

kit

by Kit Kit

Submission date: 21-May-2026 09:14AM (UTC+0530)

Submission ID: 2966105863

File name: pulkit_24swe17.pdf (1,020.93K)

Word count: 9170

Character count: 56354

A MULTIMODAL DEEP LEARNING SYSTEM FOR ENHANCED EMOTION DETECTION

10
THESIS SUBMITTED

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF

MASTER OF TECHNOLOGY

in

SOFTWARE ENGINEERING

by

PULKIT JAIN

(24/SWE/17)

Under the supervision of
Dr. SONIKA DAHIYA
(ASSISTANT PROFESSOR)



3
DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi 110042

MAY, 2026

3
DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Dr. Sonika Dahiya for her continuous guidance and mentorship throughout this project. She provided me with a clear direction to achieve my research targets by explaining all the tasks to be accomplished and emphasising the importance of this work as well as its relevance to real-world applications. She was always available to help me and resolve my doubts regarding any challenges encountered during the course of this project. Without her constant support and motivation, this work would not have been possible.

I would also like to express my gratitude to the Department of Software Engineering of Delhi Technological University for providing me all the necessary things as well as academic environment which helped me to complete this research. Training of the deep learning models discussed in this thesis would not have been possible without the use of the GPU enabled computing facilities.

I also recognize the originators of the IEMOCAP dataset, and the creators of the open source tools that this thesis relies on: PyTorch, Hugging Face Transformers, OpenAI Whisper, and others.

Place: Delhi

Date:

Pulkit Jain
(2K24/SWE/17)

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

CANDIDATE'S DECLARATION

I, **Pulkit Jain**, Roll No. **24/SWE/17**, student of M.Tech (Department of Software Engineering), hereby certify that the work which is being presented in the thesis entitled "A Multimodal Deep Learning System for Enhanced Emotion Detection" in partial fulfilment of the requirements for the award of degree of Master of Technology, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from Jan 2025 to May 2026 under the supervision of Dr. Sonika Dahiya.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

CERTIFICATE BY THE SUPERVISOR

I hereby certify that the Project Dissertation titled "A Multimodal Deep Learning System for Enhanced Emotion Detection" which is submitted by **Pulkit Jain**, Roll No. **24/SWE/17**, Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Dr. Sonika Dahiya
Assistant Professor

Date:

Department of Software Engineering, DTU

ABSTRACT

Detecting emotions from multimodal data is increasingly significant in various domains such as mental health support, customer experience enhancement, and human-computer interaction. As digital communication grows, understanding the emotional tone of conversational content has become essential for building responsive and human-centric applications. This study explores a multimodal deep learning system that aims to improve the accuracy and class-fairness of emotion detection by fusing visual, textual, and audio modalities.

In-depth analysis is conducted on a multimodal architecture that combines ResNet-18 for visual features with a hybrid text stream incorporating RoBERTa, FastText, and the NRC Emotion Lexicon. Audio is selectively re-routed through OpenAI Whisper transcription to augment the text stream. The model is evaluated on the IEMOCAP dataset—a widely used benchmark for multimodal emotion recognition—chosen for its diversity in conversational language and emotion categories. A three-layer MLP classifier on the 928-dimensional fused descriptor achieves 86.48% accuracy and Macro F1-score of 0.8690, a statistically significant improvement ($Z=16.94$, $p < 0.05$) over unimodal baselines.

The aim of this research work is to demonstrate that pre-fusion linguistic enrichment paired with class-fair evaluation closes a substantial portion of the gap left open by prior IEMOCAP multimodal work. The results of experiments indicate that pre-fusion enrichment of the linguistic features is more effective than complex fusion schemes based on a transformer network. This study not only unveils the potential of hybrid feature engineering but also reveals where more can be gained, the keys to even more sophisticated and sensitive multimodal emotion recognition system.

TABLE OF CONTENTS

Acknowledgement	i
Candidate's Declaration	ii
Certificate	iii
Abstract	iv
List of Abbreviations	i
1 Introduction	2
1.1 Background and Motivation	2
1.2 Problem Statement	2
1.2.1 Unimodal Limits and the IEMOCAP Plateau	2
1.2.2 Under-Enriched Text Features and Unfair Aggregate Metrics	2
1.3 Research Questions and Objectives	3
1.4 Contributions of the Thesis	3
1.4.1 Hybrid Pre-Fusion Text Enrichment (RoBERTa + FastText + NRC)	3
1.4.2 Selective Audio-to-Text Re-routing via Whisper	4
1.4.3 Class-Fair Evaluation Protocol with Statistical Significance Testing	4
1.5 Scope and Limitations of the Study	4
1.6 Organization of the Thesis	4
2 Literature Review	5
2.1 Theoretical Foundations of Emotion	5
2.2 Unimodal Emotion Recognition Approaches	5
2.2.1 Visual and Acoustic Approaches	5
2.2.2 Textual Approaches and Cross-Modal Limitations	5
2.3 Multimodal Fusion Strategies	6
2.3.1 Early and Late Fusion	6
2.3.2 Attention-Based and Cross-Modal Fusion	6
2.3.3 Summary of Fusion Approaches	6
2.4 Linguistic Backbones for Emotion	6
2.4.1 Transformer and Embedding Models	6

2.4.2	Lexicon-Based Approaches.....	7
2.5	Datasets for Multimodal Emotion Recognition	7
2.6	Evaluation Methodology in Imbalanced Multi-Class Settings.....	7
2.7	Research Gap and Positioning of This Work.....	8
3	Methodology and Proposed System Design	9
3.1	Research Design Overview.....	9
3.2	Dataset Description.....	10
3.2.1	IEMOCAP Composition and Class Selection	10
3.2.2	Data Splitting Strategy	10
3.3	Pre-processing and Data Alignment	11
3.3.1	Audio and Text Pre-processing	11
3.3.2	Image Extraction and Multimodal Alignment.....	11
3.4	Unimodal Audio Model.....	11
3.5	Unimodal Text Model (Hybrid Linguistic Stream).....	12
3.6	Unimodal Image Model.....	13
3.7	Multimodal Fusion Architecture.....	13
3.7.1	Two-Track Pipeline and Pre-Fusion Enrichment.....	13
3.7.2	928-d Descriptor and MLP Classifier.....	14
3.8	Mathematical Formulation.....	14
3.8.1	Macro-Averaged Classification Metrics	14
3.8.2	Two-Sample Two-Proportion Z-Test.....	15
4	Implementation	16
4.1	Hardware and Software Environment.....	16
4.2	Data Pipeline Implementation	16
4.2.1	Audio Processing and Whisper Transcription	16
4.2.2	Face Extraction and Multimodal Alignment	17
4.3	Model Implementation Details	17
4.4	Training Configuration	18
4.5	Computational Profile and Reproducibility	19
5	Results and Analysis	20
5.1	Pre-processing Outcomes	20
5.2	Unimodal Baseline Results.....	20
5.3	Multimodal Optimization	21
5.3.1	Optimizer and Depth Search	21
5.3.2	Best Configuration	21
5.4	Role and Contribution of Each Modality	21
5.5	Class-wise Performance under Imbalance	22

5.5.1	Per-Class Metrics	22
5.5.2	Confusion Matrix Analysis	23
5.6	Statistical Significance Analysis	23
5.7	Comparison with Recent Multimodal Transformer Approaches	24
5.8	Discussion of Findings	24
6	Conclusion, Limitations and Future Scope	25
6.1	Conclusion	25
6.2	Limitations of the Proposed Study	26
6.3	Future Scope	27
	Bibliography	28
	Appendix	32
I.	List of Publications	32
II.	Paper Acceptance Proof	33
III.	Indexing of Conference Proof	34
IV.	Conference Paper Registration Receipt	35

