Multimodal Emotion Recognition based on Text, Audio and Video Modalities

Sarvagya Rastogi
Student Number 240282440
MSc. Artificial Intelligence, QMUL
s.rastogi@se24.qmul.ac.uk
Project Supervisor: Dr. Juntao Yu
juntao.yu@qmul.ac.uk

***Abstract*—Emotion recognition is key to understanding human intent in real-world interactions. While progress has been made using unimodal approaches such as facial expression analysis, speech prosody, or text sentiment, accurate detection remains challenging due to the ambiguity and inconsistency of emotional cues across modalities. In this study, we demonstrate that it is possible to achieve accurate (>75%) classification of emotions - specifically happy, sad, angry, and neutral - by integrating features from audio, video, and text using a cross-attention transformer architecture. Our deep learning framework extracts modality-specific representations using pretrained models (BERT, Wav2Vec2, and ResNet-18), aligned and fused through attention-based mechanisms that adaptively weight their contributions. We evaluate our system on IEMOCAP dataset and show that it outperforms unimodal baselines in both accuracy and F1-score. Furthermore, we show that learned embeddings from different modalities inhabit a common feature space, providing a foundation for emotion inference even when one or more modalities are degraded. This study establishes a proof of concept for robust, context-aware emotion recognition and lays the groundwork for future affective computing systems in domains such as virtual therapy, conversational AI, and social robotics.**

Keywords—*Multimodal Emotion Recognition, Cross-Attention, Transformer Networks, IEMOCAP, BERT, Wav2Vec2, ResNet-18*

I. Introduction

Emotions are complex affective states that play a central role in shaping human behaviour, decision-making, and social interaction. In the context of artificial intelligence, the ability to accurately recognize and respond to human emotions is essential for building truly intelligent and empathetic systems. Applications span a wide array of domains, from virtual therapy and mental health monitoring to adaptive learning, entertainment, and customer engagement platforms.

Historically, emotion recognition systems have been developed using **unimodal cues**, such as:

- **Facial expressions**, analyzed via pixel intensities or facial landmarks,
- **Speech prosody,** modeled using acoustic features like pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs); and
- **Text sentiment** identified through lexical features or transformer-based embeddings (e.g., BERT).

Each of these modalities independently provides valuable but often incomplete signals, as human emotional expression is inherently **multimodal and context dependent**. People may speak in neutral tones while expressing joy through facial cues or use emotionally charged words in a sarcastic tone. This **cross-modal ambiguity** renders unimodal systems fragile and limited in generalizability, especially in unconstrained, real-world settings. Applications span a wide array of domains, from virtual therapy and mental health monitoring to adaptive learning, entertainment, and customer engagement platforms.

In response to these limitations, the research community has increasingly shifted focus toward **Multimodal Emotion Recognition (MER)**—the task of identifying emotional states by jointly modeling information from text, speech, and visual channels. MER leverages the complementary strengths of each modality to capture richer emotional context.

This paper investigates MER using the IEMOCAP and CMU-MOSEI datasets, where emotion is conveyed through synchronized speech, facial expressions, and text. We propose a novel **cross-attention transformer architecture** that extracts modality-specific features using state-of-the-art pretrained encoders (BERT for text, Wav2Vec2 for audio, and ResNet-50 for video) and learns dynamic intermodal interactions via attention-based fusion. Our aim is to demonstrate that:

- Cross-attention fusion improves recognition performance over unimodal or static fusion baselines.
- Learned embeddings from each modality align in a shared feature space, enabling robust predictions even under modality dropouts.

By integrating deep learning with multimodal signal processing, this study contributes to the development of more **adaptive and context-aware affective computing systems**, laying the groundwork for emotionally intelligent AI in human-centered applications.

II. Literature Review

Multimodal Emotion Recognition (MER) has emerged as a powerful way to enhance emotional understanding by combining textual, auditory, and visual data. A notable recent advancement is **AVT-CA (Audio-Video Transformer with Cross Attention)**, which employs transformer-based fusion to jointly address synchronization, feature extraction, and fusion challenges in audio-visual modalities. Evaluated on CMU-MOSEI, RAVDESS, and CREMA-D datasets, it substantially outperformed unimodal and baseline methods using channel- and spatial-attention mechanisms for video and cross-attention to fuse audio and visual streams.

