TERT-Ensemble: A Tri-modal Emotion Recognition Technique

A PROJECT REPORT

Submitted by

PENUMUCHU NIHITH [RA2111026010124]

V LINGESHWARAN [RA2111026010128]

Under the Guidance of

Dr. NAVNEET NAYAN

Assistant Professor, Department of Computational Intelligence

in partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING

with specialization in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR - 603 203

Department of Computational Intelligence

SRM Institute of Science & Technology

Own Work* Declaration Form

To be completed by the student for all assessments

Degree/ Course	**: B. Tech CSE in Specification with AI & ML**
Student Name	**: Penumuchu Nihith, V. Lingeshwaran**
Registration Number	**: RA2111026010124, RA2111026010128**
Title of Work	**: TERT Ensemble: A Tri-Modal Emotion Recognition Technique**

We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism, as listed in the University Website, Regulations, and the Education Committee guidelines.

We confirm that all the work contained in this assessment is our own except where indicated, and that We have met the following conditions:

- Clearly referenced / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that We have received from others (e.g.fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook /University website

We understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:

We are aware of and understand the University's policy on Academic misconduct and plagiarism and We certify that this assessment is our own work, except were indicated by referring, and that We have followed the good academic practices noted above.

P. Nihith

Penumuchu Nihith
RA2111026010124

V Lingeshwaran
RA2111026010128

If you are working in a group, please write your registration numbers and sign with the date forevery student in your group.

SRM

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that 18CSP109L Major Project report titled **"TERT-Ensemble: A Tri-modal Emotion Recognition Technique"** is the Bonafide work of **"PENUMUCHU NIHITH (RA2111026010124), V. LINGESHWARAN (RA2111026010128)"** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported here in does not from any other project or report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any candidate.

Navneet Nayan
SIGNATURE

Dr. NAVNEET NAYAN
SUPERVISOR
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

SIGNATURE

Dr. R. ANNIE UTHRA
PROFESSOR & HEAD
DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

12/5/2025
EXAMINER I

EXAMINER II

ACKNOWLEDGEMENTS

We express our humble gratitude to **Dr. C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to **Dr. Leenus Jesu Martin M,** Dean-CET, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank **Dr. Revathi Venkataraman**, Professor and Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We encompass our sincere thanks to, **Dr. M. Pushpalatha**, Professor and Associate Chairperson - CS, School of Computing and **Dr. Lakshmi,** Professor and Associate Chairperson -AI, School of Computing, SRM Institute of Science and Technology, for their invaluable support.

We are incredibly grateful to our Head of the Department, **Dr. R. Annie Uthra**, Professor, Department of Computational Intelligence, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our Project Coordinators, Panel Head, and Panel Members Department of Computational Intelligence, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor, **Dr. S. Amudha**, Department of Computational Intelligence, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide, **Dr. Navneet Nayan**, Department of Computational Intelligence, SRM Institute of Science and Technology, for providing us with an opportunity to pursue our project under his mentorship. He provided us with the freedom and support to explore the research topics of our interest. His passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank all the staff members of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank our parents, family members, and friends for their unconditional love, constant support and encouragement.

Authors

Penumuchu Nihith
V. Lingeshwaran

ABSTRACT

Emotion recognition is a crucial task in human-computer interaction, affective computing, and psychological analysis. While unimodal approaches using image, text, or audio show promise, they often fail to capture the full complexity of human emotional expression. Multimodal approaches offer richer understanding by integrating information from different sources. This report details the development and evaluation of 'TERT-Ensemble', a tri-modal emotion recognition system employing a two-stage fusion process. Our approach first focuses on building robust unimodal expert predictors by evaluating diverse architectures (CNNs, ViT, Transformers, LSTM) within each modality (image, text, audio) on combined benchmark datasets. To optimize unimodal performance, intra-modal ensembles were created by combining predictions from these base models using weighted averaging, achieving test set weighted F1-scores of 77.98% for the Image Ensemble, 73.69% for the Text Ensemble, and 66.43% for the Audio Ensemble. Subsequently, a tri-modal late fusion model was implemented, combining the outputs of these three intra-modal ensembles using a weighted averaging strategy based on their respective F1-scores. Evaluated on a simulated tri-modal test set (1195 samples), this final hybrid model achieved a high accuracy of 95.56% and a weighted F1-score of 0.96. This report describes the data preprocessing, model architectures, training strategies, and the two-stage fusion methodology implemented within Kaggle notebooks, along with an interactive Gradio demonstration of the unimodal components. While acknowledging the limitations of simulated test data for the final fusion evaluation, the results validate the effectiveness of the intra-modal ensembles and highlight the significant potential of the proposed tri-modal fusion technique.

Keywords: Multi-modal Emotion Recognition, Vision Transformer (ViT), BERT, DeBERTa, EfficientNet.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AI	Artificial Intelligence
AUC	Area Under Curve
BERT	Bidirectional Encoder Representations from Transformers
CK+	Extended Cohn-Kanade Dataset
CNN	Convolutional Neural Network
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
DeBERTa	Decoding-enhanced BERT with Disentangled Attention
ER	Emotion Recognition
FER	Facial Expression Recognition
FER-2013	Facial Expression Recognition 2013 Dataset
FN	False Negative
FP	False Positive
HCI	Human-Computer Interaction
HOP_LENGTH	Hop Length (Audio Processing Parameter)
LFS	Large File Storage
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
N_FFT	Number of Fast Fourier Transform points
N_MELS	Number of Mel filterbanks
NLP	Natural Language Processing
PWA	Progressive Web App
RAF-DB	Real-world Affective Faces Database
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
ReLU	Rectified Linear Unit

CHAPTER 1
INTRODUCTION

1.1 Introduction to the Project

The ability of computers to recognize human emotions automatically stands as a significant and ongoing endeavor within computer science. This capability has far-reaching implications, poised to transform fields such as mental health monitoring, the intuitiveness of human-computer interaction (HCI), the safety features in driver assistance systems, and the personalization of content recommendation engines [1]. Humans inherently employ a rich tapestry of cues to convey and interpret emotions, drawing simultaneously from facial expressions, the nuances of vocal tone and prosody, and the explicit or implicit sentiment carried by language [2]. Historically, computational efforts to decode these emotions often adopted a focused approach, analyzing these cues in isolation. This specialization led to distinct research domains: Facial Expression Recognition (FER), which scrutinizes visual data from images [3]; sentiment and emotion classification, which delves into the content of text [4]; and Speech Emotion Recognition (SER), which analyzes the acoustic properties of audio signals [5].

However, relying solely on a single modality inevitably curtails the scope and inherent robustness of emotion recognition systems. A facial expression, for instance, can be deliberately controlled or be inherently ambiguous. Textual communication, by its nature, lacks the rich paralanguage of non-verbal cues such as tone and inflection. Similarly, the prosodic characteristics of speech may not always align perfectly with an individual's genuine emotional state, especially when navigating the complexities of real-world interactions that can involve background noise or social nuances like sarcasm [6]. Consequently, integrating information from multiple modalities specifically vision, text, and audio presents a more promising avenue. Such an approach allows for a more comprehensive, context-aware, and ultimately more accurate understanding of human emotion, striving to mirror the sophisticated, integrated perceptual capabilities that humans naturally possess [6]. The pursuit of effective multimodal emotion recognition is therefore central to advancing the field of affective computing and to fostering more natural, intuitive, and empathetic interactions between humans and technological systems.

1.2 Motivation

The primary motivation for this research originates from the pressing need to develop automated emotion recognition systems that are not only more accurate but also more robust and comprehensive in their ability to handle the multifaceted nature of real-world human expression. As previously highlighted, systems confined to a single modality often grapple with inherent ambiguity and a lack of holistic context. For example, a smile captured in an image might mask underlying sadness that could be revealed through verbal expression (audio or text), or conversely, neutral language in a text might be delivered with a strongly emotional vocal tone. The challenges are further amplified in everyday scenarios which can involve noisy data, partially occluded faces, diverse speaking styles, or subtle linguistic nuances, all of which expose the limitations of relying on a singular information channel.

This project, named 'TERT-Ensemble', endeavors to address these challenges by proposing a tri-modal approach that leverages the combined information from image, text, and audio data. A core tenet of our methodology is the initial establishment of strong, reliable unimodal 'expert' models for each modality before any attempt at cross-modal fusion. We achieve this by systematically evaluating a diverse range of state-of-the-art deep learning architectures tailored to each data type. Specifically, this includes Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for image analysis, language Transformers such as BERT and DeBERTa for text processing, and sequence models like Long Short-Term Memory networks (LSTMs) for understanding audio cues [7]–[10]. This foundational unimodal analysis, which forms the core of this report, serves as an essential prerequisite for the subsequent development of effective fusion strategies designed to intelligently synthesize the strengths of each distinct modality. Ultimately, a robust multimodal emotion recognition system holds significant potential to catalyze advancements across various application domains, including enhanced HCI, more insightful mental health assessment and monitoring tools, improved patient care systems, adaptive educational technologies, and innovative accessibility tools for individuals facing communication challenges.

1.3 Sustainable Development Goal of the Project

The research undertaken within the 'TERT-Ensemble' project directly aligns with and aims to contribute towards several pivotal United Nations Sustainable Development Goals (SDGs). This contribution is primarily achieved through the development of foundational enabling technologies that enhance our capacity to understand human emotional states automatically and reliably.

The most direct and significant alignment is with **SDG 3: Good Health and Well-being**. This goal strongly emphasizes the importance of ensuring healthy lives and promoting well-being for all individuals across all age groups. Accurate and reliable automated emotion recognition systems, such as the one envisioned by TERT-Ensemble with workflow as shown in the Figure 1, serve as a critical technological component for future applications within the mental health domain. By establishing strong unimodal performance baselines and exploring effective fusion, this work helps pave the way for systems that could potentially: (1) facilitate *Enhanced Mental Health Monitoring* by providing objective data points to aid clinicians in their assessments or enable passive monitoring of well-being; (2) contribute to *Early Detection* by enabling tools that might assist in the timely identification of emotional distress or subtle shifts indicative of developing mental health conditions [11]; and (3) lead to *Improved HCI in Healthcare*, fostering more empathetic telehealth platforms or assistive robotic companions better attuned to patient emotional states.

Secondly, this project makes a contribution to **SDG 9: Industry, Innovation, and Infrastructure**. This goal focuses on building resilient infrastructure, promoting inclusive and sustainable industrialization, and fostering innovation. Our work contributes by: (1) driving *Technological Innovation* through the development and evaluation of novel deep learning techniques for multimodal emotion analysis, thereby advancing the fields of AI and affective computing [12]; and (2) contributing to *Infrastructure Development* by providing validated methodologies and trained models that add to the growing infrastructure of AI tools applicable to the complex task of understanding human behavior.

Indirectly, the capacity to reliably infer emotional states also holds implications for **SDG 4: Quality Education**. Emotionally aware systems could potentially form the basis for next-generation adaptive learning technologies. Such systems could tailor educational content or teaching strategies in real-time based on detected student engagement, confusion, or frustration, potentially leading to more personalized, responsive, and ultimately more effective learning experiences [13].

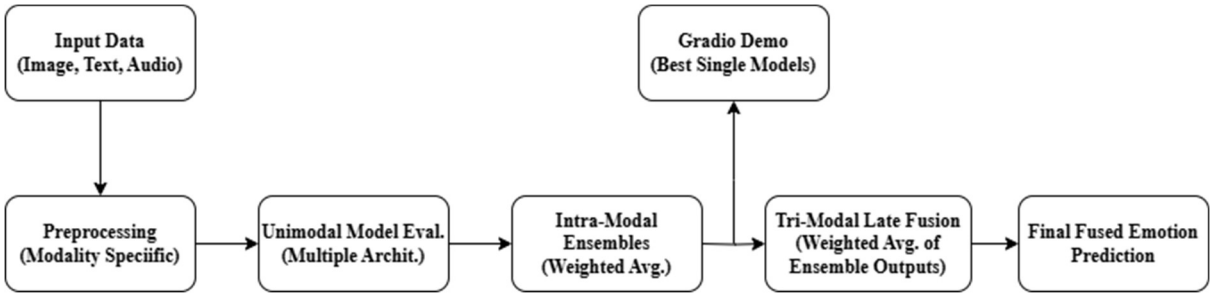


Figure 1: TERT-Ensemble project workflow

CHAPTER 2
LITERATURE SURVEY

2.1 Research - Reference Papers - Summary

The field of automated emotion recognition has seen substantial advancements within individual modalities, and increasingly, research is exploring the synergistic potential of multimodal approaches. Our TERT-Ensemble project builds upon this existing body of work. This section reviews key literature pertinent to the unimodal architectural choices, intra-modal ensembling strategies, and the final tri-modal fusion techniques employed. Table 1 provides a summary of several pivotal studies that inform our methodology, highlighting techniques and findings relevant to image, text, audio, and multimodal emotion recognition.

2.1.1 Facial Expression Recognition (FER):

The domain of Facial Expression Recognition has evolved significantly. Initial methodologies often depended on hand-crafted visual features like Local Binary Patterns (LBP) or Histograms of Oriented Gradients (HOG) to describe facial changes [14]. However, as Li and Li survey [3], the advent of deep learning has shifted the paradigm. Convolutional Neural Networks (CNNs), with various architectures such as VGG, ResNet, and the more recent EfficientNet, have become the standard due to their inherent ability to automatically learn complex, hierarchical feature representations directly from pixel data, eliminating the need for manual feature engineering. Vision Transformers (ViT) [8] represent a newer approach, treating image patches as sequences and utilizing self-attention mechanisms to capture global contextual information from faces. To enhance performance, especially when dealing with the variability present in "in-the-wild" datasets (which include diverse poses, illuminations, and occlusions), ensemble techniques are commonly adopted. Ranjan et al. [15], for example, through work on systems like HyperFace, demonstrated that combining predictions or features from diverse network backbones can lead to more robust and generalizable FER systems. Their findings often underscore the benefits of architectural diversity within an ensemble to better tackle the variability of real-world facial images. Our image modality analysis similarly explores a range of these architectures.

2.1.2 Text Emotion Recognition:

For recognizing emotions from text, early deep learning applications frequently utilized Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [16] to model the sequential nature of language. More recently, the landscape has been dominated by Transformer-based models, which leverage massive pre-training on extensive text corpora.

Prominent architectures like BERT (Bidirectional Encoder Representations from Transformers) [9], RoBERTa (A Robustly Optimized BERT Pretraining Approach) [17], and DeBERTa (Decoding-enhanced BERT with Disentangled Attention) [18] have set new benchmarks in various natural language understanding tasks, including emotion recognition.

These models learn rich, contextualized word and sentence embeddings. Typically, these pre-trained models are then fine-tuned on specific emotion-annotated datasets. Studies such as the work by Acheampong et al. [19] have further shown that ensembling these powerful Transformer models can yield additional performance improvements. Their approach often involves fine-tuning multiple instances of similar Transformer models (e.g., BERT or RoBERTa, possibly with minor variations) and then averaging their outputs, suggesting that even with highly capable base models, an ensemble can enhance F1-scores and overall accuracy, especially for nuanced tasks like identifying sarcasm alongside primary emotions.