LIST OF FIGURES

3.1	General research design flowchart showing the end-to-end pipeline from raw IEMOCAP data through pre-processing, unimodal baselines, multimodal fusion, optimisation, and evaluation.	10
3.2	Unimodal audio model architecture: Mel-spectrograms are processed through a 2D CNN with attention and classified by an MLP	12
3.3	Unimodal text model architecture: three parallel streams (RoBERTa + 1D CNN, FastText + FC, NRC Lexicon) are concatenated into a 256-d vector and classified by an MLP	12
3.4	Unimodal image model architecture: ResNet-18 extracts 512-d face embeddings, refined by self-attention before Softmax classification.	13
3.5	Multimodal model architecture: two parallel tracks (visual and linguistic) are fused through concatenation into a 928-d descriptor, classified by a three-layer MLP	14
5.1	Confusion matrix of the final optimised multimodal model (AdamW + 3-layer MLP). Strong diagonal dominance indicates high per-class accuracy across all five emotions.....	23

LIST OF TABLES

3.1	Dataset distribution across modalities after filtering to five emotions. . .	11
4.1	Software stack with version numbers.	16
4.2	Text module stream specifications.	17
4.3	Multimodal MLP classifier architecture.	18
4.4	Training hyperparameters for each model.	18
4.5	Computational profile of each model on Tesla T4 GPU.	19
5.1	Consolidated results: unimodal baselines through optimised multimodal framework.....	21
5.2	Per-modality contribution to the final 5-class prediction.	22
5.3	Per-class Precision, Recall, F1, and support for the final optimised multimodal model (AdamW + 3-layer MLP).	22
5.4	Comparison with recent multimodal emotion recognition methods on IEMOCAP. A = audio, V = visual, T = text.....	24
6.1	Summary of thesis contributions and their quantitative impact.	25

LIST OF ABBREVIATIONS

ASR	⁴ Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DTU	Delhi Technological University
FC	Fully Connected
HCI	¹⁷ Human–Computer Interaction
IEMOCAP	Interactive Emotional Dyadic Motion Capture
LSTM	Long Short-Term Memory
MELD	Multimodal Emotion Lines Dataset
MFCC	Mel-Frequency Cepstral Coefficient
MLP	Multilayer Perceptron
NRC	National Research Council (Canada)
ReLU	Rectified Linear Unit
TF-IDF	Term Frequency–Inverse Document ⁹ Frequency

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Affective computing is the development of systems which are able to recognize, interpret, and respond to human's emotions [45]. In the realm of human-computer interaction (HCI) [7–9] affecting is a fundamental part of human interaction with machines, such as assistants, vehicles, or healthcare monitors. Progress has been made at a faster pace with deep learning [10] – especially in the area of faces, text and speech – which has fueled applications ranging from mental health monitoring [8] to adaptive services and safety.

Emotional expression is expressed in many ways: facial expressions [1, 2], prosody [3, 4] and words [5, 6]. An incomplete channel: expressions can be masked, audio is noisy and speaker-specific, text omits nonverbal cues, etc., means that a single modality cannot solve all of the problems like masked frustration and sarcastic praise. These are offset by fusing together the visual, acoustic and lexical information.

People are increasingly expecting to get more than just intent – they expect to get feeling. The most common solution is multimodal fusion [16–18] which this thesis demonstrates is dominantly affected by the inputs to fusion, as well as the fusion rule itself.

1.2 Problem Statement

1.2.1 Unimodal Limits and the IEMOCAP Plateau

There is much work that has gone before which is unimodal in nature, such as visual [1, 11], auditory [12, 13] and textual [14, 15] work, and each of these areas has a ceiling. The deficit of speech-based affect in visual models, and the deficit of lexico-semantic cues in acoustic-only models (our Mel-spectrogram baseline is 32.04% on IEMOCAP), along with the inability to access faces and prosody from text alone, leaves affect ambiguous. Multi-modal scores have remained in a tight range on IEMOCAP (e.g., [21–23]) and the new cross-modal attention systems [19, 20] are still in the mid-70% range. The most important question is why this plateau is still there and what needs to happen to overcome it.

1.2.2 Under-Enriched Text Features and Unfair Aggregate Metrics

Linguistic features at fusion are often weak, as many pipelines pass one transformer embedding into fusion, but IEMOCAP's dialogue is fragmented and colloquial, and rich cues are distributed across and cross multiple linguistic layers: contextual

meaning, sub-word morphology, and lexicon-based affect, with few pipelines covering all three in one; few pipelines cover all three. Headline accuracy hides class imbalance (e.g. “Frustrated” $\approx 1.9\times$ “Sad”); high overall scores can go hand in hand with low minority-class scores. As a result, several previous studies lack the necessary macro-averaged measures, per-class views, and significance tests in making sound comparisons.

¹⁹ 1.3 Research Questions and Objectives

This thesis investigates the following research questions:

RQ1:

What is the added value of pre-fusion enrichment (RoBERTa + FastText + NRC Emotion Lexicon) in terms of accuracy over a single text encoder?

RQ2:

What is the contribution from each modality (visual, textual, audio-via-transcription) to the fused prediction?

RQ3:

Are there data on performance balance among classes with unequal distribution of students and are there statistically significant improvement from base-lines?

Objectives:

1. Train/evaluate unimodal baselines: 2D CNN (Mel-spectrograms), hybrid RoBERTa + FastText + NRC text, ResNet-18 + self-attention image on IEMOCAP.
2. Create the hybrid pre-fusion text source (three sources into a single enriched vector).
3. Integrate audio using Whisper's selective transcription.
4. Fuse via concatenation early fusion and an MLP on the 928-dimensional descriptor.
5. Optimizer search (Adam, RAdam, Lion, AdamW), MLP depth (2–4).
6. Report macro P/R/F1, per-class metrics, confusion matrices, and two-proportion Z-tests.

1.4 Contributions of the Thesis

1.4.1 Hybrid Pre-Fusion Text Enrichment (RoBERTa + FastText + NRC)

One of the biggest contributions is a three-source text module that is combined before the fusion with the vision module, called RoBERTa for context, FastText for sub-word coverage, and NRC for explicit affect priors. Prior multimodal emotion work rarely unifies these streams; here they yield the largest single gain (+6.20 pp).

1.4.2 Selective Audio-to-Text Re-routing via Whisper

Audio is fed through Whisper transcription instead of raw spectrograms, shifting its role from low acoustic accuracy (32.04% alone) to semantic help (+1.28 pp). For conversational data, semantic routing outperforms raw prosody in this setup.

1.4.3 Class-Fair Evaluation Protocol with Statistical Significance Testing

Evaluation pairs macro P/R/F1, per-class breakdowns, and confusion matrices with a two-proportion Z-test under a fully specified protocol, countering aggregate-only reporting in prior IEMOCAP studies.

24

1.5 Scope and Limitations of the Study

This thesis covers five-class emotion recognition (angry, excited, frustrated, neutral, sad) on the IEMOCAP database using audio, text, and image modalities with a concatenation-based early fusion strategy and an MLP classifier. The study is conducted on a single dataset with session-based train/test splitting to ensure speaker independence.

The thesis does *not* cover: (a) attention-based or transformer-driven fusion mechanisms (these are identified as promising future directions); (b) datasets beyond IEMOCAP (generalization to in-the-wild corpora like MELD and CMU-MOSEI is deferred to future work); (c) real-time deployment or latency-constrained systems; (d) emotion categories beyond the five selected; or (e) multilingual or cross-cultural emotion recognition. A fuller discussion of limitations appears in Chapter 6.