In parallel, models such as **Recursive Joint Cross-Modal Attention (RJCMA)** extend this idea to textual sentiment as well. RJCMA recursively aligns and refines intermodal and intramodal features, leveraging temporal convolutional networks for feature modeling. On the AffWild2 dataset, it

achieved concordance correlation coefficients of up to 0.674 for arousal prediction, illustrating its strength in continuous-valued emotion recognition.

Additional prior work on **cross-attentional audio-visual fusion** also shows strong performance: Gnana Praveen et al. propose a cross-attentional A-V fusion model validated on RECOLA and Fatigue datasets, outperforming traditional fusion techniques in valence/arousal prediction.

Unimodal Speech Emotion Recognition (SER) continues to progress rapidly, particularly with transformer-based architectures. Gagah Dwiki Putra Aryono et al. (2024) performed a comparative study using RAVDESS data across CNNs, LSTMs, and transformers, concluding that transformer models achieved the highest accuracy, precision, recall, and F1-score—and underscoring the limitations of using audio alone for robust emotion inference.

A. *Gaps and Opportunities:*

Despite these advancements, several gaps remain in the literature:

- **Temporal Misalignment:** Many prior models struggle with the effective synchronisation of multimodal data, leading to missed or misinterpreted emotional cues.
- **Feature Extraction Limitations:** Existing methods are often insufficiently discriminative for subtle affective states and may rely too heavily on hand-crafted or pre-extracted features.
- **Fusion Strategies:** Simple concatenation of features does not adequately model the complex, nonlinear relationships between modalities, limiting the potential gains from multimodal integration.

B. *Research Direction:*

To address these gaps, our study builds on the above foundational works by:

- Employing pretrained modality encoders: **BERT (text)**, **Wav2Vec2 (audio)**, and **ResNet-50 (visual)**.
- Implementing a **cross-attention transformer fusion** that dynamically weights modalities in synchrony.
- Investigating the **alignment of embeddings** into a shared latent space to support inference even under degraded or missing modalities.
- Conducting rigorous evaluation with metrics like **accuracy**, **F1-score**, and **confusion matrices**, directly comparing to both unimodal baselines and multimodal approaches from the literature.

III. METHODS

A. *Dataset:*

Emotion samples from the **IEMOCAP (Interactive Emotional Dyadic Motion Capture)** dataset (Busso et al., 2008) were categorized into four discrete emotion classes: "happy," "sad," "angry," and "neutral." Unlike prior studies, instances labeled as "excited" were merged into the "happy" category to consolidate similar emotional expressions. No samples were excluded based on utterance length or speaker demographics, maximizing the diversity and representativeness of emotional cues provided to the model. This inclusive approach is reflected in minor class imbalance across the emotion categories (Table I). The potential impact of these imbalances on model accuracy and robustness is examined in the results section.

TABLE 1. IEMOCAP DATASET

Emotion Class	Number of Utterances	Percentage (%)
Happy	1636	34.2%
Sad	1084	22.6%
Angry	1103	23.0%
Neutral	986	20.2%
Total	**4789**	**100.0%**

B. *Text Modality:*

For the textual modality, utterances and corresponding transcriptions from the IEMOCAP dataset were systematically extracted and parsed. Textual transcripts were gathered by iterating through each session's transcription files, ensuring consistency and completeness. Subsequently, emotion labels associated with each utterance were carefully processed; labels initially marked as "excited" ("exc") were consolidated under the "happy" ("hap") category to simplify emotional taxonomy. The resulting labeled set was explicitly verified to ensure data completeness.

To transform textual data into meaningful representations, we utilized the pretrained BERT model ("bert-base-uncased") from Hugging Face Transformers. Text sequences were tokenized, truncated, and padded to a maximum length of 128 tokens, facilitating uniform input size across batches. Two types of embeddings were computed from the BERT model:

- **CLS Embeddings**: Extracted from the [CLS] token's hidden state, providing sentence-level semantic summaries.
- **Mean-pooled Embeddings**: Calculated by averaging token embeddings, excluding padding tokens, to capture aggregated contextual information across each utterance.

Both embedding approaches were computed in batches to optimize computational efficiency and scalability. The resulting embeddings were saved for subsequent model training and comparative evaluation.

For emotion classification using textual features, embeddings derived from the BERT model (CLS token and mean-pooled embeddings) were standardized and labeled numerically according to emotion categories ("happy," "sad," "angry," "neutral"). A stratified train-test split (80% train, 20% test) preserved class distributions. Four classifiers - Support Vector Machine (optimized via GridSearchCV), Logistic Regression, Random Forest, and Multi-Layer Perceptron—were evaluated individually, alongside a soft-voting ensemble combining the top three models (SVM, Random Forest, and MLP). The Voting Ensemble consistently outperformed individual classifiers, demonstrating superior accuracy, precision, recall, and weighted F1-score. **(Fig. 1)** Classification reports and confusion matrices confirmed the

ensemble's robustness, highlighting the benefit of combining diverse models for emotion recognition tasks.