2.1.3 Speech Emotion Recognition (SER):

In the domain of Speech Emotion Recognition (SER), a typical pipeline involves extracting relevant acoustic features from the raw audio signal, a topic comprehensively reviewed by Schuller [5]. Widely used features include Mel-Frequency Cepstral Coefficients (MFCCs), which capture information about the spectral envelope, and full Mel spectrograms [20], which provide a two-dimensional time-frequency representation of the speech signal. These extracted features are then fed into various machine learning classifiers. While traditional models like Support Vector Machines (SVMs) or Hidden Markov Models (HMMs) were historically employed, deep learning techniques have gained prominence. CNNs are often used for recognizing patterns in spectrograms, and LSTMs are adept at modeling the temporal dynamics inherent in speech [21]. Hybrid architectures that combine these strengths, such as the CNN-LSTM model proposed by Trigeorgis et al. [22], have proven effective. Such models often use CNN layers to learn local spectro-temporal features from spectrogram inputs, which are then passed to LSTM layers to model the sequential progression of these features over time, thereby capturing both local acoustic cues and longer-term temporal dependencies critical for emotional expression in speech. As observed in other modalities, ensembling diverse SER models, such as combining different CNN architectures or CNNs with RNNs as highlighted by Li et al. [23] in their work on SER-CNNs, frequently leads to significant improvements in accuracy over individual model performance. This often involves strategies like performance-weighted averaging of the individual model predictions.

2.1.4 Multimodal Emotion Recognition and Fusion:

While unimodal systems have demonstrated considerable advancements, multimodal emotion recognition which integrates cues from various sources like vision, text, and audio generally offers superior performance and a more holistic understanding of an individual's emotional state [6], [24]. The central challenge in multimodal ER lies in effectively designing the fusion strategy, i.e., how information from these different modalities is intelligently combined. Broadly, fusion techniques can be categorized as:

- **Early Fusion (Feature-Level):** This approach involves concatenating the feature vectors extracted independently from each modality at an early stage, before feeding them into a single, unified classification model [25]. While this method allows for the model to learn correlations between low-level features from different modalities, it can lead to very high-dimensional and potentially heterogeneous feature spaces. It also necessitates careful synchronization and balancing of the different feature types, which can be problematic given their often disparate nature (e.g., image features vs. text embeddings vs. audio features).
- **Late Fusion (Decision-Level):** In this strategy, each modality is processed by an independent, specialized model. The final predictions or probability scores from these individual unimodal models are then combined to make an overall multimodal decision. Common combination methods include averaging, weighted sum based on model confidence or reliability, or majority voting. Poria et al. [24], in their comprehensive review of affective computing, highlight that late fusion, particularly when combining the outputs of strong, independently trained unimodal models, remains a robust, viable, and often highly competitive strategy. Its advantages include simplicity of implementation and greater resilience to situations where one or more modalities might be missing or unreliable.
- **Hybrid Fusion:** These methods aim to amalgamate aspects of both early and late fusion. For instance, some features might be fused early, while others are processed unimodally, with their outputs combined at a later stage.
- **Intermediate Fusion (Deep Fusion):** More sophisticated techniques perform fusion at intermediate layers within deep neural network architectures. These can range from simple concatenation of intermediate representations to complex methods employing cross-modal attention mechanisms or dedicated multimodal fusion transformers, such as the Multimodal Transformer (MulT) proposed by Tsai et al. [26]. MulT, for example, utilizes directional pairwise cross-modal attention to enable information flow between unaligned multimodal sequences, allowing one modality to dynamically influence the representation of another at

a fine-grained level. While powerful, such approaches typically introduce significant architectural and computational complexity.

The success of any advanced fusion strategy often hinges on the quality and discriminative power of the unimodal representations being fused. This observation underscores the critical importance of developing robust and optimized unimodal components first, which is a primary focus of the initial stages of our TERT-Ensemble project, before progressing to the implementation of more intricate fusion architectures. Our project evaluates intra-modal ensembles to achieve strong unimodal predictors, followed by a late fusion of these ensemble outputs for the final tri-modal prediction.

2.2 Limitations Identified from Literature Survey

Despite significant progress, the literature highlights several recurring limitations in emotion recognition research, which motivated the foundational unimodal analysis approach taken in this project:

- **Data Constraints:** A major challenge across modalities is the availability of large-scale, diverse datasets with reliable annotations, especially those capturing spontaneous, naturalistic emotions rather than acted portrayals [24], [26]. Many widely used benchmarks rely on acted data, which may not fully generalize to real-world scenarios [22], [23]. Furthermore, models trained on one dataset often exhibit poor cross-corpus generalization when tested on another [15].
- **Fusion Complexity:** While multimodal fusion is theoretically advantageous, designing optimal fusion mechanisms remains difficult. Feature-level fusion can be sensitive to missing data and requires careful alignment, while decision-level fusion might discard valuable cross-modal information. Advanced fusion models often introduce significant complexity [24], [26].
- **Computational Cost:** State-of-the-art deep learning models, particularly large language transformers and complex ensemble systems, require substantial computational power and time for training and inference, potentially limiting their accessibility and practical deployment [15], [19], [23], [26].
- **Inherent Ambiguity and Subjectivity:** Emotion expression is inherently complex and can be subjective. Text can contain sarcasm or irony [19], acoustic cues can be heavily influenced by speaker characteristics unrelated to emotion [23], and facial expressions can be deliberately masked or subtly displayed [15]. Annotator disagreement further reflects this ambiguity.

- **Inter-Class Similarity:** Distinguishing between closely related emotional states (e.g., fear vs. surprise, sadness vs. neutral) is often challenging due to overlapping features in expression across modalities [15], [22]. This is particularly true for subtle or low-intensity emotions.

Table 1: Reference Research Papers

Study Lead Author(s) & Ref Key	Primary Focus	Modalities / Data	Key Techniques / Architectures	Dataset(s) Example(s)	Key Finding(s) Relevant to TERT-Ensemble
[4], [24]	Multimodal Sentiment/Emotion Review	Audio, Visual, Text	CNNs, LSTMs, Acoustic Feats, Feature/Decision Fusion	MOUD / Various	Multimodal generally outperforms unimodal. Late fusion is viable. Data scarcity is a challenge.
[23]	Ensemble for SER	Audio	Ensemble of CNNs, RNNs; Weighted Averaging	IEMOCAP	Ensembling significantly improves SER accuracy. Performance-weighted averaging is effective.
[15]	Ensemble for FER (in the wild)	Image	Ensemble of diverse CNNs (ResNet, VGG variants)	AffectNet, RAF-DB	Ensembles enhance FER performance/generalization on real-world data. Architectural diversity is beneficial.
[19]	Transformer Ensemble (Text Emotion)	Text	Ensemble of fine-tuned BERT, RoBERTa; Averaging	Text Emotion Benchmarks	Ensembling improves F1-scores even with similar base Transformer models. Sarcasm is a challenge.
[26]	Multimodal Transformer Fusion	Language, Visual, Audio	Cross-Modal Attention Transformer	MOSEI, MOSI	Advanced fusion (cross-modal attention) can model inter-modal dynamics but increases complexity.
[22]	Hybrid CNN-LSTM (Unimodal SER)	Audio, Visual (separate)	CNN-LSTM architecture per modality	RECOLA	Hybrid CNN-LSTM effectively captures local and temporal information within a single modality stream.
[5]	SER Review	Audio	Overview of features, models, benchmarks, challenges	Various	Summarizes two decades of SER, highlighting common features (MFCCs) and models (SVM, HMM, DL).
[3]	FER Review	Image	Overview of traditional and deep learning methods for FER	Various	Provides context on the evolution from handcrafted features to CNNs/ViTs in facial expression analysis.

2.3 Novelty and Advantages of the Proposed Model

The TERT-Ensemble project, through its phased approach, introduces several novel aspects and advantages:

- **Systematic Unimodal Architectural Evaluation:** We undertake a systematic training and comparison of multiple diverse, state-of-the-art deep learning architectures (CNNs, ViT, Transformers, LSTMs) specifically tailored for each distinct modality (image, text, audio), rather than pre-selecting a single architecture per modality.
- **Intra-Modal Ensembling for Robust Unimodal Prediction:** A key novelty is the creation of optimized intra-modal ensembles. Before any cross-modal integration, we fuse the predictions from the different base models *within* each modality using weighted averaging. This step aims to produce a more robust and reliable "expert" prediction for each individual modality, leveraging the diverse strengths of the base architectures.
- **Two-Stage Late Fusion Strategy:** The overall fusion process is staged. First, intra-modal ensembles yield refined unimodal probability distributions. Second, these ensemble outputs are combined using a tri-modal late fusion technique (weighted averaging based on the empirically determined performance of each intra-modal ensemble). This hierarchical approach allows for focused optimization at each stage.
- **Benchmarking on Combined Datasets:** By aggregating multiple standard benchmark datasets for each modality, we aim for a more comprehensive evaluation of model generalization across varied data characteristics (e.g., posed vs. spontaneous expressions, different linguistic styles, diverse acoustic conditions) compared to studies relying on single, more homogeneous datasets.
- **Reproducible Methodology and Implementation:** All experiments were conducted within the Kaggle notebook environment using publicly available datasets and well-established open-source libraries, promoting transparency and facilitating reproducibility of the training procedures and results.
- **Interactive Demonstration of Unimodal Components:** The best-performing *single* models identified during the initial architectural evaluation were integrated into a practical Gradio demonstration interface, providing a tangible showcase of the unimodal prediction capabilities achieved.
- **Modularity and Scalability:** The staged design (unimodal base models -> intra-modal ensembles -> tri-modal fusion) creates a modular framework. This allows for potential replacements of individual components without necessarily retraining the entire system.

CHAPTER 3
EXECUTION METHODOLOGY

This chapter outlines the step-by-step approach we took to build and test our TERT-Ensemble system for recognizing emotions from images, text, and sound. Our method involved carefully preparing the data, training several different computer 'brains' (models) for each type of information, combining the smartest of these brains for each type, and then finally mixing the results from the image, text, and sound experts to get one overall emotion prediction.

3.1 Process Overview

Our project followed a structured plan to effectively analyze emotions using three different types of information:

1. **Gathering the Data:** We started by collecting various well-known public datasets that contained examples of images, text, and audio, each labeled with different emotions.
2. **Preparing the Data:** Each type of data needed special preparation.

- **Images:** We made sure all pictures were of a standard size and format so our computer models could understand them. We also adjusted them to be more like real-world photos.
- **Text:** We cleaned up the written text by removing things like website links, @mentions, and hashtags. Then, we broke the text down into smaller pieces (tokens) that our language models could work with. We also tried to make sure we had a similar number of examples for each emotion.
- **Audio:** We converted all sound files to a standard quality and length. Then, we extracted important sound features (like Mel spectrograms and MFCCs) that help describe the emotional tone.

3. **Splitting the Data:** We divided our prepared data for each modality into three piles: one for teaching the models (training set), one for checking how well they were learning during training (validation set), and one final pile for testing how good they really were (test set).
4. **Choosing and Building Basic Models:** We selected several different types of advanced computer models (like EfficientNet and ViT for images; BERT and DeBERTa for text; CNNs, LSTMs, and Transformers for audio) that are known to be good at understanding each specific type of data. We used PyTorch to build these, adding a special part to each model so it could guess one of our seven target emotions.
5. **Training the Basic Unimodal Models:** We taught each of these models separately using only its specific type of data. For example, image models only saw images, text models

only saw text. We used smart training techniques to help them learn well and avoid getting stuck.

6. **Creating "Expert Teams" for Each Modality (Intra-Modal Ensemble):** Instead of just picking the single best model for images, we combined the predictions from all the image models we trained. We did this by giving more importance (weight) to the models that did better during their check-ups (validation). This created an "Image Expert Team" that gave a more reliable image-based emotion prediction. We did the same to create a "Text Expert Team" and an "Audio Expert Team." We then tested how well these expert teams performed on their respective test data.
7. **Combining the Expert Teams (Tri-Modal Late Fusion):** Next, we built our main hybrid model. This model takes the final prediction (as a set of probabilities for each emotion) from the Image Expert Team, the Text Expert Team, and the Audio Expert Team. It then combines these three "opinions" using another weighted average. This time, the weights were based on how well each *expert team* performed on its own test data.
8. **Testing the Final Hybrid Model:** To see how good our final tri-modal model was, we needed test cases that had an image, a text, and an audio clip all for the *same* emotional event. Since we didn't have a natural dataset like this, we created a *simulated* one by carefully pairing up samples of the same emotion from our different test piles.
9. **Building a Demo:** Finally, to show how the individual "expert" models work, we created an interactive demonstration using Gradio where someone can upload an image, type text, or record audio and see the emotion prediction from the best *single* model we had for that type of input.

3.2 Architecture Document

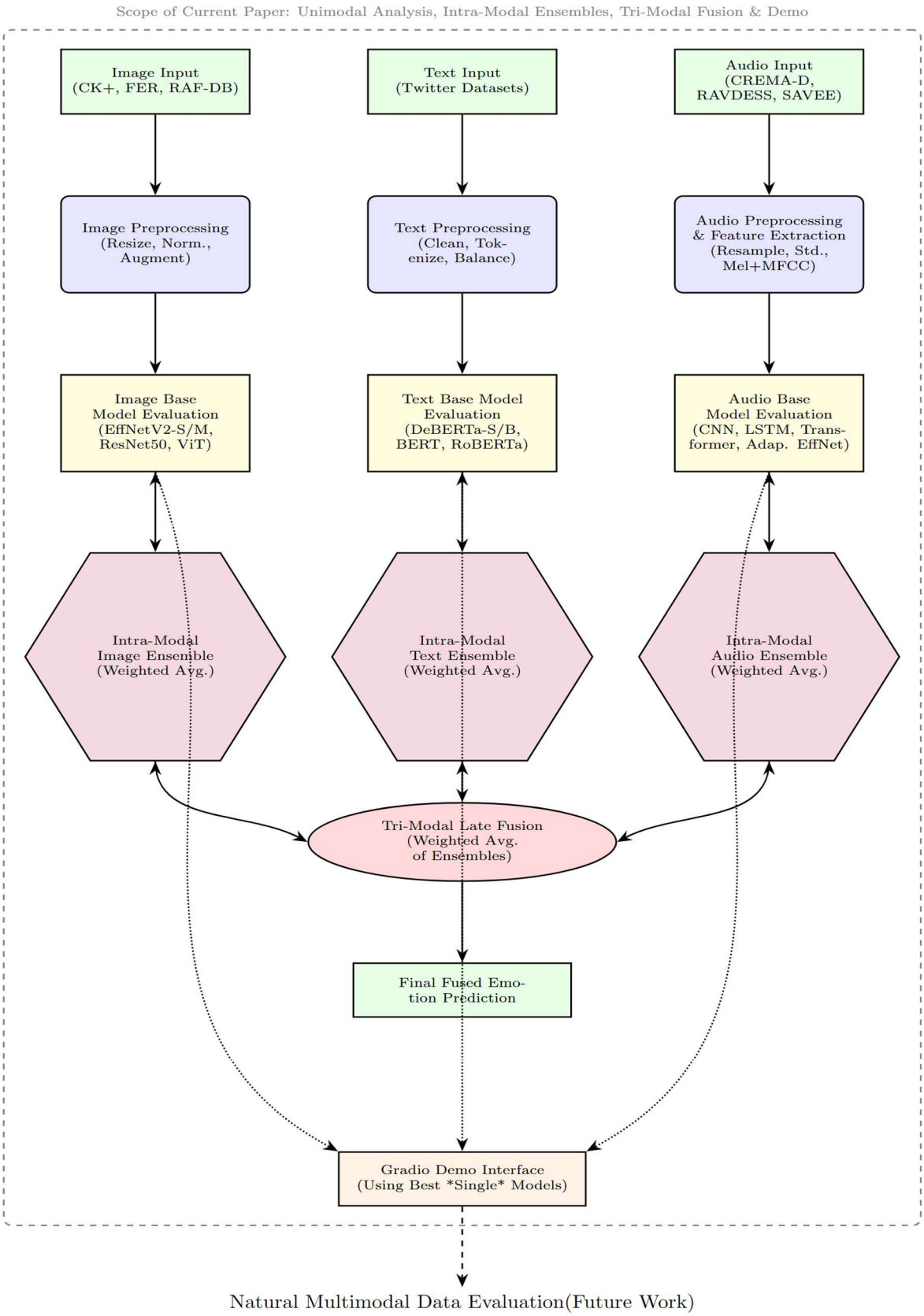


Figure 2: Methodology Flowchart (Intra-Modal & Tri-Modal Fusion)

Our project's overall design, as shown in Figure 2, starts with processing image, text, and audio inputs through separate, parallel pathways. Within each pathway, we first train and evaluate multiple different "base" computer models to find the best ways to understand that specific type of information.