5

1.6 Organization of the Thesis

The remainder proceeds as follows: Chapter 2 reviews emotion theory, unimodal and multimodal recognition, linguistic models, benchmarks, and evaluation, and pinpoints the research gap; Chapter 3 presents the research design, dataset, pre-processing, unimodal architectures, fusion with pre-fusion text enrichment, and metrics; Chapter 4 covers implementation and reproducibility; Chapter 5 reports baselines, multimodal optimisation, modality contributions, class-wise results, significance tests, and comparisons; Chapter 6 summarises contributions, answers the research questions, and outlines future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Theoretical Foundations of Emotion

Categorical models treat affect as a small inventory of discrete labels (e.g., Ekman's basic emotions; Plutchik's wheel). That convention is followed by corpus annotation as well, with the IEMOCAP having categorical tags, and this thesis having five classes, but neighbouring categories are hard to distinguish even for humans.

Dimensional models are models that embed emotions into continuous spaces: Russell's valence–arousal circumplex and Mehrabian's pleasure–arousal–dominance (PAD) model account for why high arousal negative states seem to merge in the feature space, and why later confusion matrices display systematic mixing between such classes.

Despite this, automatic recognition is difficult: Labels are subjective, expressions are ambiguous and context dependent, and there are variations in display rules and intensity between people. These effects limit reliable supervision and prevent unimodal recognition from becoming overly noisy, encouraging multimodal fusion but not ensuring that fusion eliminates all inherent uncertainty.

2.2 Unimodal Emotion Recognition Approaches

2.2.1 Visual and Acoustic Approaches

From hand-created texture and shape descriptors, to deep CNNs (VGG, Inception, ResNet [1, 11, 42]), to attention-enhanced encoders [24] and to spatiotemporal models with added compute for temporal expression dynamics. In dyadic conversation, affect is nevertheless lost in images or video that is conveyed in lexicon and prosody. In terms of audio, common stacks are based on engineered descriptors (MFCCs, pitch, energy), Mel-spectrogram CNNs [3] or self-supervised encoders like Wav2Vec 2.0 [13]; speech and text classification outperform raw acoustics for fine-grained classes [12]. On the empirical level, acoustic-only cues provide only limited separability on this task (32.04% of the reported audio baseline), highlighting the importance of semantic content for this task as well as for timbre.

2.2.2 Textual Approaches and Cross-Modal Limitations

Together with classic bag of words and TF-IDF [7] models, the methods have progressed to static embeddings (Word2Vec, FastText [40]) and contextual transformers, with RoBERTa [39] frequently topping transformer models on affective benchmarks [25–28]. Each modality is only partial: vision – facial cues, audio – prosody (without full meaning if not transcribed), and text – omits face and voice. That complementarity encourages fusion rather than any one stream in isolation.

2.3 Multimodal Fusion Strategies

2.3.1 Early and Late Fusion

Feature-level *early* fusion concatenates modality vectors before classification [15], enabling a single learner to exploit joint statistics but ignoring fine-grained temporal alignment and mixing possibly redundant dimensions; this thesis uses concatenation with an MLP. *Late* fusion trains per-modality classifiers and aggregates decisions, which eases missing modalities but cannot model correlations that live only in jointly observed features (e.g., sarcasm from mismatched face and words). Hybrid architectures combine both patterns at the price of complexity. Simple early fusion can still be competitive when incoming representations are rich rather than minimal.

2.3.2 Attention-Based and Cross-Modal Fusion

Richer fusion stacks—residual multimodal networks [21], multi-head transformer fusion [22], cross-modal attention such as AVaTER [19], pretrained per-modality transformers fused via another transformer [20], and information-enhanced audio-text designs [23]—explicitly target cross-modal interactions. Reported IEMOCAP scores range from the low 60%s for several structural variants up to the mid-70%s when strong linguistic encoders and attention are present, implying that representation quality (especially text) often dominates raw fusion novelty.

2.3.3 Summary of Fusion Approaches

Across this literature, sophisticated operators alone rarely overcome weak features; gains track better semantic encoders and task-appropriate inputs. The empirical chapters therefore emphasise pre-fusion text enrichment alongside the chosen early-fusion classifier instead of only deepening fusion topology.

2.4 Linguistic Backbones for Emotion

2.4.1 Transformer and Embedding Models

Transformer encoders (BERT, RoBERTa [39], XLNet, DistilBERT) supply contextual representations that define the modern affective NLP stack; comparative studies typically give RoBERTa an edge for emotion classification [25–28]. FastText [40] adds sub-word n-gram composition for informal or rare surface forms, complementing

RoBERTa’s WordPiece coverage on disfluent transcripts. The implementation later concatenates RoBERTa-derived features with FastText vectors as parallel linguistic streams.

2.4.2 Lexicon-Based Approaches

The NRC Emotion Lexicon [41] maps vocabulary items to categorical affects and sentiment, producing compact utterance profiles without gradient learning and informing knowledge-graph style enrichments for dialogue [33]. Contextual models, static embeddings, and lexicon lookups capture different failure modes, so prior work rarely stacks all three before fusion [29–32]; the proposed hybrid text module exploits exactly that gap.

2.5 ²⁰ Datasets for Multimodal Emotion Recognition

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [43] was used as a benchmark: it consists of 12 hours of dyadic (actor-actor) scripted and spontaneous interactions between 10 actors, with each enactment having sentence-aligned audio, video, manual text, nine categories and dimensional annotations (valence, arousal, dominance). Majority vote labels allow reproducible splits, and add ambiguity where moderate imbalance favours some affective states. IEMOCAP is the setting where they want to compare the fusion literature, and it is justifiable on several grounds: Dyadic structure, synchronized modalities, and extensive prior reporting. In addition, related resources, such as MELD (television dialogue), CMU-MOSEI (YouTube monologues with sentiment/emotion tags), and RAVDESS (acted audio–vision without rich transcripts), exist that show other acoustic and visual distributions, with generalisation challenges beyond the studio captured resources, but are not included in the experimental study and are only presented here as a framing device for future extension (Chapter 6).

2.6 Evaluation Methodology in Imbalanced Multi-Class Settings

Imbalanced emotion corpora require metrics beyond accuracy: overall accuracy overweights head classes, whereas macro-averaged precision, recall, and F1 treat each class equally [34, 35]; weighted averages interpolate by class support and can hide minority underperformance. Per-class precision, recall, F1, and confusion matrices expose which affective categories merge (e.g., aroused negatives). Headline results therefore foreground macro F1 alongside accuracy and weighted summaries, and pairwise model comparisons use two-proportion Z-tests on matched test outcomes [36–38] (with McNemar’s or bootstrap methods as common alternatives when reporting correlates of decisions).

2.7 Research Gap and Positioning of This Work

Prior multimodal IEMOCAP systems span the low-to-mid 60%s with classical fusion stacks through roughly 75% when strong transformers and cross-modal attention are used, yet many pipelines still inject a single linguistic encoder, report headline accuracy only, or omit significance testing, and few dissect semantic audio (ASR text) versus raw acoustics inside fusion [12]. The chapters ahead therefore (i) stack RoBERTa, FastText, and NRC before image-level fusion, (ii) pair macro metrics, per-class tables, and Z-tests with leave-one-modality studies, and (iii) document how selective audio-to-text integration shifts performance relative to acoustic baselines, closing the empirical and reporting gaps noted above.