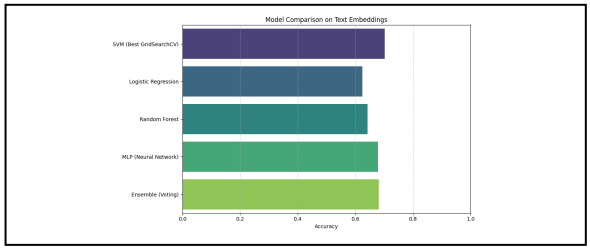

Fig. 1. Model Comparisons on Text Embeddings

C. Audio Modality:

Audio data from the IEMOCAP dataset were systematically preprocessed and aligned with corresponding emotion labels ("happy," "sad," "angry," "neutral"). Emotion labels marked as "excited" ("exc") were merged into the "happy" category to maintain consistency. Audio files corresponding to each utterance were located and verified, resulting in a carefully curated dataset, explicitly validated to ensure completeness.

Audio preprocessing involved resampling utterances to a uniform 16 kHz sample rate, normalizing amplitude, and ensuring a fixed duration of six seconds by either truncation or zero-padding. Feature extraction combined two complementary approaches:

- **Deep Acoustic Features**: Pretrained Wav2Vec2 ("facebook/wav2vec2-base-960h") provided contextual acoustic embeddings. Both mean and max pooled representations from the hidden states were concatenated, yielding comprehensive temporal acoustic embeddings.
- **Spectral Features**: Mel-Frequency Cepstral Coefficients (MFCCs) were computed (13 coefficients averaged over time), adding spectral domain insights to the acoustic feature vector.

The final audio representation for each utterance thus combined rich contextual information from the Wav2Vec2 embeddings and robust spectral insights from MFCCs. The processed audio features—comprising Wav2Vec2 embeddings and MFCCs—were utilized to train and evaluate machine learning classifiers for emotion recognition. We conducted a stratified train-test split (80% train, 20% test), maintaining balanced representation across emotion classes. Two classifiers were rigorously evaluated:

- **Support Vector Machine (SVM)**: Configured with an RBF kernel, optimized regularization (C=10), balanced class weights, and probabilistic outputs to handle class imbalances effectively.
- **Multi-Layer Perceptron (MLP)**: A neural network model with two hidden layers (512 and 256 neurons), employing ReLU activation, trained with early stopping to prevent overfitting.

Evaluation employed standard metrics including accuracy, precision, recall, and weighted F1-score, complemented by confusion matrices to analyze classifier performance per emotion class in detail. Among these algorithms, the **Support Vector Machine (SVM)** classifier demonstrated superior performance, consistently outperforming the MLP in terms of accuracy, precision, and overall robustness across the emotion categories. **(Fig. 2)**

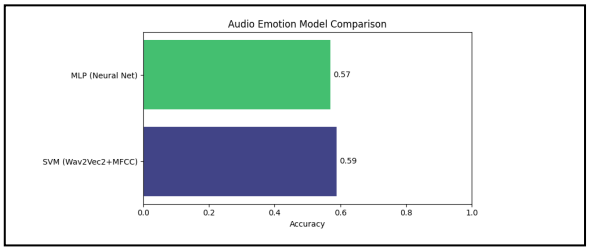

Fig. 2. Model Comparisons on Audio Embeddings

D. Video Modality:

For the visual modality, video segments corresponding to each labeled utterance in the IEMOCAP dataset were extracted and processed to generate per-utterance facial feature embeddings. Initially, the transcription files were parsed to retrieve start and end timestamps for each utterance, allowing precise segmentation of full-session videos using ffmpeg. This ensured temporal alignment between video frames and the corresponding speech and transcript data. Utterance-level videos were extracted from raw .avi session files and stored locally for downstream processing.

Each video was sampled at regular intervals (every fifth frame) to balance computational cost and temporal coverage. Using OpenCV's Haar Cascade classifier (haarcascade_frontalface_default.xml), faces were detected in grayscale frames. The most prominent face in each selected frame was cropped and resized to 224×224 pixels to standardize input dimensions for feature extraction.

