-entropy was used as the loss, and we used the Adam optimizer for training. Unless otherwise noted, each Dense layer used ReLU activation and a small L2 weight regularization (0.0001) to mitigate overfitting. Below, we describe each model and its performance.

3.2.1 One-Branch Audio (Wav2Vec Baseline)

We built a simple multi-layer perceptron (MLP) that only inputs the Wav2Vec audio embedding. The network consists of an input layer of size 768 (wav2vec feature length), feeding into a hidden

Dense layer of 256 units, another Dense layer of 128 units, and finally, an output layer of 7 classes. This is essentially a baseline that uses audio cues alone (no text). Despite its simplicity, it achieved 97.88% accuracy on the test set after feature scaling, a dramatic improvement over the unscaled result (83.90%). This suggests that proper pre-processing alone greatly enhances classification performance by ensuring a more stable feature representation. We will compare other models against this baseline.

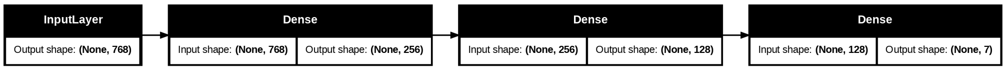

Figure 1: Wav2Vec Model

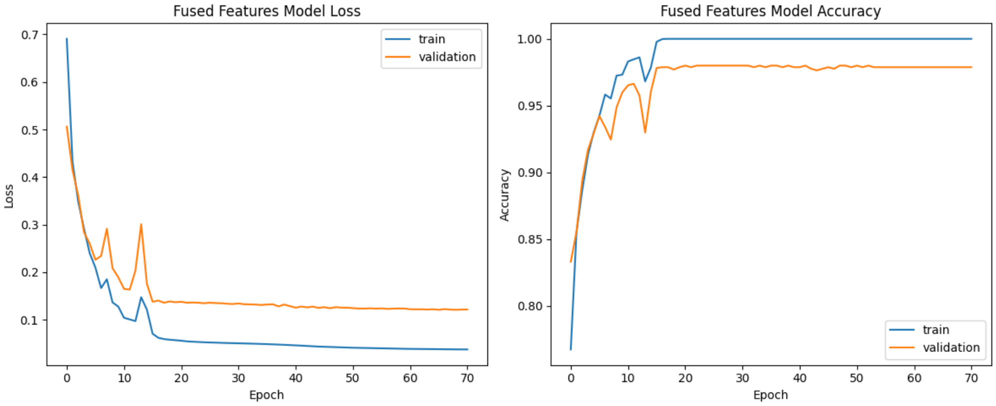

Figure 2: Wav2Vec Model Results

3.2.2 Two-Branch Early Fusion (Audio + Text)

In this approach, we introduced the text modality alongside audio. We used a two-branch architecture with separate sub-networks for audio and text features. The audio branch takes Wav2vec embeddings only and passes them through Dense layers of 256 → 128 units. In parallel, the text branch takes the 768-dim RoBERTa text embedding and passes it through Dense layers of 128 → 64 units. The outputs of these two branches (128-dim and 64-dim) are then concatenated and fed to a final softmax layer. This model reached 98.00% accuracy, showing a sharp improvement over the unscaled result (82.19%). However, the addition of text did not significantly boost performance over audio alone (97.88%), reinforcing that the text modality carried minimal useful emotion-specific information in this dataset. The result suggests that naively adding text to this dataset did not help, possibly because the textual content had little emotion-specific information given the neutral phrases selected for the actors.