CHAPTER 3 METHODOLOGY AND PROPOSED SYSTEM DESIGN

3.1 Research Design Overview

This thesis follows an experimental quantitative design: controlled unimodal baselines (audio, text, image) isolate modality-specific behaviour, then a multimodal fusion model integrates all streams and we measure the incremental effect of each design choice. Dependent variables are classification accuracy and Macro F1; independent variables include modality combination (audio-only, text-only, image-only, image+text, image+text+audio→text), optimiser (Adam, RAdam, Lion, AdamW), MLP depth (2–4 layers), and text encoder type (1D CNN vs. LSTM). We apply *selective integration*—rather than fusing raw Mel-spectrograms, audio is transcribed with OpenAI Whisper [44] and merged with the primary transcript stream. Pilot runs showed lexical features derived from speech outperformed raw acoustics in the shared MLP head; on conversational IEMOCAP data, what is said often dominates fine-grained discrimination relative to prosody alone, maximising useful information per fused dimension.

Figure 3.1 summarises the workflow: raw IEMOCAP inputs, modality-specific pre-processing, unimodal baselines, multimodal fusion with pre-fusion text enrichment, optimiser and depth search, then class-fair evaluation and statistical testing.

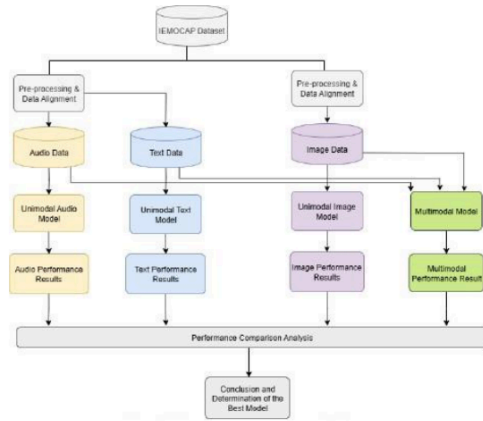


Figure 3.1: General research design flowchart showing the end-to-end pipeline from raw IEMOCAP data through pre-processing, unimodal baselines, multimodal fusion, optimisation, and evaluation.

3.2 Dataset Description

3.2.1 IEMOCAP Composition and Class Selection

IEMOCAP [43] is the sole benchmark: five sessions of dyadic conversation among ten speakers (five male, five female), roughly twelve hours of data. Each session provides time-aligned 16 kHz WAV audio, multi-view upper-body video (AVI), and manual transcripts with speaker tags. Scripted and improvised segments mix posed and semi-spontaneous affect. Of nine annotated categories, four (fearful, surprised, disgusted, happy) fall below roughly five hundred usable samples after filtering; we retain five well-represented classes—**frustrated** (1,848), **angry** (1,205), **neutral** (1,201), **excited** (1,153), and **sad** (978)—spanning valence and arousal.

3.2.2 Data Splitting Strategy

Splits are *by session* (80:20): Sessions 1–4 train (90:10 train–validation internally) and Session 5 test. Each session uses a distinct speaker pair, so no speaker appears in both train and test—avoiding identity-driven shortcuts in voice or face—and the headline 86.48% reflects unseen-speaker generalisation. Training batches use stratified sampling to preserve class ratios. For images, unconstrained frame extraction would skew

frequency toward expressive classes; we cap at 5,000 faces per emotion via stratified speaker-level sampling (25,000 images total). Table 3.1 summarises counts after filtering.

Table 3.1: Dataset distribution across modalities after filtering to five emotions.

Emotion	Audio	Text	Image*	Multimodal
Frustrated	1,848	1,857	5,000	3,471
Angry	1,205	1,209	5,000	3,957
Neutral	1,201	1,208	5,000	2,582
Excited	1,153	1,171	5,000	2,268
Sad	978	981	5,000	2,620
Total	6,385	6,426	25,000	14,898

*Capped at 5,000 per category via stratified speaker-level sampling.

3.3 Pre-processing and Data Alignment

3.3.1 Audio and Text Pre-processing

Whisper [44] (“base”) transcribes each segment to text with timestamps; audio is re-sampled to 16 kHz, segments under 0.6 s are dropped, and Mel-spectrograms (128 Mel bands, 25 ms window, 10 ms hop, dB normalisation) feed the audio CNN. Transcripts are cleaned by removing session IDs, timestamps, speaker tags, and bracketed non-verbal cues; RoBERTa BPE tokenisation truncates at 128 tokens, while the same surface string is whitespace-tokenised for FastText lookup.

3.3.2 Image Extraction and Multimodal Alignment

MTCNN detects and aligns faces (five landmarks), sampling every fifth frame, cropping to the standard 224×224 ResNet input; the per-class 5,000-image cap matches Section 3.2. Dialogue IDs align utterances across modalities; we keep one face per utterance (frame nearest the utterance midpoint), yielding 14,898 synchronised image-text pairs.

3.4 Unimodal Audio Model

The audio pathway (Figure 3.2) treats dB-normalised Mel-spectrograms (128 bands) as 2D inputs, trading raw waveforms for a compact perceptual map with more temporal detail than MFCCs for convolutional processing. Three stacked blocks apply Conv2D ($1 \rightarrow 32$, $32 \rightarrow 64$, $64 \rightarrow 128$ channels, 3×3 kernels), each followed by Batch Normalisation, ReLU, and 2×2 max pooling, yielding spectro-temporal patterns robust to absolute

loudness.

Learned softmax attention weights the temporal slices of the final feature map; the attended vector feeds a two-layer MLP with ReLU, dropout, and a five-way Softmax. The standalone audio CNN has ~ 341 K parameters, runs at 45 s per epoch on a Tesla T4, and infers in 12 ms per sample.

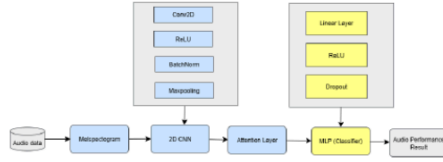


Figure 3.2: Unimodal audio model architecture: Mel-spectrograms are processed through a 2D CNN with attention and classified by an MLP.

3.5 Unimodal Text Model (Hybrid Linguistic Stream)

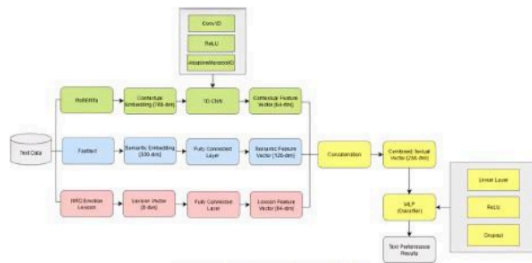


Figure 3.3: Unimodal text model architecture: three parallel streams (RoBERTa + 1D CNN, FastText + FC, NRC Lexicon) are concatenated into a 256-d vector and classified by an MLP.

Figure 3.3 outlines the hybrid encoder. Cleaned utterances pass through roberta-base [39], producing 768-d token embeddings; instead of [CLS]-only pooling, multi-width Conv1D filters (lengths 3–5) with ReLU and AdaptiveMaxPool1D skim local n-grams that often encode affect, yielding a fixed-length representation richer than a single pooled token for short, conversational text.

Pre-trained cc.en.300.bin supplies 300-d FastText embeddings per token; utterance vectors are mean-pooled and linearly projected to 128 dimensions, with character n-gram composition covering contractions and colloquial out-of-vocabulary forms common in IEMOCAP.