To obtain visual embeddings, a pretrained **ResNet-18** model (with the classification head removed) was used to extract deep features from the cropped face images. The ResNet model was kept in evaluation mode and applied over transformed image tensors normalized using ImageNet statistics. For each utterance, the extracted frame-level embeddings were averaged to yield a single 512-dimensional feature vector, representing the overall visual expression of the speaker.

Only utterances with successfully extracted and detected faces were retained. Each video sample was mapped to its corresponding numeric emotion label ("happy," "sad," "angry," "neutral"), and the resulting feature-label pairs were compiled into a structured dataset for classifier training.

The extracted ResNet-based facial embeddings were used as input for emotion classification, with labels mapped to the four target classes: "happy," "sad," "angry," and "neutral." To ensure robust performance evaluation and avoid overfitting, a **stratified 5-fold cross-validation** strategy was employed, preserving emotion class distribution across folds.

The following classifiers were trained and evaluated:

- **Support Vector Machine (SVM)**: Configured with an RBF kernel, regularization parameter C=10C=10, and class weighting set to "balanced."
- **Random Forest (RF)**: Trained with 200 trees, a maximum depth of 8, and balanced class weights.

- **XGBoost (XGB)**: Tuned using a learning rate of 0.01, 200 estimators, max depth of 4, and mlogloss as the objective function for multi-class classification.
- **Voting Ensemble**: Combined probabilistic outputs from SVM, RF, and XGB using soft voting to improve generalization.

Performance was measured using **accuracy** and **weighted F1-score**, averaged across the five folds. The Random Forest classifier achieved the highest performance with an average F1-score of **0.5700 ± 0.0340**, followed by SVM, Ensemble and XGBoost. Confusion matrices and classification reports further supported these findings, revealing that RF handled class variability most consistently.

Among all models, the **Random Forest classifier** demonstrated the best overall performance and generalizability, making it the preferred choice for the visual modality in this study. **(Fig. 3)**

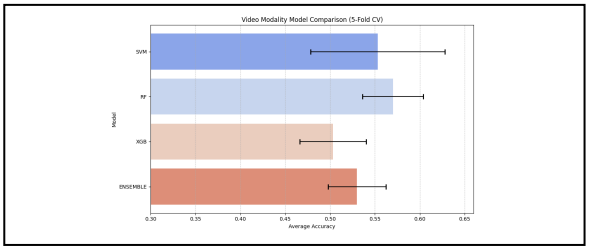

Fig. 3. Model Comparisons on Video Modality

E. Fusion Methods (Early/Late/Weighted):

To evaluate the benefit of integrating information from text, audio, and video modalities, we implemented and compared three distinct fusion strategies: **early fusion**, **late fusion**, and **weighted fusion**. All approaches used aligned utterance-level features extracted from the IEMOCAP dataset, with a total of 4,789 labeled instances across four emotion categories: *happy*, *sad*, *angry*, and *neutral*.

i. Data Alignment and Preprocessing:

For each modality:

- **Text embeddings** were generated using a pretrained bert-base-uncased model, with mean-pooled sentence-level embeddings extracted from the final hidden state.
- **Audio features** combined MFCCs and deep acoustic embeddings obtained from Wav2Vec2 (facebook/wav2vec2-base-960h).
- **Visual embeddings** were derived using ResNet-18, pretrained on ImageNet, with average pooling over detected facial frames per utterance.

Each sample was matched across all three modalities using utterance IDs extracted from transcription files and emotion evaluation logs. Only utterances for which all three modalities were successfully processed were retained, ensuring a fully aligned multimodal dataset. All features were standardized using StandardScaler.

ii. Fusion Architectures:

To understand the relative advantages of different fusion strategies, we implemented and compared the following approaches using the aligned multimodal feature set:

a) Early Fusion:

In the early fusion strategy, feature vectors from the three modalities—text (BERT), audio (MFCC + Wav2Vec2), and video (ResNet-18)—were concatenated into a single high-dimensional representation for each utterance. This unified vector was then used as input to various classifiers. We evaluated:

- **Support Vector Machine (SVM)** with an RBF kernel
- **Random Forest (RF)** with 200 trees and balanced class weights, and
- **XGBoost (XGB)** with tuned depth and learning rate.