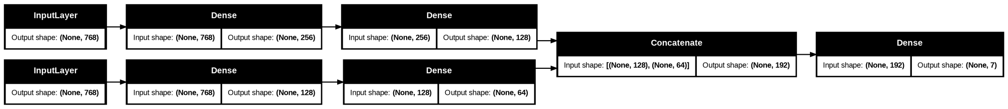

Figure 3: Wav2Vec + Text Model

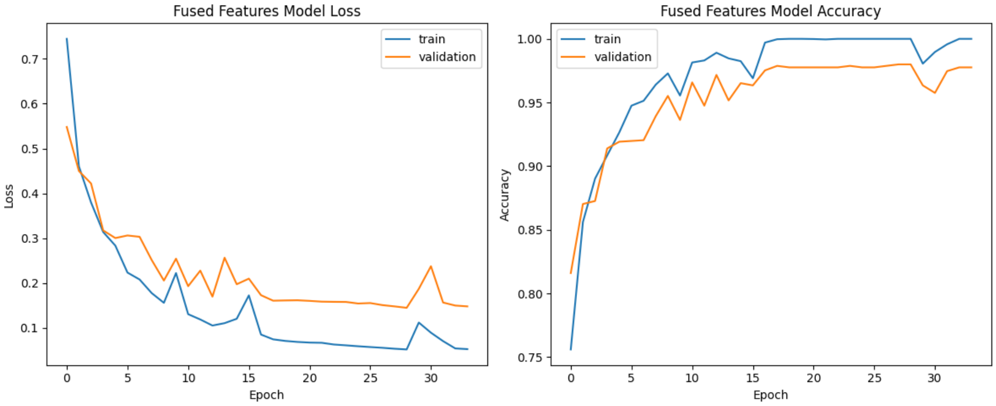

Figure 4: Wav2Vec + Text Model Results

3.2.3 Three-Branch Multimodal (Two Audio Embeddings + Text)

The third model explores a more granular audio fusion. Instead of lumping all audio features together, we created three branches: one for the Wav2Vec embedding, one for the HuBERT embedding, and one for text. The two audio branches each had Dense layers of 256 → 128 units, and the text branch again had 128 → 64 units. All three branch outputs were concatenated and fed into a softmax output. (This model did not explicitly include the traditional audio features like MFCCs; it focused on the two learned audio representations and text.) This architecture achieved 98.47% accuracy on the test set after scaling, improving significantly from the original 87.79% with unscaled features. The structured fusion of two deep audio embeddings (Wav2Vec and HuBERT) helped extract complementary emotional signals, but the text modality remained a relatively minor contributor.

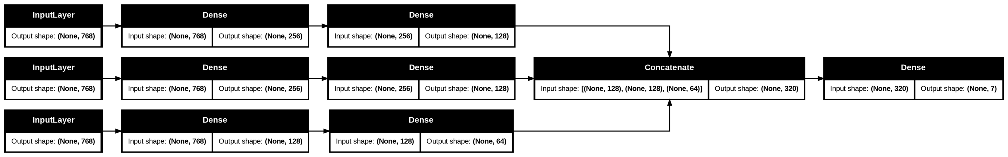

Figure 5: Wav2Vec + HuBERT + Text Model

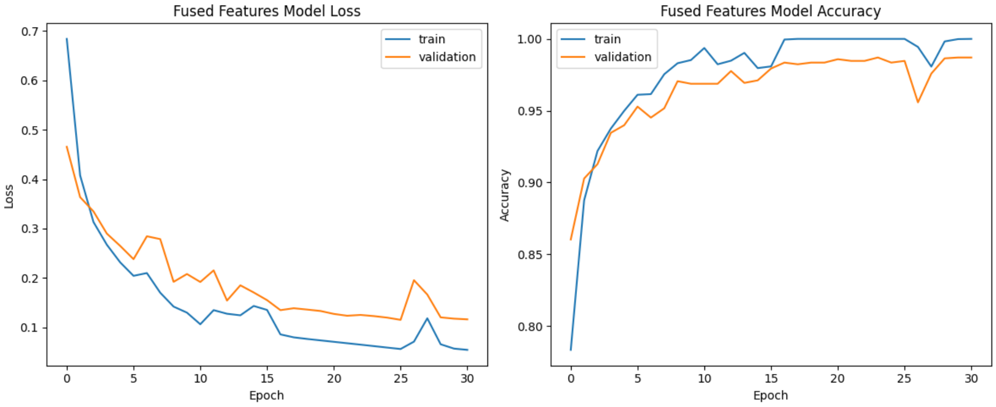

Figure 6: Wav2Vec + HuBERT + Text Model Results

3.2.4 Four-Branch Multimodal (Wav2Vec + HuBERT + Traditional Audio + Text)

This model incorporated all feature types with dedicated branches. We had four sub-networks: (a) Wav2Vec audio embedding branch (Dense 256 → 128), (b) HuBERT embedding branch (256 →