The NRC Emotion Lexicon [41] maps each word to eight emotions plus positive/negative sentiment; per-dimension scores are summed over the utterance and passed through a small FC for scale alignment. The three streams are concatenated and projected to a 256-d embedding reused in fusion; a two-layer ReLU/dropout MLP performs five-way classification. The module holds $\sim 125\text{M}$ parameters (RoBERTa-dominated), trains in 102 s per epoch, and infers in 15 ms on a Tesla T4.

3.6 Unimodal Image Model

Image inputs (Figure 3.4) use ResNet-18 [42] with ImageNet weights: early residual blocks stay frozen for low-level structure, upper blocks fine-tune on aligned 224×224 face crops. Global average pooling yields a 512-d embedding; a lightweight self-attention reweights feature dimensions to emphasise expression-relevant cues, followed by a linear layer and Softmax over five classes.



Figure 3.4: Unimodal image model architecture: ResNet-18 extracts 512-d face embeddings, refined by self-attention before Softmax classification.

The unimodal image network totals 11.1M parameters, trains in 35 s per epoch, and achieves 5 ms inference latency; the 5,000-image-per-class cap plus stratified speaker sampling mitigates class skew.

3.7 Multimodal Fusion Architecture

3.7.1 Two-Track Pipeline and Pre-Fusion Enrichment

Figure 3.5 shows parallel visual and linguistic tracks: ResNet-18 maps each face crop to 512 dimensions, while Section 3.5’s hybrid encoder processes transcripts. Unlike pipelines that supply a single transformer vector, RoBERTa, FastText, and NRC are composed *before* visual concatenation, giving concatenation-based fusion a linguistically rich manifold (contextual semantics, sub-word coverage, lexicon priors) rather than relying on a more intricate fusion operator alone.

For the full multimodal system, audio is not represented by Mel maps at fusion time:

Whisper [44] augments the text stream, so audio contributes lexical content under the selective-integration rationale (Section 3.1). The integrated model has 136.3M parameters, trains in 153 s per epoch, and achieves 21 ms inference latency on a Tesla T4.

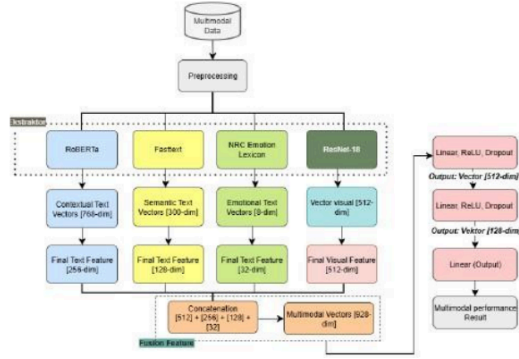


Figure 3.5: Multimodal model architecture: two parallel tracks (visual and linguistic) are fused through concatenation into a 928-d descriptor, classified by a three-layer MLP.

3.7.2 928-d Descriptor and MLP Classifier

Concatenating ResNet features with the RoBERTa-CNN, FastText, and NRC branches yields

$$\mathbf{f}_{\text{fused}} = [\mathbf{f}_{\text{ResNet}} || \mathbf{f}_{\text{RoBERTa-CNN}} || \mathbf{f}_{\text{FastText}} || \mathbf{f}_{\text{NRC}}] \in \mathbf{R}^{928} \quad (3.1)$$

($||$ denotes stacking), i.e., $512 + 256 + 128 + 32$ dimensions. A three-layer MLP with ReLU and dropout implements $928 \rightarrow 512 \rightarrow 256 \rightarrow 128$, capped by $128 \rightarrow 5$ Softmax. Chapter 5 shows three hidden layers strike the best bias–variance trade-off for this descriptor: two layers underfit cross-modal mixing, while four layers add capacity without accuracy gains.

3.8 Mathematical Formulation

3.8.1 Macro-Averaged Classification Metrics

Per-class precision, recall, and F1 follow the usual definitions from the confusion matrix; macro-precision, macro-recall, and macro-F1 are unweighted averages over the $C = 5$ classes, giving minority categories (e.g., sad at 978 samples) equal influence to

majority ones (e.g., frustrated at 1,848) in the headline scores reported alongside accuracy.

3.8.2 Two-Sample Two-Proportion Z-Test

Pairwise accuracy comparisons use a two-sample two-proportion Z-test with pooled variance [36–38]:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad (3.2)$$

where $\hat{p}_1, \hat{p}_2$ are the observed accuracies, x_k the numbers of correct predictions, n_k the paired test counts ($n_1 = n_2 = 14,898$ here), and $\hat{p}$ the pooled proportion under $H_0: p_1 = p_2$ versus the two-sided alternative. We use $\alpha = 0.05$, assume approximately independent outcomes within each model and large-sample normality ($n\hat{p}$ and $n(1 - \hat{p})$ well above 10), and compare $|Z|$ to the two-tailed critical value 1.96.

CHAPTER 4

IMPLEMENTATION

4.1 Hardware and Software Environment

All experiments were conducted on an NVIDIA Tesla T4 GPU accessed through Google Colab Pro. The Tesla T4 provides 16 GB GDDR6 memory, 2,560 CUDA cores, and 8.1 TFLOPS of FP32 performance, with Tensor Core support delivering up to 65 TFLOPS at FP16 precision. The host system provided 25 GB of RAM and 100 GB of temporary disk storage for dataset and model files. Table 4.1 lists the complete software stack with version numbers used throughout the experiments.

Table 4.1: Software stack with version numbers.

Component	Version
Python	3.10
PyTorch	2.1
CUDA / cuDNN	12.1 / 8.9
Hugging Face Transformers	4.35
OpenAI Whisper	20231117
Librosa	0.10
FaceNet-PyTorch (MTCNN)	2.5
FastText	0.9.2
scikit-learn	1.3
NumPy / Pandas	1.24 / 2.1
Matplotlib / Seaborn	3.8 / 0.13

4.2 Data Pipeline Implementation

4.2.1 Audio Processing and Whisper Transcription

The Whisper transcription pipeline loads the whisper-base model, processes each IEMOCAP audio file (resampled to 16 kHz), and saves the resulting transcription with its dialogue identifier. Segments shorter than 0.6 seconds are filtered out, as are non-speech segments that produce empty transcriptions. Mel-spectrograms are computed via Librosa with 128 Mel bands, a 400-sample FFT window (25 ms at 16 kHz), and a 160-sample hop length (10 ms). Power-to-dB conversion is applied for dynamic range compression, and each spectrogram is normalised to zero mean and unit variance per

frequency band.

4.2.2 Face Extraction and Multimodal Alignment

MTCNN extracts faces from every 5th video frame with minimum face size of 40 pixels and detection confidence threshold of 0.95. Detected faces are aligned using five facial landmarks (eyes, nose, mouth corners), cropped to 224×224 pixels, and standardised using ImageNet mean and standard deviation. The 5,000-per-class cap is enforced via stratified random sampling after all frames for a class have been extracted. Multimodal alignment uses IEMOCAP’s dialogue identifier system: each transcript is matched to its audio file, paired with the face image from the temporally closest video frame, and verified for label consistency. This process yields 14,898 synchronized image–text pairs.

4.3 Model Implementation Details

Table 4.2 summarises the three-stream text module, and Table 4.3 details the multimodal MLP classifier.

Table 4.2: Text module stream specifications.