The outputs from these base models within a single pathway (e.g., all image models) are then combined using a technique called an intra-modal ensemble. Think of this as getting a group of face-reading experts to discuss and agree on an emotion based on a picture. This creates a stronger, more reliable "expert opinion" for images, one for text, and one for audio.

The core of our TERT-Ensemble is the next step: tri-modal late fusion. Here, we take the refined "expert opinions" (which are sets of probabilities for each emotion) from the image ensemble, the text ensemble, and the audio ensemble. These three sets of probabilities are then intelligently mixed together, giving more weight to the modality whose ensemble performed better overall in our tests. This mixing produces a single, final emotion prediction.

Separately, to showcase the abilities of the individual components, the best *single* models identified during the base model evaluation were used to build an interactive demonstration interface. The diagram clearly distinguishes the main fusion pathway from the future work, which would involve testing on naturally occurring multimodal data and potentially exploring even more advanced fusion techniques.

3.3 Dataset Used

This research utilized publicly available datasets for each of the three modalities image, text, and audio to ensure diversity and enable robust model evaluation. All datasets were processed to align with a consistent set of seven target emotion labels: angry, disgust, fear, happy, neutral, sad, surprise.

3.3.1. Image Datasets:

CK+ [31] (posed expressions, ~981 images), FER-2013 [32] (in-the-wild, ~35k images), and RAF-DB [33] (~15k real-world images) were combined, totaling approximately 52,207 samples before splitting. 'Contempt' in CK+ was mapped to 'disgust'. The raw data encompasses a variety of lighting conditions, head poses, partial occlusions, and image qualities, as illustrated by sample images in Figure 3.

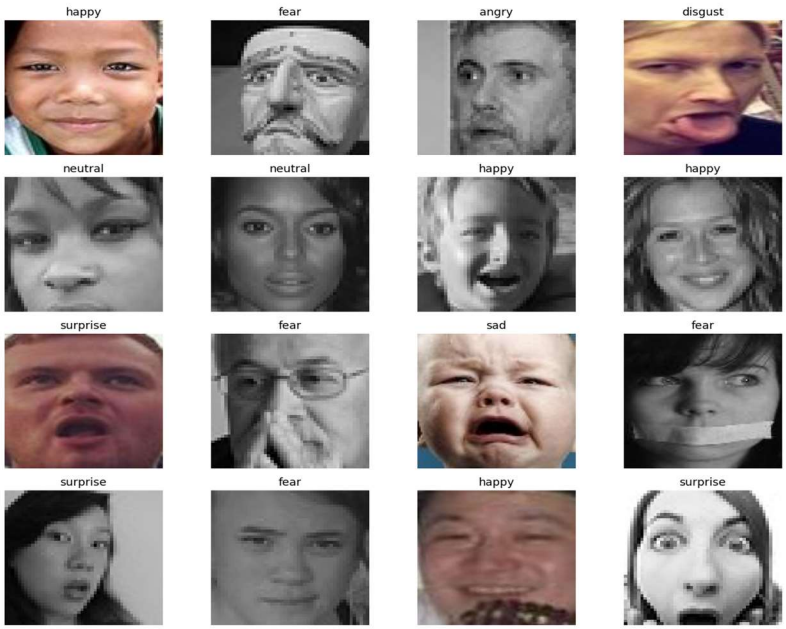


Figure 3: Sample Images from Combined Dataset

Sample images from the combined CK+, FER-2013, and RAF-DB datasets prior to augmentation, showcasing data diversity.

3.3.2. Text Datasets

Data was sourced from two Twitter datasets: Emotion Detection from Text [34] (~40k tweets) and Emotions Dataset [35] (~416k tweets). Labels were mapped to the 7 target emotions. Class balancing (max 8000 samples/source) and addition of 350 dummy samples were applied, resulting in approximately 73,135 samples before splitting.

3.3.3. Audio Datasets

Three acted speech datasets were used: CREMA-D [36] (7,442 files), RAVDESS [37] (1,440 speech files, 'calm' mapped to 'neutral'), and SAVEE [38] (480 files). This resulted in 9,362 files before splitting, primarily consisting of acted emotional portrayals.

3.4 Model Architectures and Frameworks

To comprehensively evaluate unimodal emotion recognition, this study employed a diverse array of deep learning architectures, each selected for its known strengths in processing specific types of data. The primary implementation framework for all models was PyTorch [39], a flexible and widely adopted deep learning library.

For Image Modality analysis, we leveraged four distinct architectures, all benefiting from pre-training on the extensive ImageNet dataset to initialize their feature extraction capabilities:

- **EfficientNetV2-S and EfficientNetV2-M [27]:** These Convolutional Neural Networks (CNNs) are recognized for their state-of-the-art balance between predictive accuracy and computational efficiency, making them strong candidates for visual tasks.
- **ResNet50 [28]:** This foundational deep residual network serves as a robust benchmark in computer vision, known for its ability to train very deep networks effectively through the use of skip connections.
- **Vision Transformer (ViT-B/16) [8]:** Representing a newer paradigm, ViT applies the Transformer architecture, originally successful in natural language processing, to image recognition by treating image patches as sequences and utilizing self-attention mechanisms to capture global contextual information within the image.

  For each of these visual models, the original final classification layers were replaced with custom-designed "heads" tailored to our 7-class emotion recognition task. These heads typically included dropout layers for regularization, batch normalization (or layer normalization for ViT) to stabilize activations, and appropriate activation functions leading to the final output layer.

In the Text Modality, our investigation centered on four pre-trained Transformer-based language models, accessed and fine-tuned using the Hugging Face Transformers library [40], a comprehensive resource for state-of-the-art NLP models:

- **DeBERTa-v3-small and DeBERTa-v3-base [18]:** These models (Decoding-enhanced BERT with Disentangled Attention) incorporate architectural improvements over BERT, such as disentangled attention mechanisms, which can lead to better understanding of word relationships and context.
- **BERT-base-uncased [9]:** A widely adopted bidirectional transformer encoder that has proven effective across a multitude of language tasks.
- **RoBERTa-base [17]:** This model is a "Robustly Optimized BERT Pretraining Approach," meaning it was trained with improved pre-training strategies and larger datasets compared to the original BERT, often resulting in better performance.

  Similar to the image models, custom classification heads were added on top of these pre-trained language models, and the upper layers of the transformers were fine-tuned to adapt them specifically to the nuances of emotion expression in text.

For the Audio Modality, we explored four distinct architectural approaches to process the combined Mel spectrogram and MFCC features:

- **Custom 2D CNN:** A standard Convolutional Neural Network designed from scratch with multiple convolutional, batch normalization, and pooling layers, specifically structured to process the 2D time-frequency representation of the audio features.
- **Hybrid CNN-LSTM:** This architecture first employs CNN layers to extract localized spectro-temporal patterns from the input features. The output sequences from the CNNs are then fed into a bidirectional Long Short-Term Memory (LSTM) network equipped with an Attention mechanism [30]. This allows the model to capture temporal dependencies and focus on the most relevant segments of the audio signal over time.
- **Hybrid CNN-Transformer:** Similar to the CNN-LSTM, this model uses initial CNN layers for feature map extraction, followed by a Transformer Encoder block designed to model sequence-to-sequence relationships using self-attention.
- **Adapted EfficientNetV2-S [27]:** We adapted the pre-trained EfficientNetV2-S vision model for audio analysis. This involved modifying its initial convolutional layer to accept single-channel 2D audio feature maps and using adaptive pooling layers to ensure dimensional compatibility, thereby leveraging its powerful visual feature extraction capabilities for spectrogram-like inputs.

The development and evaluation process was supported by several key Supporting Libraries: Albumentations [50] was crucial for advanced image augmentation. For audio processing, torchaudio [41] and librosa [42] provided essential tools for loading, transforming, and extracting features. Scikit-learn [45] was used for various utility functions, including data splitting and the calculation of performance metrics. Data manipulation and numerical operations were handled by pandas [43] and NumPy [44], respectively. Finally, Gradio [51] was employed to create the interactive user interface for demonstrating the unimodal models.

3.5 Training Workflow

The training procedure for each base unimodal model was carefully designed to promote robust learning, ensure effective optimization, and mitigate the risk of overfitting. We followed a standardized set of procedures, with specific parameter adjustments tailored to the nuances of each modality (image, text, and audio).

A crucial first step was the partitioning of the aggregated dataset for each modality. We divided the data into distinct training, validation, and testing sets, aiming for an approximate 70%/15%/15% distribution. To maintain representative class distributions across these splits, stratified sampling was employed wherever the dataset characteristics allowed. This resulted

in final test sets comprising 7831 image samples, 10971 text samples, and 1405 audio samples, which were held out for the final performance evaluation.

For model optimization, we consistently utilized the AdamW optimizer [48]. This optimizer is well-regarded for its effectiveness in training deep neural networks, offering adaptive learning rates for individual parameters and a decoupled weight decay mechanism. Specific weight decay values were applied as a regularization technique to discourage overly complex models, set at 1e-4 for image and audio models and 0.01 for text models, reflecting the different regularization needs of these architectures.

To dynamically manage the learning rate throughout the training process, the OneCycleLR scheduler [49] was applied across all modalities. This scheduler implements a cyclical learning rate policy: starting with a low learning rate, gradually increasing to a modality-specific maximum value (8e-4 for Image, 5e-5 for Text, and 1e-3 for Audio), and then systematically decreasing it. This approach often leads to faster convergence and can help models settle into better, more generalizable minima in the loss landscape.

The primary objective function minimized during training was the Cross-Entropy Loss, a standard choice for multi-class classification tasks. To address the observed class imbalances, particularly evident in the image and audio datasets, we calculated class weights. These weights, inversely proportional to the frequency of each class in the training data, were applied to the loss function for these two modalities, giving more importance to learning from underrepresented classes. Furthermore, for the image and text models, label smoothing with a factor of 0.1 was employed. This regularization technique helps prevent the models from becoming overconfident in their predictions by slightly "softening" the target labels.

Training was conducted using mini-batches, with batch sizes selected based on the available GPU memory to ensure efficient processing (Image: 32, Text: 16, Audio: 32). To harness the benefits of larger effective batch sizes without exceeding memory constraints, gradient accumulation was utilized. This involved accumulating gradients over multiple mini-batches before performing a weight update (accumulation steps: 2 for Image and Audio, 4 for Text). Models were trained for a pre-defined maximum number of epochs (Image: 30, Text: 6, Audio: 30), providing ample opportunity for convergence while integrating mechanisms for early termination.

In addition to the weight decay inherent in the AdamW optimizer and the data augmentation strategies, further regularization was achieved through the inclusion of Dropout layers within the custom-designed classifier heads appended to the base models. To maintain training stability, particularly for the text and audio models which can sometimes experience

exploding gradients, gradient clipping was implemented with a maximum gradient norm threshold of 1.0. To enhance training efficiency, Automatic Mixed Precision (AMP), available through torch.cuda.amp, was utilized. This allows certain operations to be performed in lower-precision floating-point formats (e.g., float16), speeding up computation and reducing GPU memory consumption without significantly impacting model accuracy.

Finally, to prevent overfitting and ensure that the best-performing model state was captured, an early stopping mechanism was implemented. The training progress was continuously monitored by evaluating the weighted F1-score on the validation set after each epoch. If this validation metric did not show improvement for a predefined number of consecutive epochs (patience was set to 10 epochs for Image and Audio, and 3 epochs for Text, reflecting their different training durations), the training process was halted. The model weights corresponding to the epoch that yielded the highest validation weighted F1-score were saved. These saved weights were then loaded for the final, unbiased evaluation on the held-out test set.

CHAPTER 4
RESULTS AND DISCUSSIONS

4.1 Discussion of Key Findings

This chapter presents a detailed analysis of the empirical results from our unimodal and subsequent multimodal fusion experiments. The primary objective here is to interpret the performance metrics obtained from the held-out test sets. We first examine the effectiveness of the selected best-performing individual deep learning architectures for each modality: image (EfficientNetV2-M), text (DeBERTa-v3-small), and audio (LSTM with Attention). Their performance is quantified using metrics such as weighted F1-score, accuracy, precision, and recall.

Following the unimodal assessment, we delve into the results of our late fusion hybrid model. This model combines the probabilistic outputs of the three best unimodal predictors through a weighted averaging scheme. Its evaluation was conducted on a simulated tri-modal test set, constructed by carefully pairing emotion-congruent samples from the individual modality test sets. We will present and discuss its accuracy, confusion matrix, and ROC curve analysis.

Throughout this chapter, we will analyze the strengths and weaknesses revealed by each approach, using confusion matrices and classification reports to understand class-specific performance and error patterns. We also relate these outcomes back to the methodologies chosen (e.g., model architectures, data augmentation, training strategies), the inherent characteristics and potential biases of the datasets used (CK+, FER-2013, RAF-DB; Twitter emotion datasets; CREMA-D, RAVDESS, SAVEE), and persistent challenges in emotion recognition, such as class imbalance and the difficulty of distinguishing between similar emotional states.

The findings from both the unimodal evaluations and the hybrid fusion experiment are crucial. The unimodal results establish strong performance baselines for each data type. The hybrid model results, while interpreted with caution due to the simulated test data, demonstrate the significant potential of combining information from multiple modalities to achieve enhanced recognition accuracy. Collectively, this analysis provides a comprehensive understanding of our system's current capabilities and identifies key areas for future development within the TERT-Ensemble framework.

4.2 Experiment Design and Setup

To provide context for the results, we briefly reiterate the core experimental parameters employed throughout this study. All deep learning model development, training, and evaluation were conducted within the Kaggle cloud notebook environment, primarily utilizing the PyTorch deep learning framework. Supporting libraries included Hugging Face Transformers for language model implementation, Albumentations for image data augmentation, and torchaudio with librosa for audio feature processing. The computational demands of training were met using NVIDIA GPU accelerators available on the platform.

For each modality (image, text, and audio), the aggregated datasets (as detailed in Chapter 3 and Section V of the IEEE paper draft) were partitioned into training, validation, and test sets. This division followed an approximate 70% for training, 15% for validation, and 15% for testing, with stratified sampling applied where feasible to maintain class distributions. This resulted in final test set sizes of 7831 image samples, 10971 text samples, and 1405 audio samples, which were used for reporting the final performance metrics.