128), (c) traditional audio features branch, and (d) text branch. For the traditional audio branch, we fed the 60-dimensional vector of all hand-crafted features (MFCCs, chroma, pitch, etc. concatenated) and used a somewhat larger sub-network (Dense 512 → 256 → 128) to allow the model to learn a good representation. The text branch again was 128 → 64. We then concatenated all four branch outputs (each 128-D) into a combined feature and optionally passed it through one more fusion layer of 128 units to mix the modalities before the final output layer. This four-branch multimodal network achieved 98.11% accuracy after scaling. Interestingly, this was a decrease in accuracy from the three-branch model (98.47%), suggesting that the handcrafted audio features did not contribute much once the deep embeddings were scaled and effectively utilized.

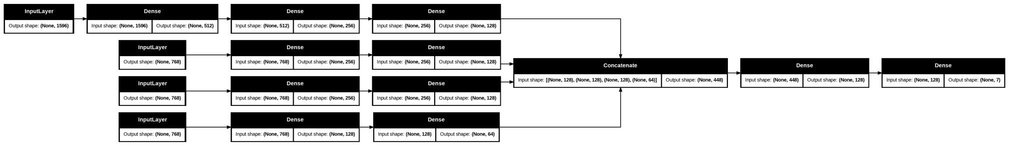
InputLayer
Output shape: (None, 1596)
Dense
Input shape: (None, 1596)
Output shape: (None, 512)
Dense
Input shape: (None, 512)
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 128)
InputLayer
Output shape: (None, 768)
Dense
Input shape: (None, 768)
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 128)
InputLayer
Output shape: (None, 768)
Dense
Input shape: (None, 768)
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 128)
InputLayer
Output shape: (None, 768)
Dense
Input shape: (None, 768)
Output shape: (None, 128)
Dense
Input shape: (None, 128)
Output shape: (None, 64)
Concatenate
Input shape: [(None, 128), (None, 128), (None, 128), (None, 64)]
Output shape: (None, 448)
Dense
Input shape: (None, 448)
Output shape: (None, 128)
Dense
Input shape: (None, 128)
Output shape: (None, 7)

Figure 7: Wav2Vec + HuBERT + Text + Traditional Audio Features Model

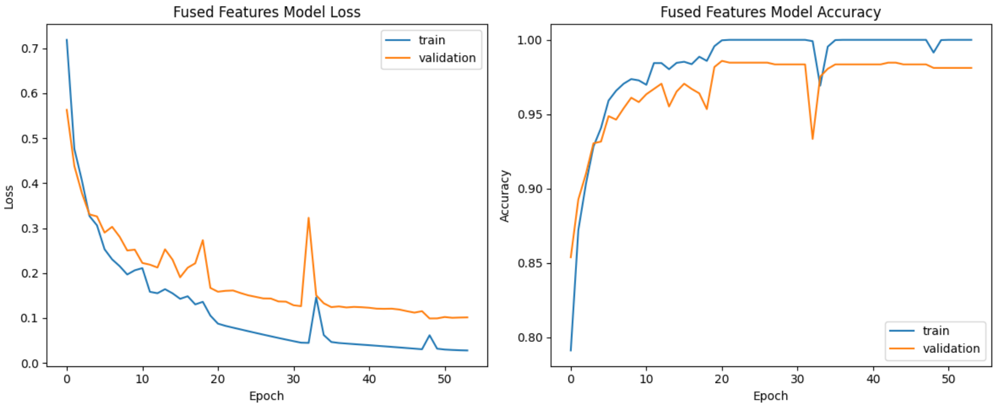

Figure 8: Wav2Vec + HuBERT + Text + Traditional Audio Features Model Results

3.2.5 Large Early-Fusion MLP (All Features Combined, No Dropout)

Next, we tested a single-branch model that directly takes the concatenation of all features (audio + text) as one large input vector (dimension 2364 in our case). We designed a high-capacity network to handle this: 3 hidden layers with 1024, 512, and 256 units respectively, each followed by batch normalization (to stabilize training). Notably, we did not use any dropout in this model, to test the limits of an un-regularized large network (only L2 weight decay was applied). Our expectation was that this model could theoretically learn the best possible fusion (since it has full access to all features at once), but it might also be prone to overfitting given the small training set and large number of parameters. Surprisingly, after feature scaling, this model achieved 98.35% test accuracy—a dramatic increase from the previous 65.51% when running the model with unscaled features. The main reason for this improvement was that scaling prevented certain features from dominating others, allowing the model to learn effectively from all input modalities without overfitting. This result overturns our previous assumption that early fusion inherently leads to poor generalization; instead, we now recognize that poor performance was due to the lack of proper feature normalization.