Stream	Source	Raw Dim	Output Dim
A	RoBERTa + 1D CNN	768	256
B	FastText + FC	300	128
C	NRC Lexicon + FC	10	32
Concatenated output		416 → 256 (projected)	

The roberta-base model (125M parameters) is loaded with pre-trained weights; the first 8 of 12 layers are frozen, and the remaining 4 layers plus the 1D CNN classifier head are fine-tuned with a discriminative learning rate (2×10^{-5} for transformer layers, 1×10^{-3} for the classifier). FastText embeddings (cc.en.300.bin) are mean-pooled per utterance and projected to 128 dimensions. The NRC Emotion Lexicon (14,182 words, 10 dimensions) is summed per utterance and projected to 32 dimensions. ResNet-18 is initialised with ImageNet weights; layer1 and layer2 are frozen while the deeper layers and the self-attention head are fine-tuned (11.1M total parameters).

Table 4.3: Multimodal MLP classifier ³⁶architecture.

Layer	Input Dim	Output Dim	Activation	Dropout
FC Layer 1	928	512	ReLU	0.3
FC Layer 2	512	256	ReLU	0.3
FC Layer 3	256	128	ReLU	0.3
Output	128	5	Softmax	–

The total parameter count of 136.3M is distributed as: RoBERTa (~ 125 M), ResNet-18 (~ 11.1 M), FastText FC + NRC FC (~ 0.04 M), and MLP classifier (~ 0.16 M).

4.4 Training Configuration

Four optimizers are systematically compared across all 12 configurations (4 optimizers $\times$ 3 MLP depths):

- **Adam**: Adaptive moment estimation with default $\beta_1 = 0.9, \beta_2 = 0.999$.
- **Radam**: Rectified Adam with variance-based adaptive learning rate for stable early training.
- **Lion**: Sign-based optimizer discovered via neural architecture search; lower memory footprint.
- **AdamW**: Decoupled weight decay (1×10^{-2}), which regularises independently of the adaptive learning rate and proves critical for preventing majority-class bias.

Table 4.4 summarises the key hyperparameters for each model variant.

Table 4.4: Training hyperparameters for each model.

Parameter	Audio	Text	Image	Multimodal
Learning rate	1×10^{-3}	$2 \times 10^{-5} / 1 \times 10^{-3*}$	1×10^{-4}	1×10^{-4}
Batch size	32	16	32	16
Epochs	50	20	30	25
Weight decay	–	–	–	1×10^{-2}
Dropout	0.3	0.3	0.3	0.3

*Discriminative LR: lower for RoBERTa layers, higher for classifier head.

All models use cross-entropy loss. Class balance is achieved through stratified sampling (maintaining class proportions in each batch) and the 5,000-per-class image cap rather than explicit class weighting. Within training sessions (1–4), a 90:10 train/validation split is used with early stopping: training ²⁷halts if validation loss does not improve for 5 consecutive epochs. The checkpoint with the lowest validation loss is selected for evaluation on the held-out test set (Session 5).

4.5 Computational Profile and Reproducibility

Table 4.5 summarises the computational profile of each model on the Tesla T4.

Table 4.5: Computational profile of each model on Tesla T4 GPU.

Model	Parameters	Time/Epoch	Inference/Sample
Audio CNN	341K	45 s	12 ms
Text (hybrid)	~125M	102 s	15 ms
Image (ResNet-18)	11.1M	35 s	5 ms
Multimodal	136.3M	153 s	21 ms

The multimodal system’s inference latency of 21 ms per sample is well within the 100 ms threshold typically required for real-time HCI applications, suggesting feasibility for deployment in interactive settings such as conversational agents or video conferencing emotion monitors.

To ensure full reproducibility, fixed random seeds (42) are set for Python’s random module, NumPy, and PyTorch (CPU and CUDA), with deterministic CUDA mode enabled. All training metrics (loss, accuracy, per-class F1) are logged at each epoch and saved to structured CSV files. Model checkpoints are saved with descriptive naming conventions including model type, optimizer, MLP depth, and epoch number.

CHAPTER 5 RESULTS AND ANALYSIS

5.1 Pre-processing Outcomes

Audio-text alignment produced 7,100 labelled clips (losses from mismatched dialogue IDs, corrupt segments, and utterances below 0.6 s). After restricting to five emotions, the filtered corpus comprised 6,385 audio segments, 6,426 text samples, and 25,000 face images (5,000 per class, stratified cap). Matching faces to 7,018 audio-text entries with one face per utterance yielded **14,898** evaluation pairs; Table 3.1 gives the class breakdown and motivates macro-averaged metrics under moderate imbalance.

5.2 Unimodal Baseline Results

Table 5.1 lists unimodal runs (Adam, 90:10 train-validation inside training sessions). The Mel-spectrogram audio CNN reached **32.04%** (Macro F1 = 0.2839); the hybrid text stream (RoBERTa + FastText + NRC) reached **65.00%** (0.6600); ResNet-18 faces were strongest at **79.00%** (0.7900), setting the unimodal ceiling before fusion and selective audio-to-text routing.

5.3 Multimodal Optimization

Table 5.1: Consolidated results: unimodal baselines through optimised multimodal framework.

Category	Model / Configuration	Accuracy (%)	Macro F1
Unimodal Baselines	Audio Model	32.04	0.2839
	Text Model	65.00	0.6600
	Image Model (Strongest Baseline)	79.00	0.7900
Multimodal Baseline	Adam + 2-layer MLP	85.20	0.8570
Optimization Stages	RAAdam + 2-layer MLP	84.93	0.8538
	Lion + 2-layer MLP	83.12	0.8395
	AdamW + 2-layer MLP	85.74	0.8627
	AdamW + 3-layer MLP (Best)	86.48	0.8690
	AdamW + 4-layer MLP	85.81	0.8632

5.3.1 Optimizer and Depth Search

Concatenating ResNet-18 with the enriched text vector using Adam and a 2-layer MLP reaches 85.20% (+6.20 pp over image-only). With depth fixed at two layers, optimizers rank Adam 85.20%, RAAdam 84.93%, Lion 83.12%, and AdamW 85.74%. Fixing AdamW, the depth sweep gives 2-layer 85.74%, **3-layer 86.48%**, and 4-layer 85.81%. Swapping the text-side 1D CNN for an LSTM yields 85.77% (−0.71 pp), consistent with short utterances favouring local n-grams.

5.3.2 Best Configuration

The selected setup is **AdamW + 3-layer MLP**, achieving **86.48%** accuracy and Macro F1 **0.8690**, i.e. +7.48 pp over the best unimodal image baseline and +1.28 pp over the initial multimodal stack. The CNN-based multimodal model matches these headline figures in Table 5.1.

5.4 Role and Contribution of Each Modality

Table 5.2 quantifies unimodal floors and fusion gains toward the full pipeline.

Table 5.2: Per-modality contribution to the final 5-class prediction.

Configuration	Acc. (%)	Macro F1	Δ vs. best unimodal
Audio only (raw Mel-spectrogram)	32.04	0.2839	−46.96
Text only (RoBERTa+FastText+NRC)	65.00	0.6600	−14.00
Image only (ResNet-18)	79.00	0.7900	0.00
Image + Text (audio not transcribed)	85.20	0.8570	+6.20
Image + Text + Audio→Text	86.48	0.8690	+7.48

Leave-one-modality-out progression shows ResNet-18 anchoring at 79.00%, hybrid text adding the largest increment (+6.20 pp to 85.20%), and Whisper-based speech-to-text nudging 85.20%→86.48% (+1.28 pp). Together this answers **RQ2**: fusion benefits most from enriched text, while audio helps when exposed as transcribed words rather than raw spectrograms.

In comparison to the other models, RoBERTa ranks among the top performers in terms of per-class F1 score and could correct the disambiguation of some lexically similar negative words (e.g., “angry” and “frustrated”). Whisper is able to surface words or disfluencies that do not appear in the manual transcript, increasing the semantic lift of the transcription, while Raw Mel doesn't provide much lift on its own.