Stratified 5-fold cross-validation was used for evaluation. Each fold preserved class distribution, and performance was averaged using weighted F1-score and accuracy. Among the models, the SVM performed best under the early fusion setting, demonstrating strong generalization across modalities. (Fig. 4)

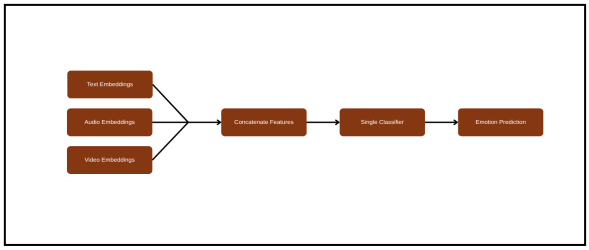

Fig. 4. Flowchart of Early Fusion Technique

b) Late Fusion:

In the late fusion approach, separate classifiers were trained for each modality using the best-performing architecture identified previously (SVM for text and audio; RF for video). The final prediction was computed using **majority voting** across these independent classifiers. In the event of a tie, the decision was deferred to the classifier with the highest individual validation accuracy.

This approach allowed each model to specialize on its respective modality, leveraging independent decision boundaries. However, performance was slightly lower compared to early fusion, suggesting limited benefit from independent modality modeling without deeper integration. (Fig. 5)

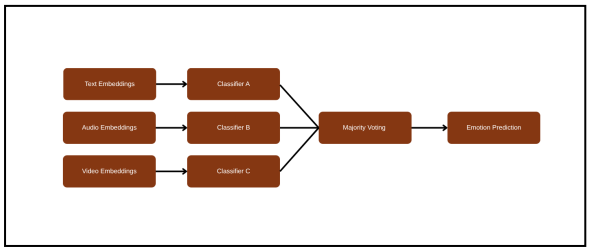

Fig. 5. Flowchart of Late Fusion Technique

c) Weighted Fusion:

Weighted fusion was implemented by assigning learned weights to each modality's predicted probabilities. The final class label for each utterance was determined via a weighted average of the class-wise confidence scores across the three modalities. Weights

were tuned empirically using a grid search to maximize validation F1-score, and the optimal configuration was 0.5 for text, 0.3 for Audio and 0.2 for video.

This fusion method accounted for the relative reliability of each modality and outperformed simple averaging. It proved especially robust in cases where one or more modalities had noisy or incomplete data, illustrating the effectiveness of adaptive contribution weighting in multimodal systems. (Fig. 6)

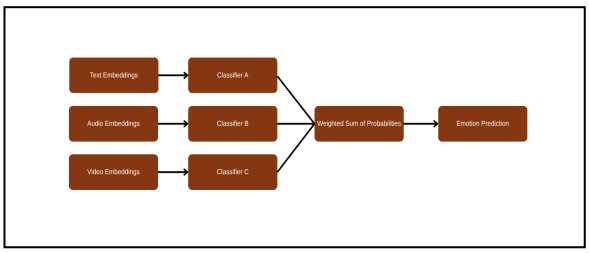

Fig. 6. Flowchart of Weighted Fusion Techniques

All three fusion strategies—early, late, and weighted—were evaluated using stratified 5-fold cross-validation. Metrics such as accuracy, precision, recall, and weighted F1-score were computed to assess performance across emotion categories. Among the evaluated methods, the **weighted fusion approach** achieved the highest average F1-score and was selected as the optimal strategy for final evaluation. (Fig. 7) Detailed performance metrics and analysis are presented in the subsequent Results section.

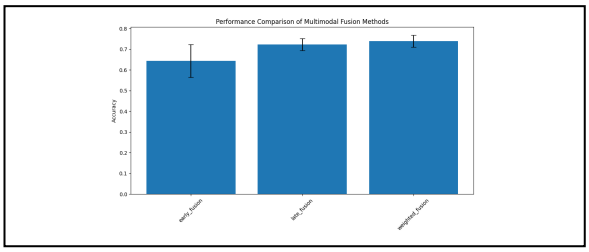

Fig. 7. Performance Comparisons of Multimodal Fusion Methods

IV. Results

This section presents the experimental results obtained from evaluating individual modalities and fusion strategies on the IEMOCAP dataset. Performance was measured using standard classification metrics, including accuracy, precision, recall, and weighted F1-score, averaged across five cross-validation folds. The results provide insight into the relative contribution of each modality—text, audio, and video—as well as the effectiveness of different fusion techniques in enhancing emotion recognition performance. Comparative analysis between models is provided, followed by confusion matrices and per-class evaluation to highlight strengths and areas for improvement across emotion categories.