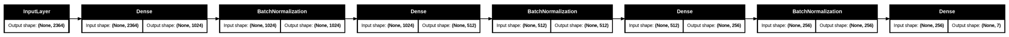
InputLayer
Output shape: (None, 2364)
Dense
Input shape: (None, 2364)
Output shape: (None, 1024)
BatchNormalization
Input shape: (None, 1024)
Output shape: (None, 1024)
Dense
Input shape: (None, 1024)
Output shape: (None, 512)
BatchNormalization
Input shape: (None, 512)
Output shape: (None, 512)
Dense
Input shape: (None, 512)
Output shape: (None, 256)
BatchNormalization
Input shape: (None, 256)
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 7)

Figure 9: Early Fused Model

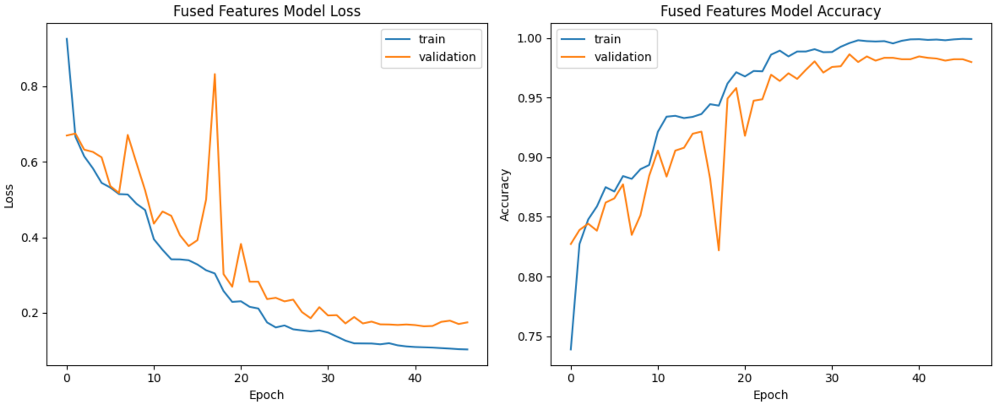

Figure 10: Early Fused Model Results

3.2.6 Regularized Large MLP (All Features + Dropout)

Finally, we introduced stronger regularization to the large early-fusion model. We kept the same architecture (1024 → 512 → 256 hidden units) but inserted dropout layers (20%) after each hidden layer, in addition to batch norm and L2 penalties. With feature scaling applied, this model reached 98.53% accuracy, making it the best-performing model. Interestingly, while dropout still provided a small performance boost from the prior model, its impact was much less pronounced than in the unscaled case, where it was essential to prevent overfitting. The revised results indicate that feature scaling alone addressed much of the overfitting previously observed, though dropout remains a useful regularization technique.

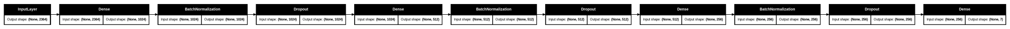
InputLayer
Output shape: (None, 2364)
Dense
Input shape: (None, 2364)
Output shape: (None, 1024)
BatchNormalization
Input shape: (None, 1024)
Output shape: (None, 1024)
Dropout
Input shape: (None, 1024)
Output shape: (None, 1024)
Dense
Input shape: (None, 1024)
Output shape: (None, 512)
BatchNormalization
Input shape: (None, 512)
Output shape: (None, 512)
Dropout
Input shape: (None, 512)
Output shape: (None, 512)
Dense
Input shape: (None, 512)
Output shape: (None, 256)
BatchNormalization
Input shape: (None, 256)
Output shape: (None, 256)
Dropout
Input shape: (None, 256)
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 7)