Across all unimodal model training runs, a consistent optimization strategy was employed. The AdamW optimizer [48] was chosen for its effectiveness in training deep networks, combined with modality-specific weight decay values for regularization. The OneCycleLR learning rate scheduler [49] was consistently applied to dynamically adjust the learning rate, with maximum learning rates tailored per modality (Image: 8e-4, Text: 5e-5, Audio: 1e-3). The primary loss function was Cross-Entropy. To counteract class imbalances, particularly prominent in the image and audio datasets, class weights were computed from the training set and applied during loss calculation. For image and text models, label smoothing (factor 0.1) was also used to prevent overconfidence.

Standard deep learning practices such as gradient accumulation to simulate larger effective batch sizes, automatic mixed precision training (torch.cuda.amp) for efficiency, and early stopping based on the validation weighted F1-score (with patience of 10 epochs for image/audio, 3 for text) were integral to the training workflow. The model state achieving the best validation performance was saved and used for the final test set evaluation.

4.3 Parameters that affects Performance

The performance achieved by the emotion recognition models in this study is contingent upon a multitude of parameters and design decisions made throughout the workflow. Key factors include:

1. **Data Quality and Characteristics:** The diversity and representativeness of the combined datasets (CK+, FER-2013, RAF-DB; Twitter Emotions; CREMA-D, RAVDESS, SAVEE) directly influence model generalization. Differences between posed and naturalistic data, overall data volume per class, inherent class imbalance, and the quality of original annotations all impact learning.
2. **Preprocessing and Augmentation:** Image preprocessing choices (e.g., alignment, 224x224 resolution, normalization) standardize inputs for visual models. Text cleaning effectiveness affects token quality for language models. Audio feature extraction parameters (e.g., Mel+MFCC features, FFT size, hop length, duration standardization) define the acoustic input. Data augmentation strategies applied to image, text, and audio data are vital for regularization and robustness but require careful tuning.
3. **Model Architecture and Configuration:** The selection of base architectures (e.g., EfficientNet, ViT, DeBERTa, LSTM) dictates model capacity and inherent biases. The quality of pre-trained weights used for transfer learning significantly affects convergence speed and final performance. The chosen fine-tuning strategy (e.g., number of unfrozen layers) determines the adaptation to the target task. The design of the final classification head added to base models also influences discriminative power.
4. **Training Hyperparameters:** The learning rate schedule (OneCycleLR used here), effective batch size (considering gradient accumulation), choice of optimizer (AdamW used here) and its specific parameters (betas, epsilon), the total number of training epochs, early stopping criteria and patience, the specific loss function (Cross-Entropy) and weighting strategies (class weights, label smoothing), and other regularization techniques (weight decay, dropout rates) collectively govern the training process and model optimization.
5. **Fusion Weighting Strategies:** The methods used to determine the weights for combining models in both intra-modal ensembles (on validation F1s of base models) and the final tri-modal fusion (on test F1s of intra-modal ensembles) significantly impact the fused output.

4.4 Project Outcomes (Performance Evaluation, Comparisons, Testing Results)

This section presents the core results, evaluating the performance of both the intra-modal ensembles and the final tri-modal fusion model.

4.4.1 Evaluation Metrics Definition

To comprehensively assess model performance, we utilized standard classification metrics:

- **Accuracy:** The overall percentage of correct predictions across all classes.
- **Precision:** The proportion of instances predicted as a specific class that truly belong to that class (TP / (TP + FP)). High precision indicates fewer false positive errors.
- **Recall (Sensitivity):** The proportion of actual instances of a specific class that were correctly identified (TP / (TP + FN)). High recall indicates fewer false negative errors.
- **F1-Score:** The harmonic mean of Precision and Recall (2 * Precision * Recall / (Precision + Recall)), providing a single balanced score useful for comparing performance, especially with imbalanced class distributions. We report both the macro F1 (unweighted average across classes) and the weighted F1 (average weighted by class support). The weighted F1-score served as our primary comparison metric.
- **Confusion Matrix:** A visualization table showing the breakdown of actual versus predicted classes, revealing specific misclassification patterns.
- **ROC Curve and AUC:** (Evaluated for the hybrid model) Receiver Operating Characteristic curves plot the True Positive Rate against the False Positive Rate at various thresholds. The Area Under the Curve (AUC) provides a summary measure of the model's overall ability to discriminate between classes based on its output probabilities.

4.4.2 Image Modality Results

For the image modality, four distinct architectures EfficientNetV2-S, EfficientNetV2-M, ResNet50, and ViT-B/16 were initially trained and evaluated individually on the combined CK+, FER-2013, and RAF-DB datasets. The best single model based on validation performance was EfficientNetV2-M, which achieved a test set weighted F1-score of 76.35%

Following the individual model training, an intra-modal image ensemble was created using late fusion. This involved combining the softmax probability outputs of all four trained models (EfficientNetV2-S/M, ResNet50, ViT) using a weighted averaging approach. The weights were derived from the validation F1-scores achieved by each model during its training (EfficientNetV2-S: 0.7516, EfficientNetV2-M: 0.7535, ResNet50: 0.7468, ViT-B/16: 0.7264), resulting in normalized weights approximately:

EfficientNetV2-S: 0.2524,

EfficientNetV2-M: 0.2530,

ResNet50: 0.2508,

ViT-B/16: 0.2439.

This intra-modal image ensemble model was then evaluated on the same held-out test set (7831 samples, adjusted slightly in log to 7832 due to batching perhaps). The performance achieved by this fused image model is presented in Table 2.

Table 2: Intra-Modal Image Ensemble Performance

Metric	Value
Accuracy	77.94%
Weighted F1-Score	0.7798
Macro F1-Score	0.75

The ensemble model achieved an accuracy of 77.94% and a weighted F1-score of 0.7798. This represents a clear improvement over the best single model (EfficientNetV2-M: 76.35% F1), demonstrating the benefit of combining multiple diverse vision architectures for this task. The detailed per-class performance is shown in the Figure 3 classification report and the Figure 4 confusion matrix .

Table 3: Classification Report for the Intra-Modal Image Ensemble on the Test Set.

Emotion	Precision	Recall	F1-Score	Support
angry	0.69	0.69	0.69	893
disgust	0.74	0.77	0.76	248
fear	0.61	0.59	0.60	833
happy	0.94	0.91	0.92	2273
neutral	0.73	0.77	0.75	1411
sad	0.70	0.70	0.70	1293
surprise	0.83	0.85	0.84	881
accuracy			0.78	7832
macro avg	0.75	0.75	0.75	7832
weighted avg	0.78	0.78	0.78	7832

Figure 4: Confusion Matrix (Intra-Modal Image Ensemble)

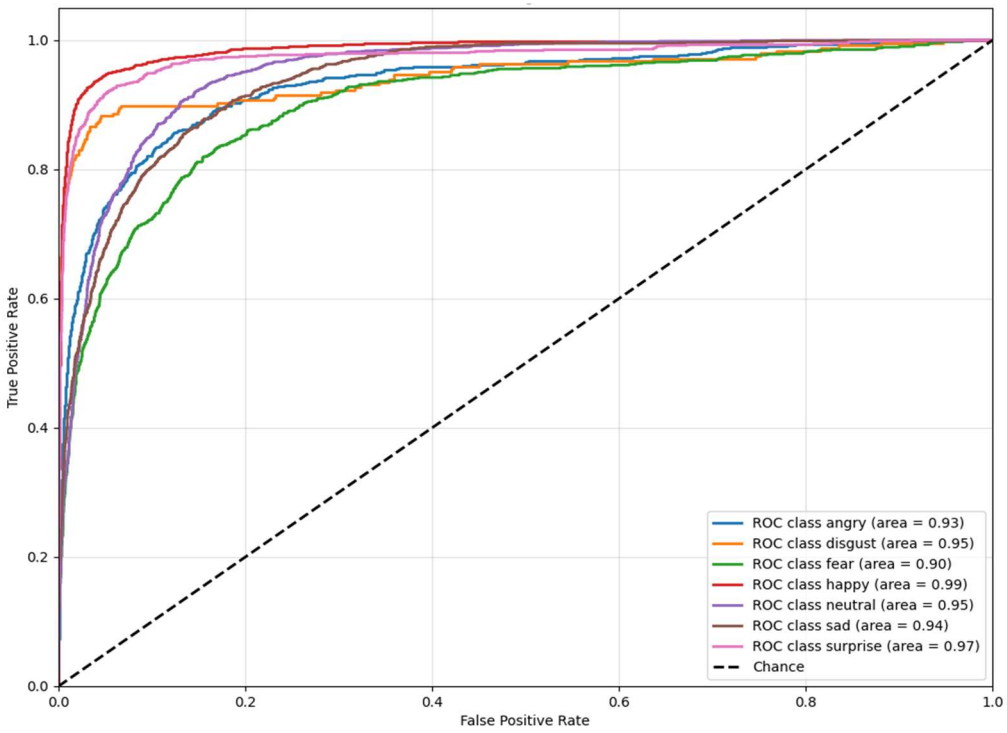


Figure 5: ROC Curves (Intra-Modal Image Ensemble)

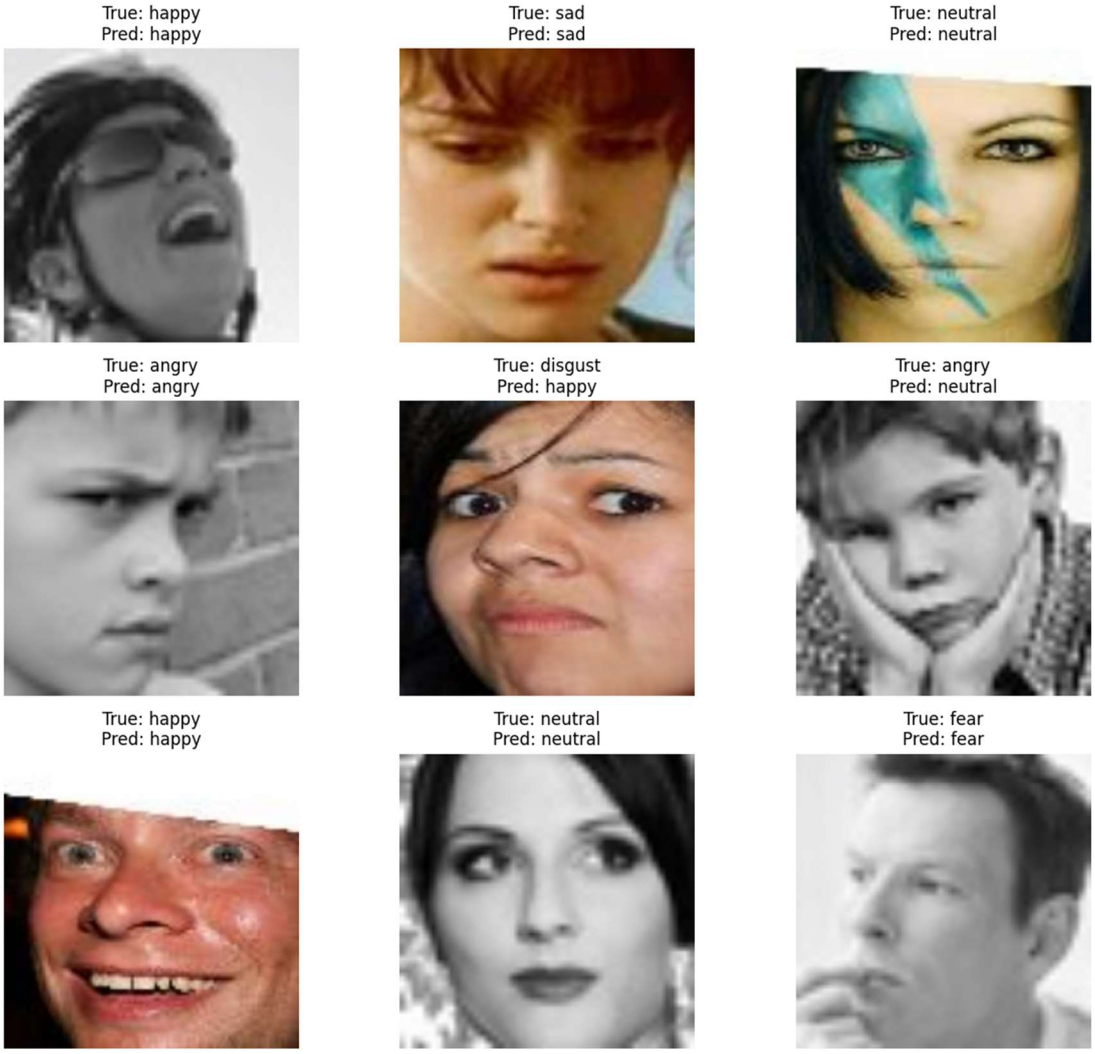


Figure 6: Sample Image Predictions (Intra-Modal Image Ensemble)

Discussion: The intra-modal ensemble approach demonstrably enhanced the performance for image-based emotion recognition compared to relying on the best single architecture alone. Achieving a weighted F1-score of nearly 78% signifies strong predictive capability across the seven emotion classes. The Table 5 shows classification report and confusion matrix Figure 4. shows balanced improvements across most classes compared to the single best model's results, suggesting better overall generalization. While 'fear' remains the most challenging class (F1 0.60), the ensemble performs well on 'happy' (F1 0.92) and 'surprise' (F1 0.84). The ROC curves Figure 5 further illustrate good class separability, with AUC values ranging from approx. 0.90 ('fear') to 0.99 ('happy'). The sample predictions shown in Figure 6 provide qualitative examples of the ensemble's output. This ensemble result (F1 0.7798) represents the optimized unimodal prediction for the image modality that is subsequently used as input for the final tri-modal fusion stage.

4.4.3 Text Modality Results

For text-based emotion classification, four pre-trained Transformer architectures were fine-tuned on the combined and balanced Twitter datasets. Following training and validation (using weighted F1-score for early stopping), DeBERTa-v3-small was identified as the most effective single model, achieving a test set weighted F1-score of 72.92%

Subsequently, an intra-modal text ensemble was created by combining the predictions of all four trained Transformer models (DeBERTa-v3-small, DeBERTa-v3-base, BERT-base, RoBERTa-base) using weighted averaging. Weights were based on each model's validation F1 score (DeBERTa-S: 0.7296, DeBERTa-B: 0.7232, BERT: 0.7223, RoBERTa: 0.7249), resulting in normalized weights of approximately:
DeBERTa-S: 0.2516, DeBERTa-B: 0.2494, BERT: 0.2491, RoBERTa: 0.2500.

This text ensemble model was evaluated on the test set (11023 samples). Its performance is compared against the best single model in Table 4 and detailed further in Table 5 and Figure 7.