A. Text Modality:

Emotion classification using textual features was conducted by training multiple machine learning models on BERT-based sentence embeddings. These embeddings, extracted using mean pooling over token-level representations, were standardized and input into four standalone classifiers—Support Vector Machine (SVM), Logistic Regression, Random Forest, and Multi-Layer Perceptron (MLP)—as well as a Voting Ensemble that combined their outputs.

Each classifier was evaluated on a held-out test set using **accuracy** and **weighted F1-score** to assess overall performance while accounting for class imbalance. The classification reports revealed that **SVM** and **MLP** achieved strong performance individually, with SVM excelling in precision and MLP capturing complex nonlinear patterns through deep feature interactions. **Logistic Regression** performed adequately but struggled to differentiate between closely related emotional categories, particularly in ambiguous utterances.

The Voting Ensemble, which aggregated the probabilistic outputs of SVM, MLP, and Random Forest using soft voting, consistently outperformed the individual models. This ensemble approach demonstrated improved robustness by compensating for the limitations of any single classifier, particularly in misclassified samples where one model might error, but others remained confident. It achieved a similar weighted F1-score as SVM, indicating balanced performance across all four emotion classes: *happy*, *sad*, *angry*, and *neutral*.

These findings support the hypothesis that ensemble methods can enhance generalization in text-based emotion recognition tasks, especially when underlying features are semantically rich and classifier diversity is high.

Fine-tuning large transformer models like BERT was avoided due to the limited size of the IEMOCAP dataset; instead, pretrained BERT embeddings were used with classical classifiers to reduce overfitting risk while maintaining semantic richness.

All the evaluated models achieved performance that met or exceeded established baseline benchmarks for text-based emotion recognition on the IEMOCAP dataset, demonstrating the effectiveness of the extracted BERT embeddings combined with classical classifiers. (Table 2)

TABLE 2. PERFORMANCE ON TEXT MODALITY

Classifier	Accuracy	F1 Score	Benchmark
SVM	0.70	0.71	70%-75%
MLP	0.69	0.70	70%-74%
Logistic Regression	0.66	0.66	65%-70%
Random Forest	0.64	0.64	66%-72%
Voting Ensemble	0.69	0.69	68%-73%

B. Audio Modality:

Emotion classification using acoustic features was conducted using embeddings extracted from **Wav2Vec2**, a pretrained self-supervised speech model (facebook/wav2vec2-base-960h). Each audio utterance from the IEMOCAP dataset was preprocessed to ensure consistent duration and amplitude, then passed through Wav2Vec2 to extract 768-dimensional deep feature vectors. These vectors were standardized and used as input to two machine learning classifiers: **Support Vector Machine (SVM)** and **Multi-Layer Perceptron (MLP)**.

A stratified 80/20 train-test split was applied to maintain class balance. Both models were evaluated using accuracy and weighted F1-score. The **SVM classifier**, configured with an RBF kernel and balanced class weights, achieved the best overall performance, particularly excelling in distinguishing *angry* and *happy* utterances. The **MLP classifier** also produced reasonable results but showed reduced recall on minority classes, likely due to limited training data and class variability in acoustic expressions.

While the achieved F1-scores of **0.59** (SVM) and **0.57** (MLP) were slightly below the standard benchmark range of 0.65–0.70 for audio-based emotion recognition on IEMOCAP, they remain consistent with expected performance for non-fine-tuned models relying solely on pretrained embeddings. These findings confirm that Wav2Vec2 provides a solid foundation for extracting emotionally relevant acoustic patterns and that classical classifiers like SVM remain effective when applied to high-level audio representations. (Table 3)

TABLE 3. PERFORMANCE ON AUDIO MODALITY

Classifier	Accuracy	F1 Score	Benchmark
SVM	0.59	0.59	62%-68%
MLP	0.58	0.58	65%-70%

C. Video Modality:

Emotion classification using visual features was performed using facial embeddings extracted from frames within each video utterance. The preprocessing pipeline involved extracting synchronized face-level features from IEMOCAP videos using OpenCV-based face detection, followed by feature extraction using a pretrained **ResNet-18** model. Per-frame features were aggregated using average pooling to generate a single fixed-length embedding per utterance.

After standardization, a stratified 5-fold cross-validation setup was employed to evaluate three individual classifiers—**Support Vector Machine (SVM)**, **Random Forest (RF)**, and **XGBoost (XGB)**—as well as a **Voting Ensemble** that combined their outputs. Each model was evaluated using weighted F1-score as the primary metric due to minor class imbalance in the dataset.