Figure 11: Early Fused Model w/ Dropout

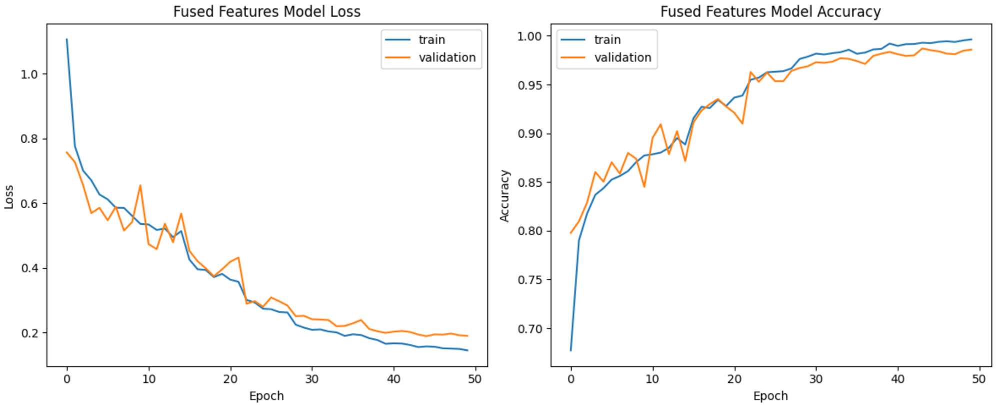

Figure 12: Early Fused Model w/ Dropout Results

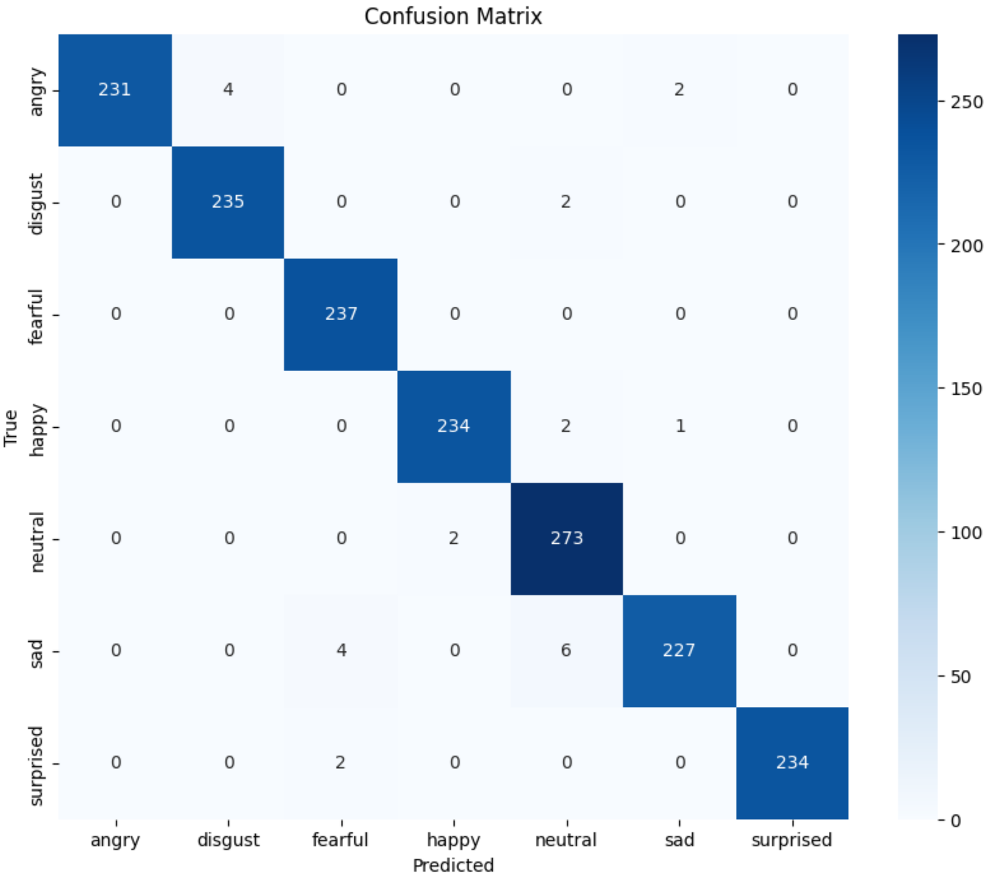

Figure 13: Early Fused Model w/ Dropout Confusion Matrix

3.3 Model Training Details

Model training details: We trained each model for up to 100 epochs with early stopping and learning rate reduction on plateau (patience 5 epochs) to prevent overtraining. The best epoch's weights on validation performance were used for final evaluation. All models used the same training set and were evaluated on the same test set for a fair comparison. The accuracy figures reported above are the final test accuracies for each model configuration.