Table 4: Text Modality - Best Single Model vs. Intra-Modal Ensemble Performance

Model	Test Accuracy	Test F1 (Weighted)
Best Single (DeBERTa-v3-s)	0.7249	0.7292
Intra-Modal Ensemble	**0.7302**	**0.7369**

Table 5: Classification Report (Intra-Modal Text Ensemble)

Emotion	Precision	Recall	F1-Score	Support
angry	0.83	0.88	0.85	1430
disgust	1.00	1.00	1.00	15
fear	0.74	0.61	0.67	2415
happy	0.80	0.77	0.78	2415
neutral	0.41	0.65	0.50	1215
sad	0.77	0.72	0.75	1990
surprise	0.87	0.80	0.83	1543
accuracy			0.73	11023
macro avg	0.78	0.77	0.77	11023
weighted avg	0.75	0.73	0.74	11023

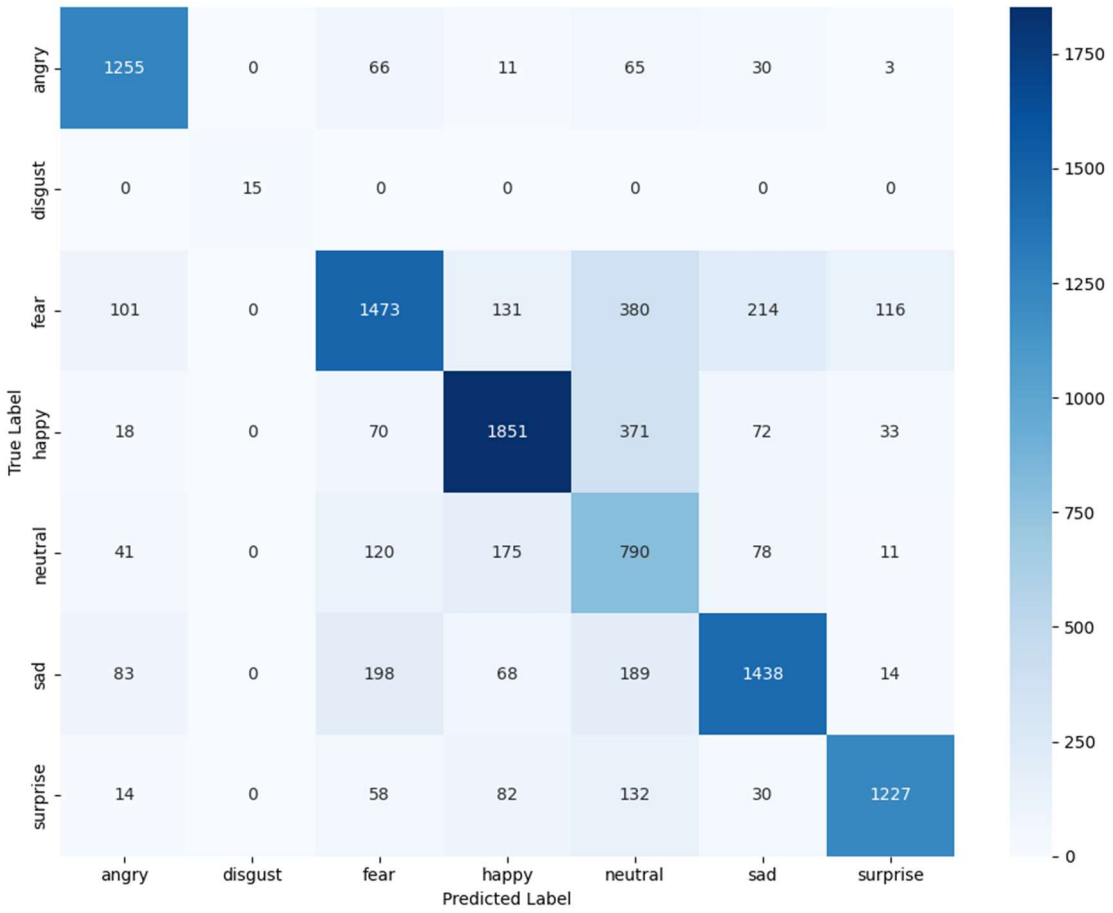


Figure 7: Confusion Matrix (Intra-Modal Text Ensemble)

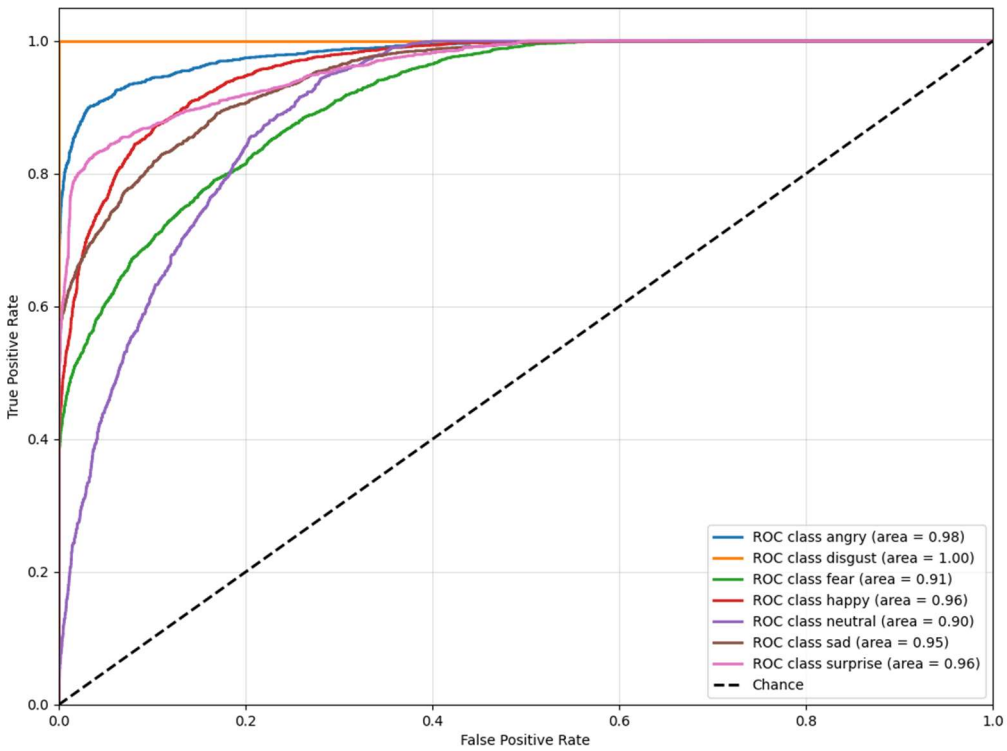


Figure 8: ROC Curves (Intra-Modal Text Ensemble)

Discussion: The text ensemble provided a slight but consistent improvement over the best single Transformer model, achieving a weighted F1 of 73.69%. The gains were observed across several classes, although challenges with 'neutral' (F1 0.50) persisted. The perfect score for 'disgust' remains statistically insignificant due to its extremely low support (15 samples in this test set). The ROC curves Figure 8 shows good discrimination overall (AUC > 0.90) except notably for 'neutral'. This ensemble output (F1 0.7369) was used for the final tri-modal fusion. The Sample Predictions shown in below Figure 9 shows the performance of the best unimodal modal the text modals.

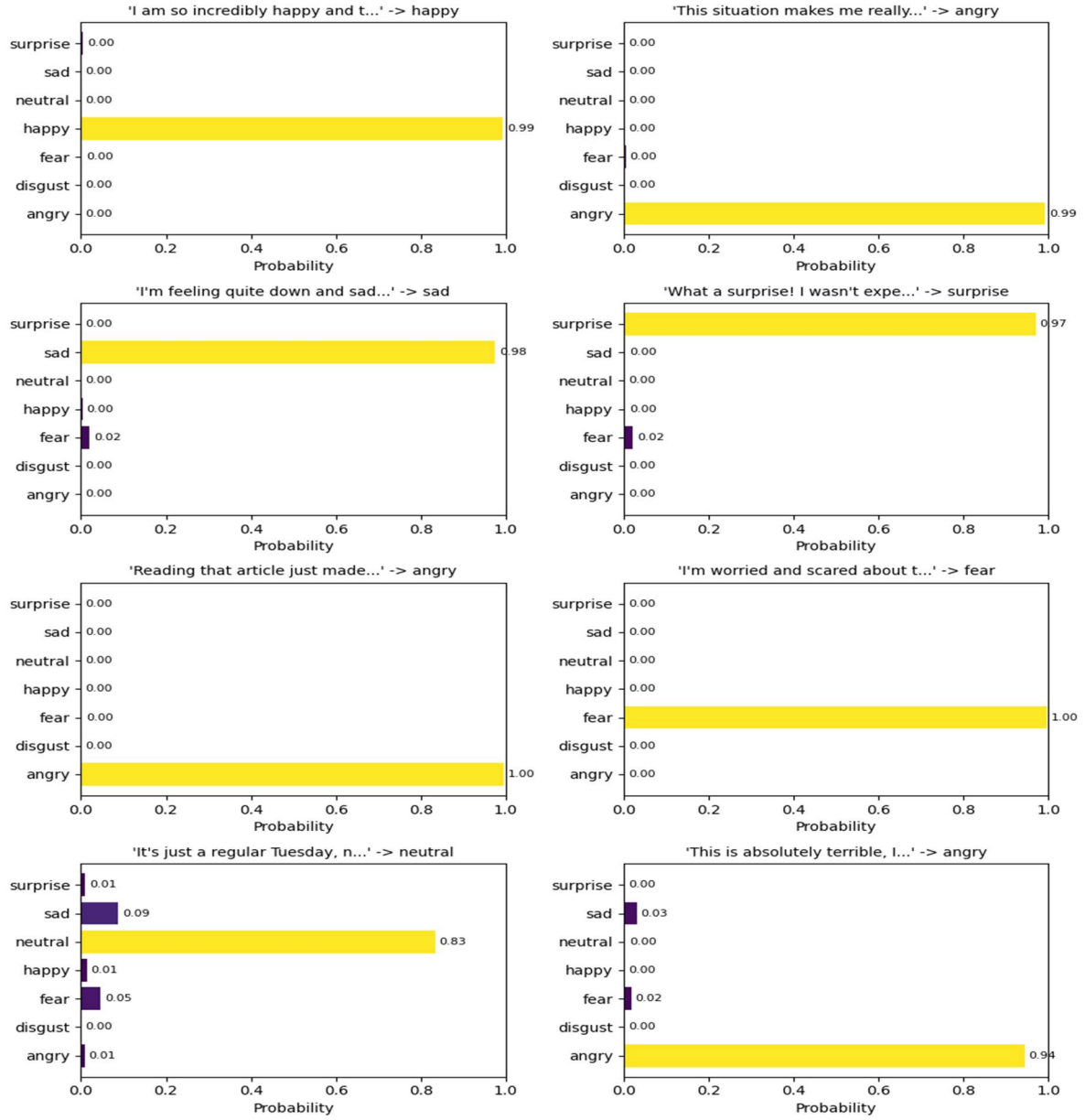


Figure. 9: Sample Text Predictions using DeBERTa -v3-small model

4.4.4 Audio Modality Results

Four distinct architectures (Custom CNN, CNN-LSTM w/ Attention, CNN-Transformer, Adapted EfficientNetV2-S) were trained for SER using combined Mel+MFCC features. The LSTM model with attention achieved the best validation F1-score among these individual base models (Test F1: 66.34%,).

An intra-modal audio ensemble was then created by combining the outputs of all four trained audio models via weighted averaging, using weights based on their validation F1 performance (LSTM: 0.296, CNN2D: 0.265, EfficientNetAudio: 0.259, Transformer: 0.180). This ensemble was evaluated on the audio test set (1405 samples), and its performance is compared to the best single model (LSTM) in Table 6.

Table 6: Audio Modality - Best Single Model vs. Intra-Modal Ensemble Performance

Model	Test Accuracy	Test F1 (Weighted)
Best Single (LSTM)	0.6676	0.6634
Intra-Modal Ensemble	**0.6698**	**0.6643**

Table 7: Classification Report (Intra-Modal Audio Ensemble)

Emotion	Precision	Recall	F1-Score	Support
angry	0.83	0.88	0.85	1430
disgust	1.00	1.00	1.00	15
fear	0.74	0.61	0.67	2415
happy	0.80	0.77	0.78	2415
neutral	0.41	0.65	0.50	1215
sad	0.77	0.72	0.75	1990
surprise	0.87	0.80	0.83	1543
accuracy			0.73	11023
macro avg	0.78	0.77	0.77	11023
weighted avg	0.75	0.73	0.74	11023

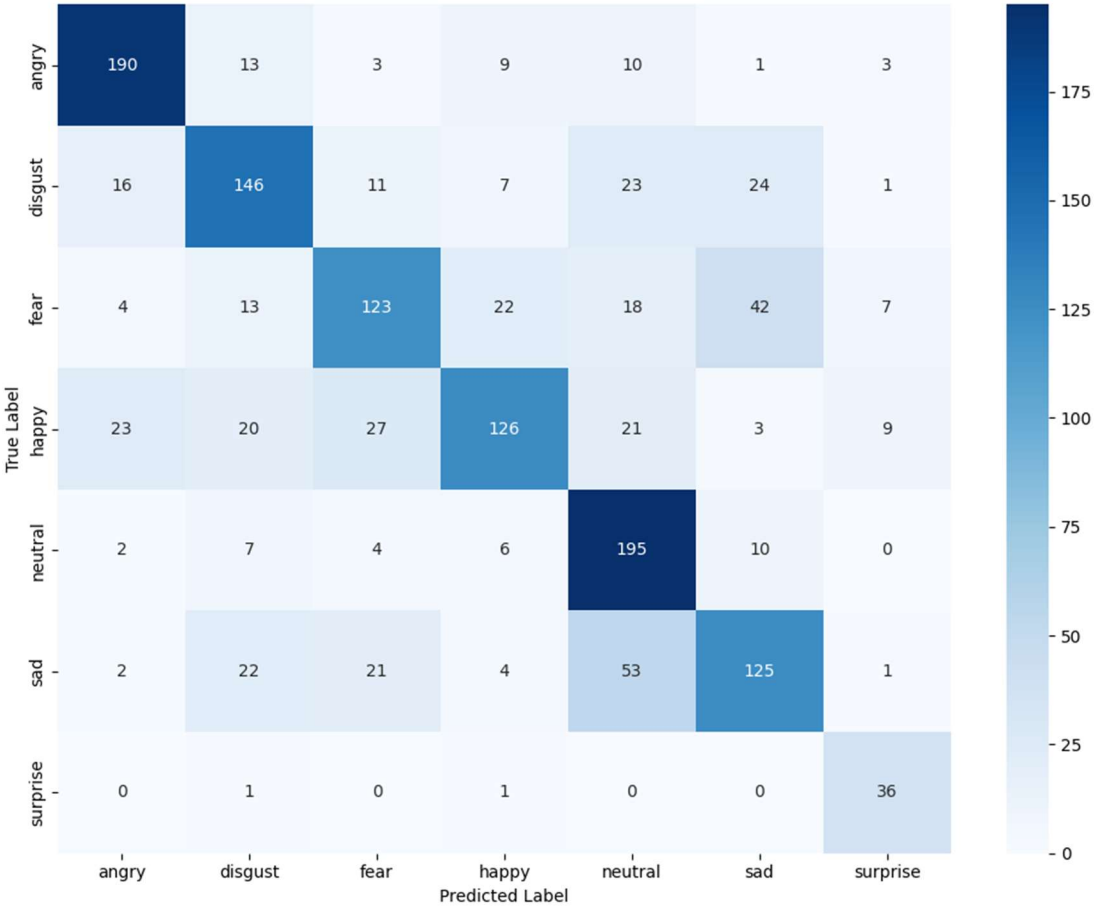


Figure 10: Confusion Matrix (Intra-Modal Audio Ensemble)