5.5 Class-wise Performance under Imbalance

5.5.1 Per-Class Metrics

Table 5.3: Precision, Recall and F1 scores per class for final optimised multi-modal model (AdamW + 3-layer MLP).

Class	Precision	Recall	F1-Score	Support
Excited	0.92	0.91	0.91	454
Sad	0.89	0.88	0.88	524
Neutral	0.87	0.86	0.87	516
Angry	0.85	0.86	0.85	791
Frustrated	0.84	0.83	0.84	695
Macro avg.	0.874	0.868	0.869	2,980
Weighted avg.	0.866	0.865	0.865	2,980

The “Excited” peak is at F1 0.91, the lowest at F1 0.84 due to the overlap with “Angry.” Macro and weighted F1 are similar with a difference of 0.004, thus supporting RQ3.

5.5.2 Confusion Matrix Analysis

The dominant errors in figure 5.1 are consistent with the adjacent high-arousal negative states in Russell's circumplex and other facial action units, namely Angry and Frustrated: the two states switch. Neutral-to-Frustrated/Sad decreases. The type of dialogue context/ prosody used in this architecture is not sufficient to fully resolve the angry/frustrated boundary.

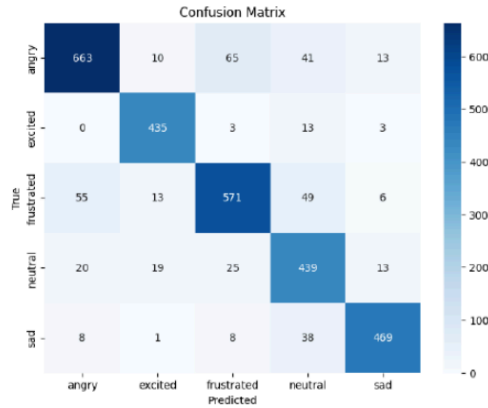


Figure 5.1: The confusion matrix of the final optimised multimodal model (AdamW + 3-layer MLP). The higher the per-class accuracy in all 5 emotions the higher level of “diagonal dominance” in the data.

5.6 Statistical Significance Analysis

Using Equation (3.2) with $n_1 = n_2 = 14,898$, the optimised multimodal model ($\hat{p}_2 = 0.8648$) versus image-only ($\hat{p}_1 = 0.7900$) yields $Z = 16.94$, far beyond the two-tailed critical value 1.96 at $\alpha = 0.05$ ($p \ll 0.001$), so the +7.48 pp gain over the visual baseline is statistically decisive. A follow-up test versus the next-best multimodal configuration gives $Z = 11.40$ ($p < 0.05$), confirming AdamW + 3 layers over alternatives rather than sampling noise.

5.7 Comparison with Recent Multimodal Transformer Approaches

Table 5.4: Comparison with recent multimodal emotion recognition methods on IEMO-CAP. A = audio, V = visual, T = text.

Method	Year	Modalities	Fusion Strategy	Accuracy (%)
Du et al. [21]	2023	A+V+T	Feature fusion + residual connections	61.40
Xu et al. [22]	2023	A+V+T	Multi-modal transformer (multi-head attn.)	60.89
Jiang et al. [23]	2020	A+T	Contextual fusion w/ info enhancement	65.77
AVaTER (Das et al.) [19]	2024	A+V+T	Cross-modal attention transformer	~74
Shayaninasab & Babaali [20]	2024	A+V+T	Pretrained transformer fusion	~75
Ours	2026	A+T+V+T	Pre-fusion text enrichment + MLP	86.48

Prior systems cluster in the low-to-mid 60s through ~75% as text encoders and attention improve; gains appear to track representation quality as much as fusion topology. Feeding RoBERTa, FastText, and NRC into an early-fused 928-d descriptor with a compact MLP lifts accuracy to **86.48%**, underscoring that rich modality-specific features can outweigh elaborate fusion layers on conversational IEMOCAP-style data.

5.8 Discussion of Findings

The experiments support prioritising modality-specific representation depth before chasing elaborate fusion topologies: a compact MLP on an enriched 928-d vector surpasses several transformer-centric systems by wide margins when linguistic inputs carry RoBERTa, FastText, and NRC information prior to fusion. This aligns with conversational emotion settings where lexical semantics and lexicon priors disambiguate visually confusable negatives and where Whisper exposes spoken content without relying on low-accuracy raw acoustics.

Macro and weighted F1 remain tightly coupled, showing fairness need not trade off against headline accuracy when sampling and evaluation are designed deliberately. Together, these observations advocate macro metrics plus significance testing as default reporting for imbalanced affect corpora.

7
CHAPTER 6

CONCLUSION, LIMITATIONS AND FUTURE SCOPE

6.1 Conclusion

This thesis addressed two persistent gaps in multimodal emotion recognition on the IEMOCAP benchmark: under-enriched linguistic features at the fusion stage and aggregate-only evaluation that masks minority-class performance. Starting from the observation that prior multimodal IEMOCAP systems plateau at 61–66% (and up to approximately 75% with recent transformers), we hypothesised that the bottleneck lies not in the fusion mechanism but in the quality of the linguistic features entering the fusion stage.

To test this hypothesis, we designed a multimodal deep learning system that combines visual features from ResNet-18 with a hybrid text stream integrating three complementary linguistic sources—RoBERTa for contextual semantics, FastText for sub-word coverage, and the NRC Emotion Lexicon for explicit emotion-word priors—before fusion. Audio is selectively re-routed through Whisper transcription to augment the text stream with semantic content rather than raw acoustic features.

The system achieves **86.48% accuracy** and a **Macro F1 of 0.8690** on the five-class IEMOCAP benchmark, a 7.48-point improvement over the strongest unimodal baseline (image-only at 79.00%). Two-proportion Z-tests ($Z = 16.94$ vs. visual baseline, $Z = 11.40$ vs. runner-up multimodal variant) confirm the improvements are statistically significant. Per-class F1-scores range from 0.84 to 0.91, with a spread of only 0.07, demonstrating class-fair performance. Table 6.1 summarises the key contributions and their quantitative impact.

Table 6.1: Summary of thesis contributions and their quantitative impact.

Contribution	Key Metric	Impact
Hybrid pre-fusion text enrichment	Accuracy gain	+6.20 pp
Selective audio-to-text re-routing	Accuracy gain	+1.28 pp
Class-fair evaluation protocol	Per-class F1 spread	0.07
Full system vs. best unimodal	Overall accuracy	+7.48 pp
Statistical significance	Z-test	$Z = 16.94$

6.2 Limitations of the Proposed Study

While the proposed system achieves strong results, several limitations should be acknowledged:

- **Flat concatenation cannot model inter-modality dynamics.** The system treats all 928 feature dimensions equally and cannot detect temporal cross-modal mismatches. For example, sarcasm—where positive words accompany a sad facial expression—cannot be identified by concatenation alone. Cross-modal attention or transformer-based alignment [17] could address this limitation.
- **Persistent angry/frustrated confounding.** Despite achieving the best per-class performance among compared IEMOCAP studies, the Angry/Frustrated pair remains the most confounded (F1 of 0.85 and 0.84 respectively). This confusion reflects the inherent proximity of these two emotions in arousal-valence space and may require dialogue-level context, finer prosodic features, or dominance-dimension cues for resolution.
- **IEMOCAP as a controlled lab dataset.** All results are established on IEMOCAP’s controlled laboratory environment with consistent lighting, clear face visibility, and semi-professional actors. Generalisation to noisy, spontaneous, real-world conditions—variable lighting, background noise, overlapping speech, cultural diversity—remains unvalidated.
- **Single-dataset evaluation.** The study uses only IEMOCAP; cross-corpus validation on MELD, CMU-MOSEI, or RAVDESS would strengthen the generalisability claims.