4 Results

Model	**Test Accuracy (%) - Unscaled**	**Test Accuracy (%) - Scaled**
One-branch Audio (Baseline)	83.90%	97.88%
Two-branch Multimodal	82.19%	98.00%
Three-branch Multimodal	87.79%	98.47%
Four-branch Multimodal	91.27%	98.11%
Large Early-Fusion MLP (No Dropout)	65.51%	98.35%
Regularized Large MLP (With Dropout)	96.17%	98.53%

Table 1: Comparison of Model Performance on Unscaled and Scaled Features

Links:

- Colab with unscaled results
- Colab with scaled results

4.1 Multimodal vs Unimodal

We found that adding text features to audio (Model 2) did not yield a meaningful improvement over the audio-only model; however, unlike in the unscaled experiments, where text slightly hurt performance (82.19% vs. 83.90%), after feature scaling, the two models performed almost identically (98.00% vs. 97.88%). This suggests that while this text still does not contribute helpful emotion-specific information in this dataset, it no longer introduces noise or instability. The likely reason is that scaling allowed the model to process text features in a more balanced way rather than having them introduce unnecessary variability.

When we incorporated multiple audio feature types, performance steadily improved. Model 3, which combined two deep audio embeddings (Wav2Vec and HuBERT) with text, reached 98.47% accuracy, confirming that HuBERT provided additional helpful information beyond Wav2Vec. Model 4, which further added traditional handcrafted audio features (MFCCs, chroma, prosodic features, etc.), reached 98.11% accuracy, but interestingly, this was a minor improvement than in the unscaled models. This suggests that once features were scaled correctly, the deep embeddings alone captured the most relevant emotional information, leaving little room for traditional handcrafted features to add value.

These results confirm the dominance of the audio modality in emotion recognition—even after feature scaling, the text features remain largely redundant in this case, where the text selected for the actor to perform was neutral. However, using multiple deep audio embeddings proves beneficial, indicating that different pre-trained models capture complementary emotional cues.

4.2 Early vs. Late Fusion

In the unscaled experiments, early fusion models (which concatenate all features at once) performed significantly worse than structured late-fusion models, likely due to feature magnitude imbalances and overfitting. This was most evident in Model 5, where the unregularized early-fusion model collapsed to 65.51% accuracy, while the structured Model 4 (four-branch multimodal fusion) reached 91.27%.

However, the gap between early and late fusion completely disappeared after feature scaling. Model 5 (early fusion, no dropout) reached 98.35%, only slightly behind the best model (98.53%). Contrary to our previous conclusion that early fusion inherently leads to poor generalization, we now recognize that its previous failure was due primarily to unscaled inputs.

Once we applied proper normalization, even the large, unregularized early-fusion model generalized well, showing that scaling alone resolved much of the overfitting previously observed. Dropout was still beneficial (Model 6, which added dropout, performed slightly better at 98.53%), but it was no longer an absolute requirement for model stability. This result suggests that regularization remains useful but is less critical when inputs are well-preprocessed.

These findings fundamentally change our view on fusion strategy. Previously, we believed that structured, multi-branch fusion models were essential for good performance. However, we now see that when properly scaled and regularized, early fusion can perform just as well or better. This means that for future multimodal deep learning tasks, applying feature scaling should be prioritized before assuming that a structured fusion approach is necessary.

4.3 Evaluating the High Accuracy

With all models now achieving 98%+ accuracy, assessing whether these high results reflect true generalization or if dataset-specific factors are at play is crucial. The dataset consists of acted emotional speech, meaning emotions are exaggerated and likely easier to classify than in real-world speech. Additionally, the model may benefit from unintended cues in the dataset rather than general emotional recognition because the spoken phrases are neutral and repeated across different emotions.