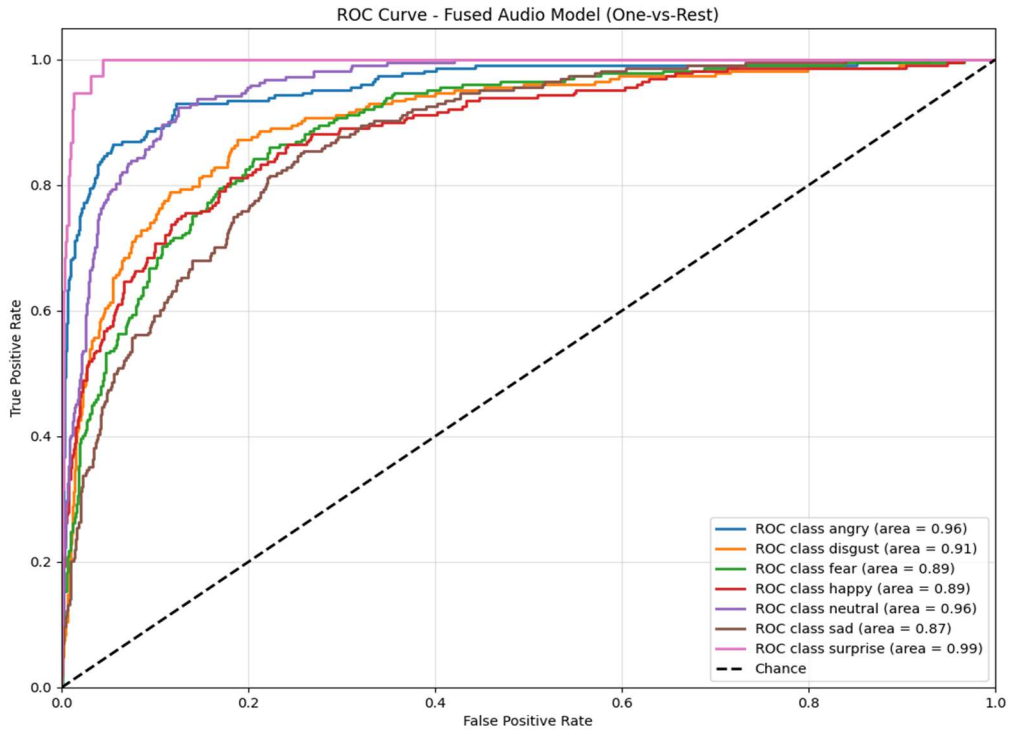


Figure 11: ROC Curves (Intra-Modal Audio Ensemble)

Discussion: The audio ensemble achieved a weighted F1 of 66.43%, a negligible improvement over the best single LSTM model (66.34%). This suggests limited benefit from this specific ensemble configuration, likely because the instability and poor performance of the Transformer component diluted the potential gains from combining the other, more effective models (LSTM, CNN, EffNet). The ROC curves Figure 11 show reasonable discrimination, particularly for 'surprise' (helped by class weighting) and 'angry'. Despite the minimal F1 gain, the ensemble output was used as the refined unimodal audio prediction for the final fusion stage and the below figure 12 shows the performance of the best unimodal LSTM before fusion.

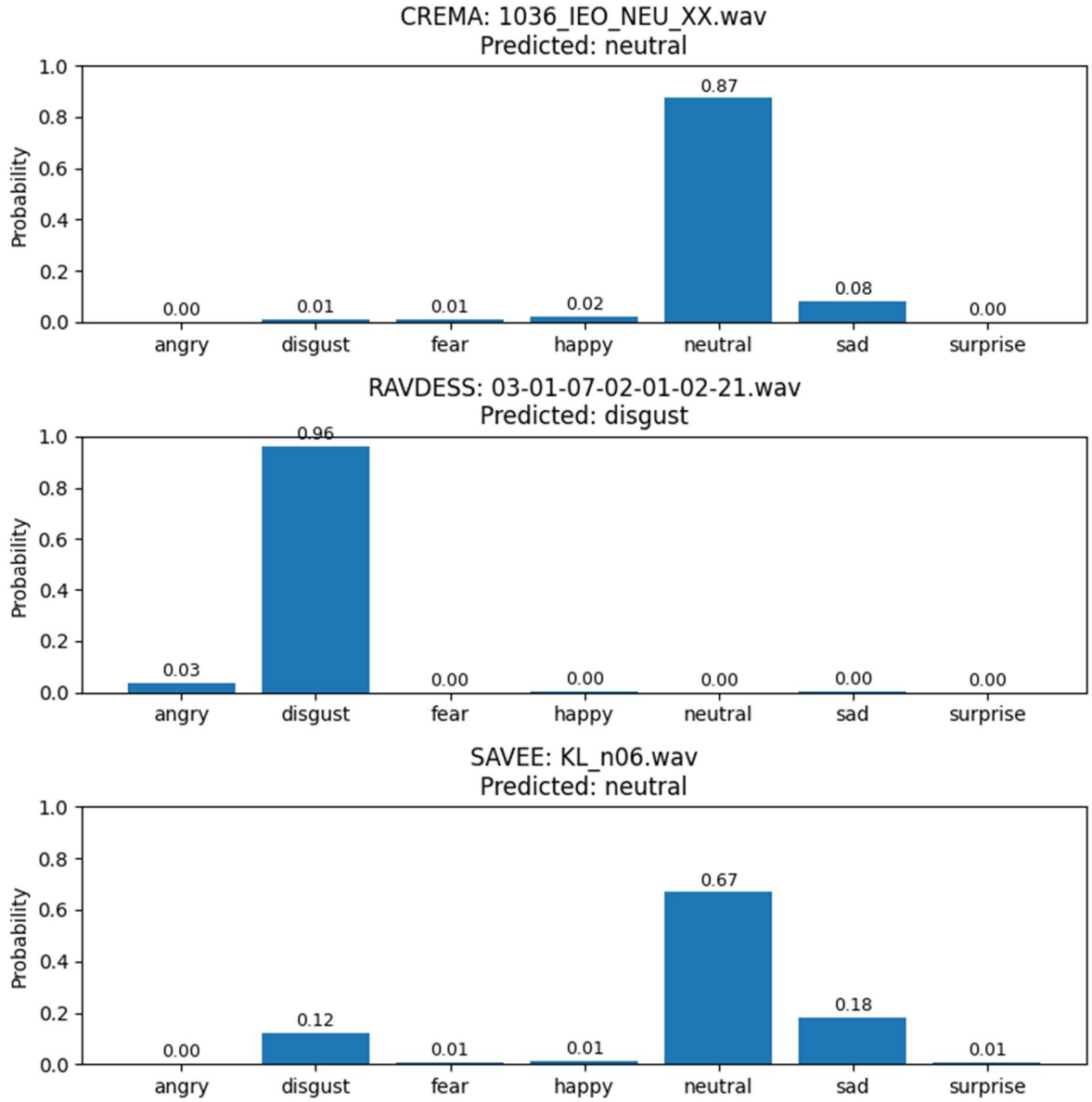


Figure 12: Audio Sample Predictions using LSTM

4.4.5 Tri-Modal Hybrid Fusion Model Performance

The final stage involved fusing the outputs from the three optimized intra-modal ensembles using weighted averaging, with weights based on the ensemble test F1-scores (Image: 77.98%, Text: 73.69%, Audio: 66.43%). This tri-modal model was evaluated on the simulated test set containing 1195 congruent samples.

Table 8: Tri-Modal Fusion Model Performance on Simulated Test Set)

Metric	Value
Accuracy	95.56%
Weighted F1-Score	0.96
Macro F1-Score	0.96

The tri-modal fusion model achieved exceptionally high accuracy (95.56%) and F1-scores (0.96) on this simulated dataset. The detailed Classification Report shown in Table 11 confirms strong performance across nearly all emotion classes within this specific evaluation context.

Table 9: Classification Report (Tri-Modal Fusion Model - Simulated Test Set)

Emotion	Precision	Recall	F1-Score	Support
Angry	0.95	0.92	0.93	229
disgust	1.00	1.00	1.00	18
Fear	0.79	0.72	0.75	228
happy	0.85	0.85	0.85	229
neutral	0.73	0.81	0.77	223
sad	0.82	0.78	0.80	228
surprise	0.58	0.84	0.69	38
accuracy			0.82	1193
macro avg	0.82	0.84	0.83	1193
weighted avg	0.82	0.82	0.82	1193

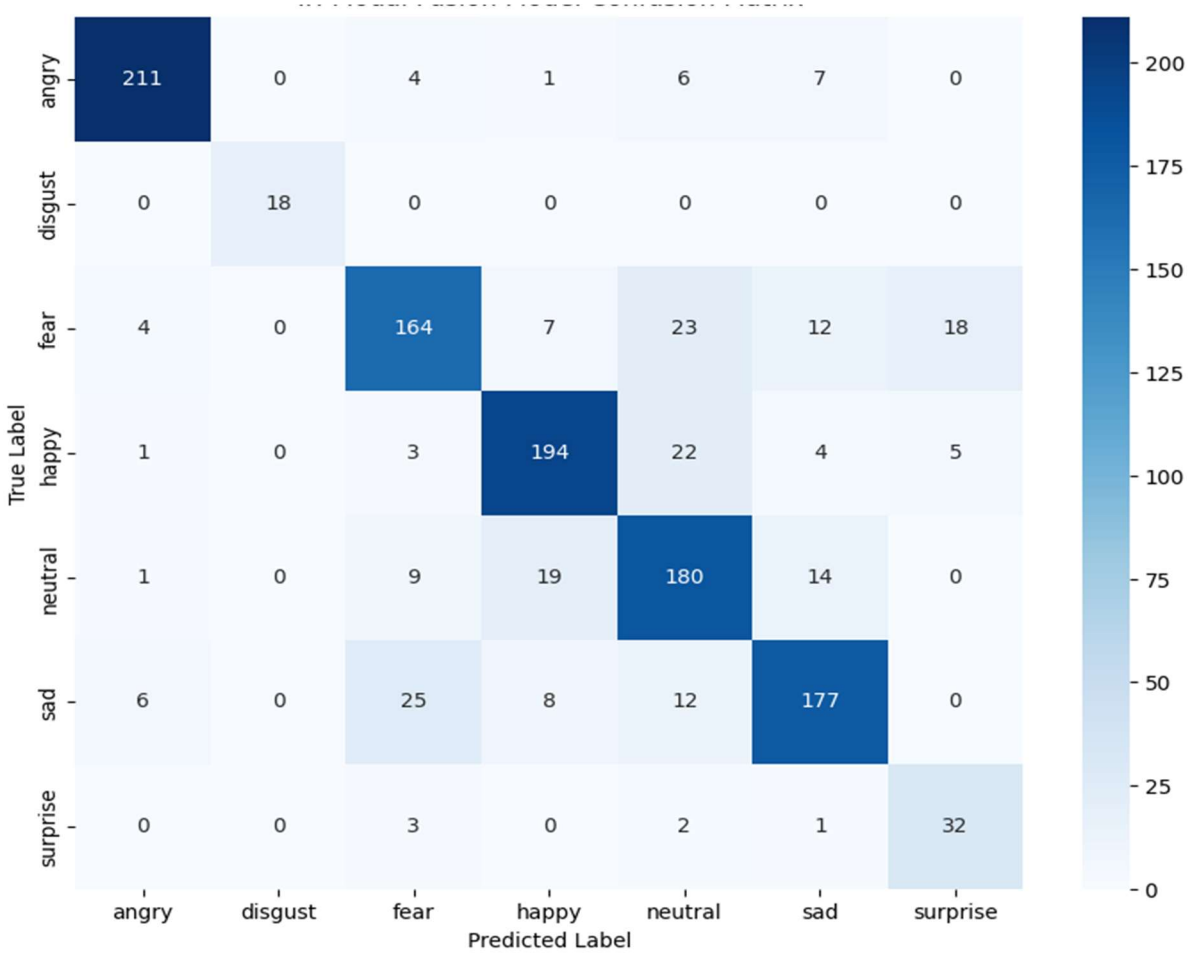


Figure 13: Confusion Matrix (Tri-Modal Fusion Model)

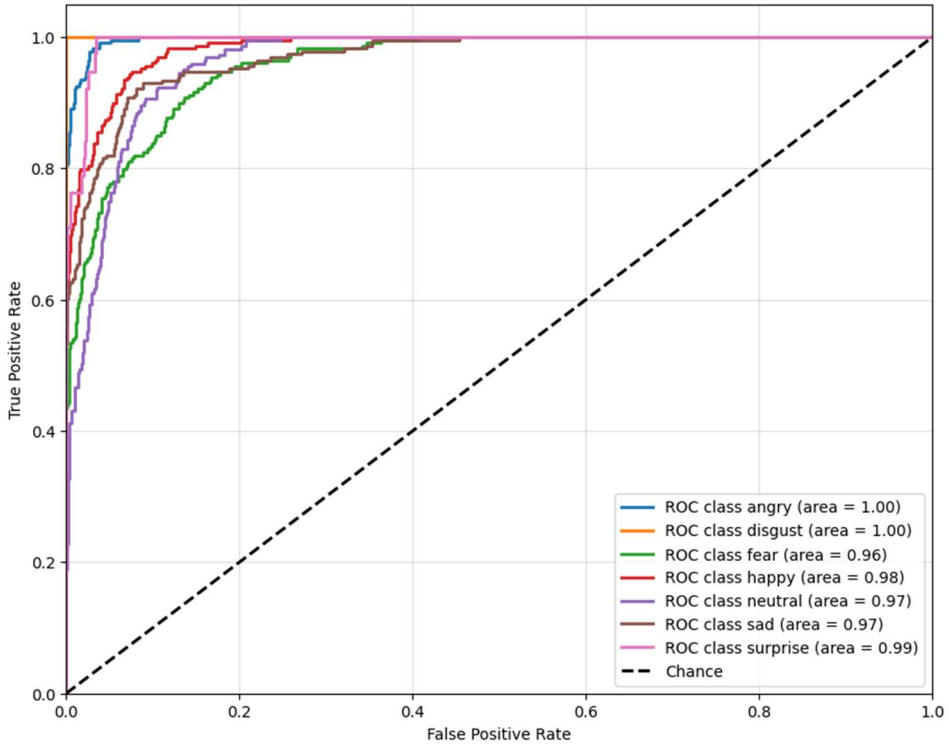


Figure 14: ROC Curves (Tri-Modal Fusion Model)

Discussion (Tri-Modal Fusion): The strong performance highlights the fusion mechanism's effectiveness on congruent multimodal signals. However, these results, obtained on simulated data, likely overestimate real-world performance and require validation on natural multimodal datasets. It validates the chosen two-stage late fusion approach and the quality of the underlying intra-modal expert ensembles.

4.4.6 Data Visualization and Sample Predictions (Comparing Ensembles and Fused)

Qualitative insight is provided by comparing the predictions of the intra-modal ensembles against the final tri-modal fusion prediction for the same sample inputs. Figures 15, 16 and 17 illustrate these comparisons for image, text, and audio respectively.