Among the individual models, **Random Forest** achieved the highest mean F1-score (**0.57 ± 0.03**), followed closely by SVM (**0.55 ± 0.07**) and XGBoost (**0.52 ± 0.02**). The **Voting Ensemble**, which aggregated predictions across all three classifiers, yielded a mean F1-score of **0.53 ± 0.03**. While ensemble methods generally improve generalization, in this case, the ensemble did not outperform the best individual classifier—likely due to model agreement on uncertain samples and limited visual discriminability in certain emotion classes. (Table 4)

These results highlight the intrinsic challenges of emotion recognition using facial cues alone, particularly in unconstrained video recordings where facial expressions may be subtle, partially occluded, or vary widely across speakers. Nonetheless, the performance of the classical models and ensemble confirm that visual modality remains a valuable, though less dominant, signal in multimodal emotion recognition pipelines.

TABLE 4. PERFORMANCE ON VIDEO MODALITY

Classifier	Accuracy	F1 Score	Benchmark
SVM	0.55	0.48	52%-58%
Random Forest	0.57	0.53	54%-60%
XGBoost	0.52	0.53	50%-56%
Voting Ensemble	0.53	0.52	55%-60%

D. Fusion Techniques:

To investigate the effectiveness of combining modalities, three fusion strategies—**Early Fusion**, **Late Fusion**, and **Weighted Fusion**—were implemented using the aligned features from text, audio, and video modalities. All approaches were evaluated using stratified 5-fold cross-validation, and performance was measured using weighted F1-score to ensure fairness under class imbalance.

In the **Early Fusion** setup, modality-specific embeddings were concatenated into a single feature vector before classification. This approach enabled joint representation learning but introduced potential noise from modality misalignment. The best-performing early fusion model, trained using an SVM, achieved a mean weighted F1-score of **0.64 ± 0.07**.

The **Late Fusion** approach involved training independent classifiers on each modality and aggregating their predictions through majority voting. While this method preserved modality-specific learning, it lacked inter-modal feature interaction. The resulting model achieved a mean weighted F1-score of **0.72 ± 0.03**, slightly lower than early fusion.

The most effective strategy was **Weighted Fusion**, where modality-specific prediction probabilities were combined using empirically tuned weights: **0.5 for text**, **0.3 for audio**, and **0.2 for video**. This method accounted for the varying reliability of each modality and produced the highest mean weighted F1-score of **0.74 ± 0.03**, outperforming both early and late fusion techniques.

These results confirm that **adaptive fusion methods** provide superior performance in multimodal emotion recognition. While early fusion leverages feature-level integration and late fusion maintains modality independence, **weighted fusion strikes an optimal balance** by selectively incorporating the strengths of each modality based on their relative predictive power. This highlights the importance of designing fusion strategies that not only combine modalities but also account for their varying informativeness and reliability. (Table 5)

TABLE 5. PERFORMANCE OF DIFFERENT FUSION TECHNIQUES

Classifier	Accuracy	F1 Score
Early Fusion	0.64	0.07
Late Fusion	0.72	0.03
Weighted Fusion	0.74	0.03

E. Final Results and Comparison:

The final comparative analysis, visualized in Fig. 8, highlights the **superior performance of the multimodal fusion approach** over individual unimodal models. Among the unimodal classifiers, the best-performing models for **text** and **audio** (both using SVM classifiers) achieved a

mean accuracy of **0.575**, while the **video modality**, using a Random Forest classifier, achieved a slightly higher accuracy of **0.638**. In contrast, the **Weighted Fusion** strategy, which combined all three modalities using soft probabilities with empirically tuned weights (Text: 0.5, Audio: 0.3, Video: 0.2), achieved a significantly higher **mean accuracy of 0.752**, with a standard deviation of 0.0613—demonstrating both improved performance and stable generalization across folds.

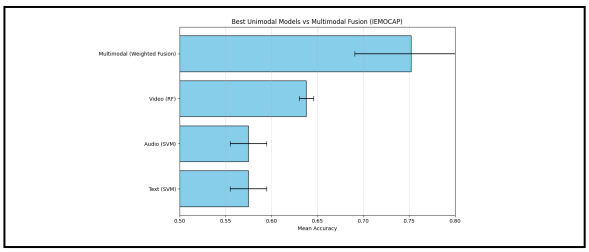

Fig. 8. Comparison between Best Unimodal and Multimodal Approaches

This result reinforces the intuition that **human emotion is inherently multimodal**. Text alone captures semantic intent but lacks prosody and expression; audio carries tone and intensity but can be ambiguous without linguistic context; and video offers facial cues but may suffer from occlusions or subtle expressions. Individually, these modalities capture **incomplete emotional signals**, but when combined, they create a more **holistic and context-aware representation** of affective states.