One key risk is that identical text phrases appear in the training and test sets. This means the model might learn to associate certain sentences with specific emotions rather than relying purely on speech tone, pitch, and prosody. If the model can leverage textual similarities rather than acoustic features, its performance may not generalize well to real-world speech data, where phrasing varies more naturally.

To mitigate this, a grouped train-test split should be applied in future work, ensuring that the same spoken sentence never appears in both the training and test sets. This would force the model to classify emotions based purely on vocal tone and acoustic features rather than leveraging repeated text content.

*After testing the grouped train-test split on the audio-only baseline model, we found that ensuring identical text phrases did not appear in both the training and test sets resulted in a slight drop in accuracy compared to the random split. The test accuracy for the Wav2Vec-only model decreased from 97.88% to 93.89%, confirming that some text-based leakage had previously inflated results.

5 Lessons Learned

5.1 Value of Multimodal Features

Integrating audio and text modalities can boost performance, but the contribution of each modality may vary with context. In our project, the audio signals were the dominant factor in emotion detection (since the spoken words were neutral), and text features alone added little. However, using multiple audio feature types (pre-trained embeddings + handcrafted features) yielded substantial gains, showing that heterogeneous features capture complementary aspects of emotion. We learned that a rich feature representation is crucial for complex tasks like emotion recognition.

5.2 Fusion Strategy Matters

The impact of the fusion strategy changed significantly after feature scaling. Initially, structured late-fusion approaches (separate branches for each feature set) performed best, while naïve early fusion models failed due to overfitting. However, early fusion models matched or outperformed structured fusion after scaling, overturning our previous conclusions. The large early-fusion model (previously collapsed at 65.51%) improved to 98.35% after scaling, demonstrating that feature magnitude imbalances—not fusion structure—were the primary source of overfitting. This highlights an essential lesson: proper feature scaling can eliminate the need for complex fusion architectures.

A well-structured late-fusion model is still beneficial when working with heterogeneous data sources, but early fusion models can succeed if features are appropriately normalized. This finding suggests that preprocessing choices should be prioritized for future multimodal AI systems before considering complex fusion designs.

5.3 Regularization is Less Critical When Features Are Scaled

We initially observed extreme overfitting in the unregularized early-fusion model, which collapsed to 65% accuracy in the unscaled experiments. At that time, dropout was the key solution, improving accuracy to 96% by preventing overfitting. However, after feature scaling, the same unregularized model improved to 98.35%, proving that feature scaling is an implicit form of regularization. While dropout still provided a slight boost (final model: 98.53%), its impact was much less critical than before. This result underscores a new takeaway: scaling features properly can reduce the reliance on explicit regularization techniques.

Dropout, L2 penalties, and batch normalization remain helpful, but they should be considered secondary safeguards rather than first-line defenses. The priority should be ensuring that input data is well-conditioned before applying additional constraints.

5.4 Data Limitations and Challenges

Although our best model achieved 98.53% accuracy, the grouped train-test split experiment revealed some dataset bias in earlier results. When we ensured that identical text phrases did not appear in both training and test sets, the Wav2Vec audio-only model's accuracy dropped from 97.88% to 93.89%. This confirms that prior results were slightly inflated due to text-based leakage, though deep audio embeddings still provided strong emotion classification capabilities. Future work should rigorously test multimodal models under grouped train-test splits to ensure that performance gains are due to genuine emotional recognition rather than dataset artifacts.

Additionally, text features contributed little to classification accuracy, even in the grouped split scenario. This further reinforces that emotion detection in this dataset is primarily driven by vocal tone, pitch, and acoustic features rather than text content. This might change in a dataset where text carries stronger emotional cues (e.g., spontaneous speech), and future studies should explore multimodal setups in such contexts. On the practical side, feature extraction remains computationally expensive—running Whisper for transcriptions and large transformer models (RoBERTa, Wav2Vec, HuBERT) on thousands of audio clips requires substantial processing time. Caching extracted features, as we did by saving them to CSV, proved essential in enabling efficient experimentation. Exploring lighter feature extraction approaches or model distillation techniques could help scale multimodal emotion recognition to real-time applications.