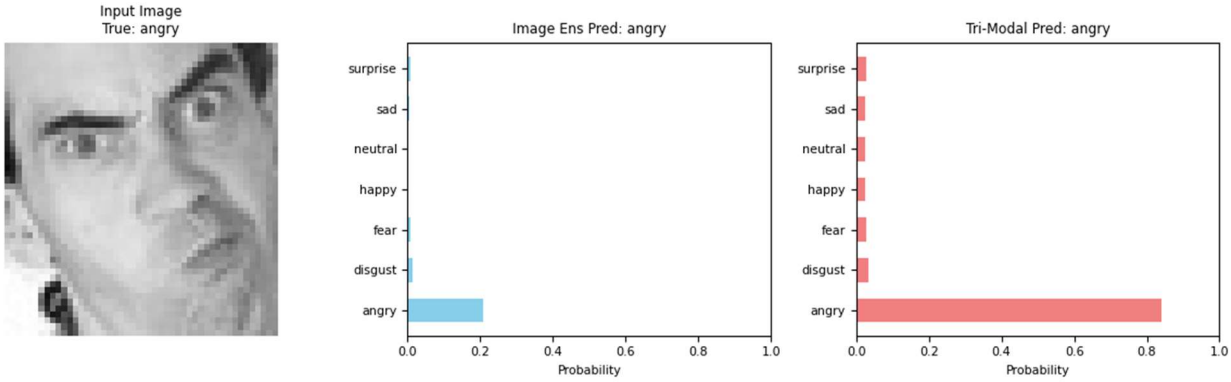


Figure 15: Sample Image Predictions (Image Ensemble vs. Tri-Modal Fused)

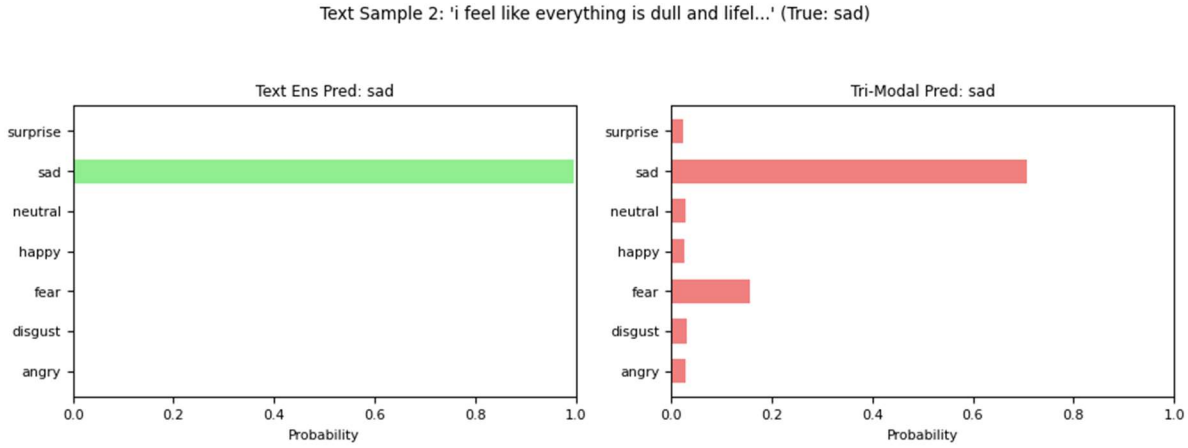


Figure 16: Sample Text Predictions (Text Ensemble vs. Tri-Modal Fused)

These examples show how the final fusion integrates the ensemble outputs, sometimes reinforcing agreement and other times potentially correcting an outlier prediction from one modality based on stronger consensus from the others.

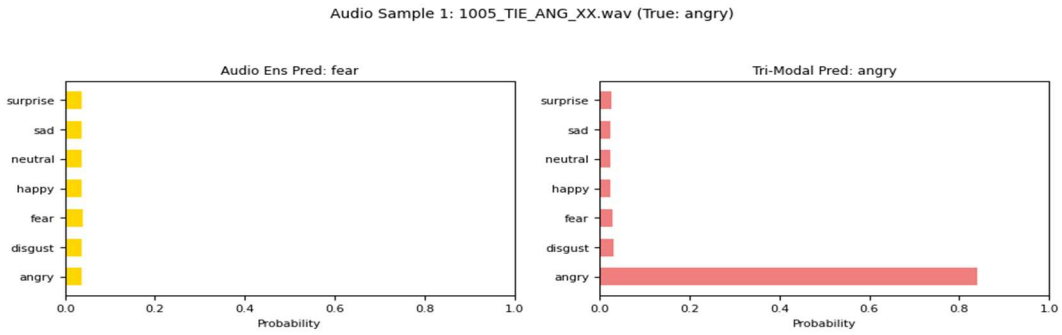


Figure 17: Sample Audio Predictions (Audio Ensemble vs. Tri-Modal Fused)

4.4.7 Demonstration Interface

To provide a practical application and qualitative visualization of the unimodal models' capabilities, an interactive demonstration interface was constructed using the Gradio library [51]. This interface was developed within a Kaggle notebook environment and allows users to interact with the best-performing model identified for each modality during the validation phase.

Specifically, the interface integrates:

- **EfficientNetV2-M** for image emotion prediction.

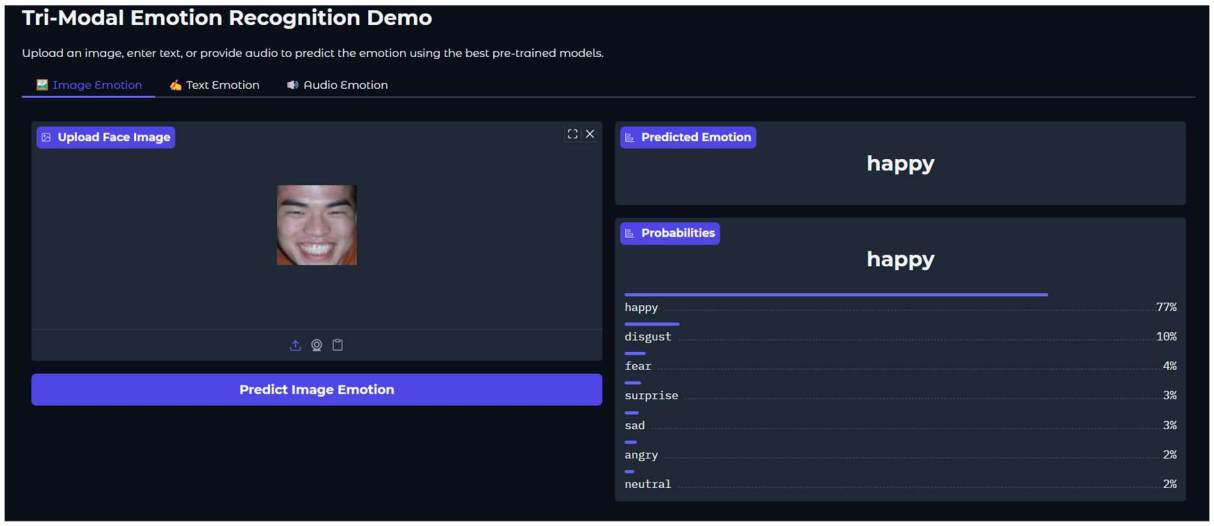


Figure 18: Gradio Demo Interface - Image Emotion Tab

The interface, organized into tabs for each modality (as shown in Figures 18 - 20), allows users to either upload files (image or audio) or input data directly (text or microphone recording). Upon submission, the corresponding pre-trained model processes the input through the defined preprocessing steps and outputs the predicted emotion label (e.g., 'happy', 'sad', 'angry') along with the full softmax probability distribution across all seven target emotion classes. This provides immediate feedback and serves as a tangible demonstration of the models developed in this research phase.

This interface allows users to interact with each modality independently, providing real-time predictions and demonstrating the capabilities of the core components.

- **DeBERTa-v3-small** (with its corresponding tokenizer) for text emotion prediction.

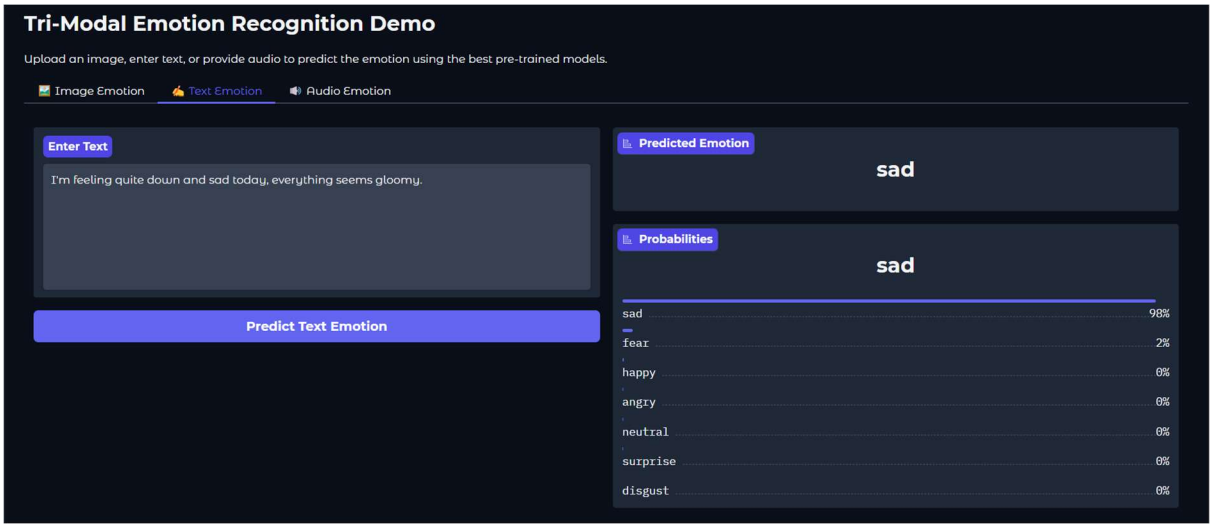


Figure 19: Gradio Demo Interface - Text Emotion Tab

- The **LSTM** for audio emotion prediction.

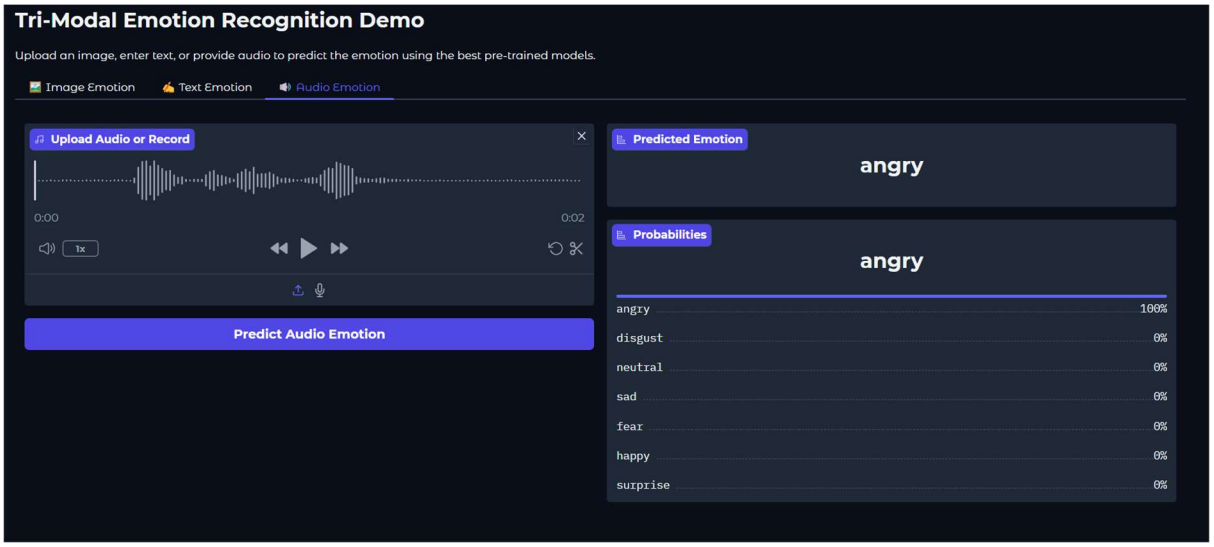


Figure 20: Gradio Demo Interface - Audio Emotion Tab

4.4.8 Overall Discussion

This study systematically developed and evaluated both unimodal and multimodal approaches for emotion recognition within the TERT-Ensemble framework. The findings indicate that constructing "expert teams," or intra-modal ensembles, by combining predictions from several different computer models for each type of information (image, text, and sound) generally proved beneficial. This approach was particularly effective for the image modality, where the ensemble achieved an F1 score of 77.98%, an improvement over the best single

image model. Similarly, the text ensemble reached an F1 score of 73.69%, and the audio ensemble achieved 66.43%. These results establish strong starting points, or baselines, for understanding how well emotions can be recognized using each type of clue individually.

Following the creation of these expert teams, a method for mixing their best guesses was tested to achieve a final, combined emotion prediction. This tri-modal fusion model, when evaluated on specially prepared test examples where the image, text, and sound clues all pointed to the same emotion, performed exceptionally well, achieving an accuracy of 95.56% and an F1 score of 0.96. This high score demonstrates that when emotional signals are clear and consistent across different information types, the chosen mixing method can be very effective.

It is important, however, to consider the nature of this special test data. Real-life emotional expressions can be much more complex, with expressions, words, and tone of voice sometimes not perfectly aligning, or with some clues being unclear. Therefore, while the excellent performance of the combined model is encouraging, further testing on more natural and potentially conflicting real-world examples is necessary to fully understand its capabilities. The sample predictions generated shown in Figures 15 - 17 and the interactive demo developed in Figures 18 - 20 help to illustrate how the models behave and where they might occasionally make mistakes.

These visualizations confirm that it remains challenging for computer models to perfectly distinguish very similar emotions or to perform well when there are few examples of a particular emotion in the learning data. In summary, this research successfully identified effective methods for understanding emotions from single types of information by using expert teams and demonstrated that a two-step mixing approach (first creating expert teams, then mixing their results) shows great promise. This provides a solid foundation and clear directions for future work on the TERT-Ensemble project, particularly in refining how image, text, and sound information are combined and tested on more natural, everyday emotion examples.

CHAPTER 5
CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

This project undertook a detailed performance analysis of various advanced deep learning methods for recognizing human emotions independently from image, text, and audio data. This investigation served as a critical foundational phase for the larger TERT-Ensemble tri-modal system, which aims to understand emotions by integrating these three information sources. By systematically training and rigorously evaluating a range of models (including EfficientNetV2, ResNet50, ViT for images; DeBERTa, BERT, RoBERTa for text; and custom CNNs, LSTMs, and Transformer variants for audio) on combined, publicly available benchmark datasets within the reproducible Kaggle environment, robust performance baselines were established.

The initial stage focused on optimizing unimodal predictions through intra-modal ensembling, where predictions from multiple architectures within a single modality were combined. This approach yielded strong results: the image ensemble achieved a weighted F1-score of 77.98%, the text ensemble reached 73.69%, and the audio ensemble obtained 66.43%. Subsequently, a tri-modal late fusion model was implemented by combining the outputs of these three optimized intra-modal ensembles using a weighted averaging strategy. Evaluated on a simulated test set constructed with congruent emotional signals, this final hybrid model demonstrated high performance, achieving an accuracy of 95.56% and a weighted F1-score of 0.96, highlighting the potential of this fusion technique.

The study confirmed the relative strengths of different modalities for emotion recognition on the used datasets and underscored persistent challenges, such as handling class imbalance and accurately discriminating between similar emotional states. The practical applicability of the developed unimodal components was further illustrated through the creation of an interactive Gradio demonstration interface. Overall, this work provides valuable insights into unimodal model capabilities and delivers a strong empirical basis for advancing to the next stage of multimodal fusion research within the TERT-Ensemble framework.

5.2 Future Enhancement

While this research has established strong baseline models and demonstrated a promising fusion approach on simulated data, several avenues for future work can further enhance the TERT-Ensemble system and advance its capabilities for real-world applications:

1. **Advanced Multimodal Fusion Strategies:** The immediate next step involves exploring and implementing more sophisticated techniques to combine information from the optimized image, text, and audio intra-modal ensembles. Beyond the current weighted averaging of probabilities, this could include:

- Investigating other late fusion methods, such as more complex voting schemes or training a meta-classifier (stacking) on the unimodal ensemble outputs.
- Exploring early fusion by attempting to combine extracted features from different modalities at an earlier stage, though this presents challenges due to differing feature types and dimensions.
- Implementing hybrid or intermediate fusion techniques, potentially using attention mechanisms or dedicated multimodal transformer models designed to learn cross-modal interactions at deeper feature levels.