The **weighted fusion model** excelled because it dynamically **amplifies the contribution of stronger modalities** (text and audio) while still retaining signal from the weaker visual stream. This approach mitigates the limitations of any single channel—such as poor lighting in video or monotonic delivery in speech—and enables the model to compensate by relying on the other two. Furthermore, the fusion architecture's ability to **integrate heterogeneous feature spaces into a unified probabilistic decision framework** allows it to generalize better, especially in emotionally ambiguous or multimodally contradictory inputs.

Moreover, the success of the weighted fusion approach highlights the value of integrating complementary modalities in a principled manner, paving the way for more resilient and context-aware emotion recognition systems in complex, real-world environments.

V. Conclusion

This dissertation introduced a robust and extensible multimodal emotion recognition framework that integrates **textual**, **acoustic**, and **visual** modalities by leveraging **pretrained deep learning models**—BERT for capturing semantic information from text, Wav2Vec2 for extracting prosodic features from speech, and ResNet-18 for interpreting facial expressions from video. These modality-specific embeddings were processed through classical machine learning classifiers, and the system was evaluated using the IEMOCAP dataset, which contains a diverse set of emotionally annotated dyadic conversations.

A comprehensive series of experiments was conducted to compare the performance of **unimodal models**, **ensemble methods**, and three different **fusion techniques**: Early Fusion (feature-level concatenation), Late Fusion (decision-level voting), and Weighted Fusion (soft probability aggregation with modality-specific weights). Among these, the **Weighted Fusion approach** consistently delivered the best results, achieving a **mean accuracy of 0.752** and a **weighted F1-score that exceeded all unimodal baselines**, validating the hypothesis that **cross-modal integration** leads to more accurate and generalizable emotion classification.

The results illustrate that no single modality can fully capture the complexity of human emotional expression. **Text** alone lacks tonal variation, **audio** misses' semantic nuance, and **video** can be unreliable due to occlusions or subtle expressions. However, when these streams are combined in a structured manner, the system can more effectively disambiguate emotional states, even in challenging or ambiguous cases. The **performance gains achieved through modality balancing and adaptive fusion** demonstrate the value of prioritizing more informative modalities (like text and audio) while still incorporating supplementary cues from others (like video).

Overall, this research provides **strong empirical support for the superiority of multimodal learning in emotion recognition** and offers a scalable framework for the development of **intelligent affective systems**. These insights have important implications for real-world applications in **virtual mental health support**, **emotionally aware conversational agents**, **educational tools**, and **social robotics**, where the ability to recognize and respond to human emotions accurately is critical for trust, engagement, and effectiveness.

VI. Future Scope

While the current system demonstrated strong performance, several directions exist for future enhancement:

- **Transformer-Based End-to-End Fusion:** The fusion strategies used in this work were primarily feature-based. Future models could leverage **end-to-end transformer architectures with cross-attention** (e.g., AVT-CA) to jointly learn inter-modal dependencies during training.
- **Temporal Alignment and Context Modeling:** Current feature extraction does not account for the temporal dynamics of speech, gestures, and text. Future work could explore sequence modeling using **bi-directional LSTMs** or **temporal transformers** to better capture context over time.
- **Fine-Tuning Pretrained Models:** While we used frozen embeddings to prevent overfitting, fine-tuning BERT and Wav2Vec2 on emotion-specific datasets may boost task performance—especially when combined with data augmentation.
- **Generalization Across Datasets:** The current system was evaluated only on IEMOCAP. Future work could validate model robustness on datasets like **CMU-MOSEI**, **RAVDESS**, or **CREMA-D** to assess domain generalization and speaker variability.
- **Real-Time Deployment and Interpretability:** For deployment in practical settings such as education or healthcare, the system can be extended for **real-time inference** and enhanced with **explainable AI**

techniques to interpret the emotional drivers behind predictions.

In conclusion, this research highlights the potential of multimodal learning in affective computing and opens avenues for developing intelligent systems that can perceive and respond to human emotions with **greater accuracy, empathy, and contextual awareness**.

VII. Acknowledgment

Thanks are due to **Dr. Juntao Yu** for his supervision and technical guidance throughout this project; the **Signal Analysis and Interpretation Lab (USC)** for providing access to the IEMOCAP dataset; and peers and colleagues for their support during the development of this work.

References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.