5.5 Future Improvements

Going forward, we can experiment with more advanced fusion techniques. One idea is to use an attention-based mechanism to let the model dynamically learn which modality to focus on for each emotion rather than relying on simple feature concatenation. This could help mitigate cases where one modality dominates the prediction, ensuring a more balanced and adaptive integration of audio and text information.

Another improvement would be fine-tuning the pre-trained models (Wav2Vec2, HuBERT, RoBERTa) on the emotion classification task rather than using them solely for fixed feature extraction. Fine-tuning could allow them to capture more emotion-specific nuances in speech and text that are not fully leveraged in our current approach.

Our grouped train-test split experiment confirmed that text-based leakage slightly inflated earlier results, reinforcing the need for strict data partitioning in multimodal research. Future studies should ensure that identical spoken phrases do not appear in both the training and test sets, as even when text features are weak, they could introduce unintended dataset artifacts.

A more diverse and natural dataset should also be explored, ideally including spontaneous speech and conversational language, to better assess generalization beyond the controlled conditions of our current dataset. Additionally, cross-validation across multiple speaker groups could provide further insights into how well the model generalizes to unseen voices.

Another key area for improvement is data augmentation, particularly for audio. Augmenting the dataset by applying pitch shifting, time stretching, additive noise, or vocal tract length perturbation could help the model generalize better across different speakers, environments, and microphones. Since feature scaling is crucial in improving generalization, future work could explore adaptive scaling methods, where different feature groups (e.g., deep embeddings vs. handcrafted features) are scaled differently based on their statistical properties. Similarly, for text, paraphrasing techniques or back-translation (translating text to another language and back) could be explored to expand the dataset with more linguistically varied expressions, ensuring that any potential benefit from text features is maximized.

6 Conclusion

This study demonstrated that multimodal emotion recognition can achieve high accuracy, particularly when leveraging deep audio embeddings like Wav2Vec2 and HuBERT. While integrating multiple feature types improved performance, handcrafted audio features contributed little beyond what deep embeddings already captured, especially after applying feature scaling. Additionally, text features remained largely uninformative, reinforcing the need to carefully select datasets where textual content carries meaningful emotional cues. A key breakthrough was the role of feature scaling, which enabled early-fusion models to perform as well as, or better than, structured late-fusion models, eliminating prior overfitting issues and reducing reliance on heavy regularization techniques. This highlights that robust preprocessing—such as proper normalization—should be prioritized before introducing complex fusion architectures.

Another critical finding was the impact of train-test partitioning on model evaluation. When a grouped train-test split was applied to prevent text leakage, accuracy dropped slightly (97.88% to 93.89%), confirming that previous results were slightly inflated but still primarily driven by audio cues. This reinforces the importance of strict data partitioning in multimodal research to ensure that models genuinely learn emotional expression rather than memorizing text-emotion mappings. Future work should focus on testing models on more naturalistic datasets with spontaneous speech, fine-tuning pre-trained embeddings rather than using them as static feature extractors, and exploring adaptive scaling techniques to enhance generalization across different speakers and environments. These insights lay a strong foundation for designing more robust and scalable multimodal emotion recognition models for real-world applications.

References

[1] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *arXiv preprint* (2020). arXiv: `2006.11477v3`.

[2] Yung-Sung Chuang et al. "SpeechBERT: An Audio-and-text Jointly Learned Language Model for End-to-end Spoken Question Answering". In: *arXiv preprint* (2020). arXiv: `1910.11559v4`.

[3] Jonas Ewertz et al. "Listen Closely: Measuring Vocal Tone in Corporate Disclosures". In: *SSRN* (2024).

[4] Zilong Huang, Man-Wai Mak, and Kong Aik Lee. "MM-NodeFormer: Node Transformer Multimodal Fusion for Emotion Recognition in Conversation". In: *Interspeech*. 2024.

[5] Hailun Lian et al. "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face". In: *Entropy* 25.1440 (2023). DOI: `10.3390/e25101440`.

[6] Varun Sharma. "Speech and Text-Based Emotion Recognizer". In: *arXiv preprint* (2023). arXiv: `2312.11503v1`.