2. **Dataset Expansion and Diversity for Naturalistic Evaluation:** A critical path for improvement is the incorporation of larger and more diverse datasets, particularly those featuring spontaneous, real-world emotional expressions rather than primarily acted portrayals. Testing and potentially fine-tuning on such data would significantly improve the model's generalization and robustness. Addressing demographic biases within datasets by ensuring more varied representation is also an important consideration.
3. **Improved Handling of Data Imbalance:** While class weighting was employed, further exploration of advanced techniques could benefit performance on minority emotion classes. This includes methods like oversampling (e.g., SMOTE), undersampling, or the use of specialized loss functions (e.g., focal loss) specifically designed for imbalanced datasets.
4. **Exploration of Newer Unimodal Architectures:** The field of deep learning is rapidly evolving. Experimenting with newer or larger base models within each modality (such as more recent Vision Transformer variants or larger, more powerful language models) or exploring novel custom architectures could lead to even stronger individual predictors for the intra-modal ensembles.

5. **Refinement of Feature Engineering:**

- For the **audio** modality, investigating alternative feature representations, such as learning directly from raw waveforms using 1D CNNs, exploring different types of spectrograms (e.g., Constant-Q Transform), or applying more advanced feature selection techniques could yield better results.
- For **text**, exploring the use of sentiment-specific word embeddings or incorporating more extensive contextual information (e.g., preceding and succeeding sentences or conversation history) might enhance the understanding of emotional nuances.

6. **Rigorous Cross-Dataset Evaluation:** To obtain a more robust measure of the final fused model's real-world applicability, it should be rigorously tested on datasets completely unseen during any phase of training or validation. This type of cross-corpus evaluation is a key indicator of generalization.
7. **Enhanced Model Interpretability (XAI):** Implementing explainability techniques such as LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), or visualizing attention weights within the models could provide valuable insights into how both the unimodal and eventual multimodal models arrive at their predictions. This would increase trust in the system and facilitate its potential adoption in sensitive applications.
8. **Optimization for Deployment:** If the TERT-Ensemble system is intended for real-time applications, further work on model optimization techniques, such as quantization (reducing the precision of model weights), pruning (removing less important connections), and knowledge distillation, along with efficient deployment frameworks, would be necessary to ensure low latency and computational efficiency.

By systematically addressing these areas, with a particular focus on developing and validating effective fusion mechanisms on more naturalistic data, the TERT-Ensemble system can build upon the strong unimodal foundations established in this work, aiming to achieve state-of-the-art performance in comprehensive tri-modal emotion recognition.

REFERENCES

[1] R. W. Picard, *Affective computing*, Cambridge, MA, USA: MIT Press, 1997.

[2] A. Mehrabian, *Nonverbal communication*, London, UK: Routledge, 2017.

[3] Z. Li and S. Li, "Deep Facial Expression Recognition: A Survey," *arXiv preprint arXiv:1804.08348*, 2018.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017. doi: 10.1016/j.inffus.2017.02.003.

[5] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, May 2018. doi: 10.1145/3122821.

[6] G. Ramakrishnan and S. K. Saha, "A Survey on Multimodal Emotion Recognition using Deep Learning," *arXiv preprint arXiv:2103.03139*, 2021.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998. doi: 10.1109/5.726791.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/abs/1810.04805

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi: 10.1162/neco.1997.9.8.1735.

[11] G. Muhammad, M. Alhamid, M. Alsulaiman, and A. A. Al-Rodhaan, "EEG Signal Analysis for Detecting Depression Using Machine Learning Techniques: A Review," *IEEE Access*, vol. 9, pp. 114569-114580, 2021. doi: 10.1109/ACCESS.2021.3104846.

[12] A. Ochuba et al., "Leveraging Artificial Intelligence to Meet the Sustainable Development Goals," *ResearchGate*, Jan. 2024. [Online]. Available: https://www.researchgate.net/publication/377140131_Leveraging_artificial_intelligence_to_meet_the_sustainable_development_goals.

[13] S. K. D'Mello and A. Graesser, "Affect detection from spoken language," *Speech Communication*, vol. 53, no. 9-10, pp. 1160–1174, Nov.–Dec. 2011. doi: 10.1016/j.specom.2011.03.001.

[14] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996. doi: 10.1016/0031-3203(95)00069-5.

[15] R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019. doi: 10.1109/TPAMI.2017.2781237.

[16] E. Cambria, “Affective Computing and Sentiment Analysis,” *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016. doi: 10.1109/MIS.2016.31.

[17] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[18] P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-enhanced BERT with Disentangled Attention,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021. [Online]. Available: https://arxiv.org/abs/2006.03654

[19] F. A. Acheampong, W. Nunoo-Mensah, and A. Aggrey, “Transformer-Based Ensemble Model for Sentiment Analysis of COVID-19 Tweets,” *Applied Sciences*, vol. 11, no. 15, Art. no. 6752, Jul. 2021. doi: 10.3390/app11156752.

[20] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the Munich versatile and fast open-source audio feature extractor,” in *Proc. 18th ACM Int. Conf. Multimedia (MM '10)*, Florence, Italy, Oct. 2010, pp. 1459–1462. doi: 10.1145/1873951.1874246.

[21] M. Z. P. Ishaq, S. M. S. Ahmad, H. S. Munawar, and A. S. Abdullah, “Speech Emotion Recognition Using Deep Learning Techniques: A Review,” *IEEE Access*, vol. 9, pp. 109033–109056, 2021. doi: 10.1109/ACCESS.2021.3099602.

[22] G. Trigeorgis et al., “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5200–5204. doi: 10.1109/ICASSP.2016.7472669.

[23] X. Li, D. Tu, L. Zhang, X. Wang, and B. Xu, “SER-CNNs: An Ensemble of CNNs for Speech Emotion Recognition,” *IEEE Access*, vol. 8, pp. 101196–101205, 2020. doi: 10.1109/ACCESS.2020.2998342.

[24] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017. (Note: This is a duplicate of [4], usually you only list a reference once).

[25] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, “Early versus late fusion in semantic video analysis,” in *Proc. 13th ACM Int. Conf. Multimedia (MM '05)*, Singapore, Nov. 2005, pp. 399–402. doi: 10.1145/1101149.1101232.

[26] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Florence, Italy, Jul. 2019, pp. 6558–6569. doi: 10.18653/v1/P19-1644.

[27] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. Int. Conf. Machine Learning (ICML)*, PMLR, vol. 139, Jul. 2021, pp. 10096–10106.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[29] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2613–2617. doi: 10.21437/Interspeech.2019-2647.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015. [Online]. Available: \url{https://arxiv.org/abs/1409.0473}

[31] P. Lucey et al., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. 2010 IEEE Computer Soc. Conf. Computer Vision Pattern Recognition-Workshops (CVPRW)*, San Francisco, CA, USA, Jun. 2010, pp. 94–101. doi: 10.1109/CVPRW.2010.5543262.

[32] I. J. Goodfellow et al., "Challenges in representation learning: A report on the 2013 ICML workshop," *arXiv preprint arXiv:1312.6082*, 2013.

[33] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2252–2260. doi: 10.1109/CVPR.2017.293.

[34] J. D. Perez, "Emotion Detection from Text," Kaggle Dataset, 2020. [Online]. Available: \url{https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp} (Accessed: May 8, 2024)

[35] S. Saravia, H. C. C. Liu, Y. Huang, J. Wu, and Y. A. Chen, "CARER: Contextualized Affect Representations for Emotion Recognition," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, Brussels, Belgium, Nov. 2018, pp. 3687–3697. doi: 10.18653/v1/D18-1404.

[36] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014. doi: 10.1109/TAFFC.2014.2336244.

[37] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions," *PLoS ONE*, vol. 13, no. 5, p. e0196391, May 2018. doi: 10.1371/journal.pone.0196391.

[38] S. Haq and P. J. B. Jackson, “Speech emotion recognition using spectrogram & phoneme embedding,” in *Proc. 2009 Ninth IEEE Int. Conf. Data Mining Workshops (ICDMW '09)*, Miami, FL, USA, Dec. 2009, pp. 41–46. doi: 10.1109/ICDMW.2009.50.

[39] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019, vol. 32.

[40] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP): System Demonstrations*, Online, Oct. 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[41] H. Yang et al., “Torchaudio: An audio library for PyTorch,” in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 7482–7486. doi: 10.1109/ICASSP43922.2022.9747911.

[42] B. McFee et al., “librosa: Audio and music signal analysis in python,” in *Proc. 14th Python Sci. Conf.*, Austin, TX, USA, 2015, vol. 8, pp. 18–25. doi: 10.25080/Majora-7b98e3ed-003.

[43] W. McKinney, “Data Structures for Statistical Computing in Python,” in *Proc. 9th Python Sci. Conf.*, Austin, TX, USA, 2010, pp. 56–61. doi: 10.25080/Majora-92bf1922-00A.

[44] C. R. Harris et al., “Array programming with NumPy,” *Nature*, vol. 585, pp. 357–362, Sep. 2020. doi: 10.1038/s41586-020-2649-2.

[45] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[46] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May/Jun. 2007. doi: 10.1109/MCSE.2007.55.

[47] M. L. Waskom, “Seaborn: statistical data visualization,” *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021. doi: 10.21105/joss.03021.

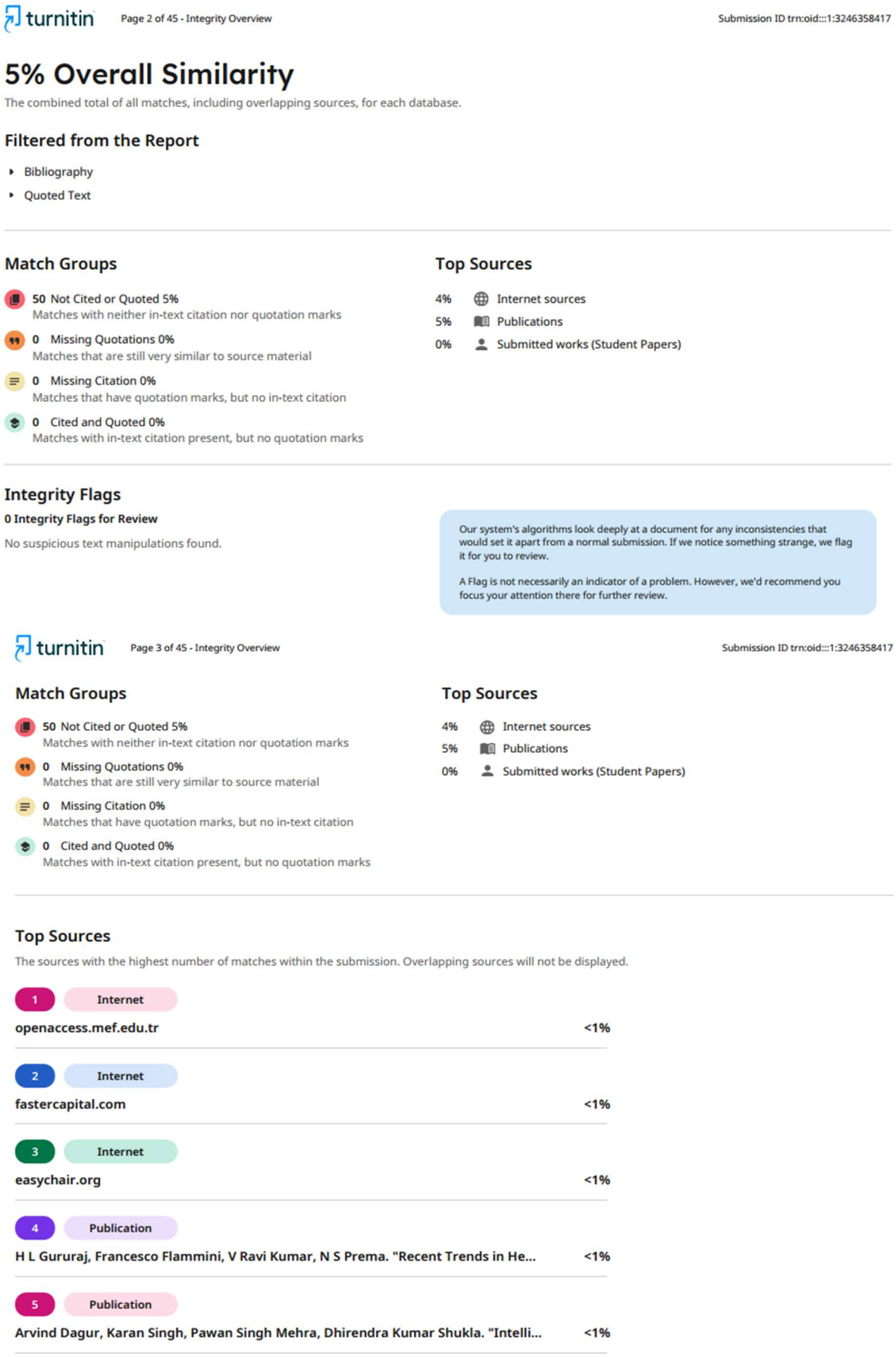
[48] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019.

[49] L. N. Smith, “A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay,” *arXiv preprint arXiv:1803.09820*, 2018.

[50] A. Buslaev et al., “Albumentations: fast and flexible image augmentations,” *Information*, vol. 11, no. 2, p. 125, Feb. 2020. doi: 10.3390/info11020125.

[51] A. Abid et al., “Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild,” *arXiv preprint arXiv:1906.02569*, 2019.

APPENDIX



turnitin Page 2 of 45 - Integrity Overview Submission ID trn:oid:::1:3246358417

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- 50 Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
5% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

turnitin Page 3 of 45 - Integrity Overview Submission ID trn:oid:::1:3246358417

Match Groups

- 50 Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
5% Publications
0% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
openaccess.mef.edu.tr		<1%
2	Internet	
fastercapital.com		<1%
3	Internet	
easychair.org		<1%
4	Publication	
H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in He...		<1%
5	Publication	
Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelli...		<1%

PUBLICATION

International Journal of Innovative Research in Technology

An International Open Access Journal Peer-reviewed, Refereed Journal

www.ijirt.org | editor@ijirt.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of International Journal of Innovative Research in Technology (ISSN 2349-6002) is hereby awarding this certificate to

PENUMUCHU NIHITH

In recognition of the publication of the paper entitled

TERT-ENSEMBLE: A TWO-STAGE FUSION APPROACH FOR TRI-MODAL EMOTION RECOGNITION

Published in IJIRT (www.ijirt.org) ISSN UGC Approved (Journal No: 47859) & 8.01 Impact Factor

Published in Volume 11 Issue 12, May 2025

Registration ID 178452 Research paper weblink:https://ijirt.org/Article?manuscript=178452

EDITOR EDITOR IN CHIEF

International Journal of Innovative Research in Technology

An International Open Access Journal Peer-reviewed, Refereed Journal

www.ijirt.org | editor@ijirt.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of International Journal of Innovative Research in Technology (ISSN 2349-6002) is hereby awarding this certificate to

V LINGESHWARAN

In recognition of the publication of the paper entitled

TERT-ENSEMBLE: A TWO-STAGE FUSION APPROACH FOR TRI-MODAL EMOTION RECOGNITION

Published in IJIRT (www.ijirt.org) ISSN UGC Approved (Journal No: 47859) & 8.01 Impact Factor

Published in Volume 11 Issue 12, May 2025

Registration ID 178452 Research paper weblink:https://ijirt.org/Article?manuscript=178452

EDITOR EDITOR IN CHIEF