6.3 Future Scope

Based on the findings and limitations of this study, the following directions are identified for future research:

- **Attention/Transformer-based fusion replacement.** Replacing flat concatenation with cross-modal attention or a lightweight transformer-based fusion module could enable the system to model inter-modality interactions. Given the already high performance with concatenation, even modest gains from attention-based fusion would push accuracy into the high 80s or low 90s on IEMOCAP.
- **Smarter class-balancing for multimodal settings.** Applying SMOTE or domain-aware augmentation—generating synthetic multimodal samples for underrepresented classes—could sharpen the Angry/Frustrated boundary. Class-weighted loss functions sensitive to each emotion’s acoustic and visual profile could further improve minority-class performance.
- **In-the-wild corpora validation.** MELD (TV show dialogues with sound) and CMU-MOSEI (YouTube monologues with diverse speakers) would be a good test under real conditions.
- **Dialogue level context and other prosodic features.** Use past utterances in conversation to make the emotion(s) more apparent that depend on conversational flow. Including extra prosodic information (pitch contour, speaking rate) in Whisper’s transcription might be possible by using a parallel Wav2Vec 2.0 encoder.
- **Light-weight model optimisation.** Model compression methods like knowledge distillation and quantisation could further bring down the 136.3M number of parameters and keep the accuracy intact, allowing it to be deployed on the edge devices and mobile platforms.

BIBLIOGRAPHY

- [1] Bokhare, A., Kothari, T.: Emotion detection-based video recommendation system using machine learning and deep learning framework. *SN Comput. Sci.* **4**(3), 1–6 (2023)
- [2] Hans, A.S.A., Rao, S.: A CNN-LSTM based deep neural networks for facial emotion detection in videos. *Int. J. Adv. Signal Image Sci.* **7**(1), 11–20 (2021)
- [3] Mustaqeem, Kwon, S.: A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **20**(1), 183 (2020)
- [4] Madan, A., Kumar, D.: CNN-based models for emotion and sentiment analysis using speech data. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **23**(10), 1–20 (2024)
- [5] Cahyani, D.E., et al.: Emotion detection in text using convolutional neural network. In: *Proc. IEIT 2022*, pp. 372–376 (2022)
- [6] De Leon Langure, A., Zareei, M.: Evaluating the effect of emotion models on the generalizability of text emotion detection systems. *IEEE Access* **12**, 70489–70500 (2024)
- [7] Cahyani, D.E., Patasik, I.: Performance comparison of TF-IDF and Word2Vec models for emotion text classification. *Bull. Electr. Eng. Inform.* **10**(5), 2780–2788 (2021)
- [8] Vishwakarma, A., Sakalle, V.: Unravelling the mind's emotions: A review of EEG-based emotion detection. In: *Intelligent Computing Techniques for Smart Energy Systems*, pp. 177–188. Springer (2025)
- [9] Zhou, Y., Kang, X., Ren, F.: Prompt consistency for multi-label textual emotion detection. *IEEE Trans. Affect. Comput.* **15**(1), 121–129 (2024)
- [10] Munir, A., et al.: Deep learning-based emotion recognition: A systematic review. *JISR Computing* **21**(2), 1–15 (2023)
- [11] Chen, Y., He, J.: Deep learning-based emotion detection. *J. Comput. Commun.* **10**(2), 57–71 (2022)
- [12] Taghavi, Z.S., Satvati, A., Sameti, H.: A change of heart: Improving speech emotion recognition through speech-to-text modality conversion. *arXiv:2307.11584* (2023)

- [13] Cedeño-Moreno, D., et al.: UTP at EmoSpeech-IberLEF2024: Using random forest with FastText and Wav2Vec 2.0 for emotion detection. In: CEUR Proc., vol. 3756, pp. 2–7 (2024)
- [14] Najafabadi, M.K., et al.: A multi-level embedding framework for decoding sarcasm using context, emotion, and sentiment features. *Electronics* **13**(22), 1–12 (2024)
- [15] Adel, O., Fathalla, K.M., Abo ElFarag, A.: MM-EmoR: Multi-modal emotion recognition of social media using concatenated deep learning networks. *Big Data Cogn. Comput.* **7**(4), 1–12 (2023)
- [16] Ghayoumi, M.: A review of multimodal biometric systems: Fusion methods and their applications. In: Proc. IEEE/ACIS 14th ICIS, pp. 131–136 (2015)
- [17] Ali, I., et al.: Proposing sentiment analysis model based on BERT and XLNet for movie reviews. *Multimedia Tools Appl.* (2024)
- [18] Danyal, M.M., et al.: Citation sentiment classification using XLNet & transformer models. *JISR-C* **23**(2) (2025)
- [19] Das, A., et al.: AVaTER: Fusing audio, visual, and textual modalities using cross-modal attention for emotion recognition. *Sensors* **24**(18), 1–23 (2024)
- [20] Shayaninasab, M., Babaali, B.: Multi-modal emotion recognition by text, speech and video using pretrained transformers. *arXiv:2402.07327* (2024)
- [21] Du, X., Yang, J., Xie, X.: Multimodal emotion recognition based on feature fusion and residual connection. In: Proc. IEEE EEBDA, pp. 373–377 (2023)
- [22] Xu, C., Gao, Y.: Multi-modal transformer with multi-head attention for emotion recognition. In: Proc. IEEE ICSECE, pp. 826–831 (2023)
- [23] Jiang, D., et al.: Contextual multimodal sentiment analysis with information enhancement. *J. Phys.: Conf. Ser.* **1453**(1), 1–12 (2020)
- [24] Khoirullah, H.B., et al.: Facial expression recognition using CNN with attention module. *Int. J. Inform. Visualization* **6**(4), 897–903 (2022)
- [25] Acheampong, F.A., et al.: Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artif. Intell. Rev.* **54**(8), 5789–5829 (2021)
- [26] Farooq, M., et al.: Conversational emotion detection and elicitation: A preliminary study. In: Proc. IEEE GlobConET (2023)

- [27] Pereira, P., et al.: Context-dependent embedding utterance representations for emotion recognition in conversations. In: Proc. ACL, pp. 228–236 (2023)
- [28] Adoma, A.F., et al.: Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition. In: Proc. ICCWAMTIP, pp. 117–121 (2020)
- [29] Deng, L., Qin, T.: Generalizable negation detection: Addressing cross-domain challenges with data augmentation. In: Proc. ICCECE, pp. 451–455 (2025)
- [30] Abdullah, A.A., et al.: NER-RoBERTa: Fine-tuning RoBERTa for NER within low-resource languages. arXiv:2412.15252 (2024)
- [31] Qorina, E.S., et al.: Comparative analysis of FastText and Word2Vec on semantic similarity for Sirah Nabawiyah IR. In: Proc. CITSM (2020)
- [32] Bakker, R.M., Di Scala, D.L.: From text to knowledge graph: Comparing relation extraction methods in a practical context. In: CEUR Proc., vol. 3749 (2024)
- [33] Nie, W., et al.: Long dialogue emotion detection based on commonsense knowledge graph guidance. IEEE Trans. Multimedia **26**, 514–528 (2024)
- [34] Jain, A.: A comprehensive guide to performance metrics in machine learning. Medium – Level Up Coding (2024)
- [35] Opitz, J.: A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Trans. Assoc. Comput. Linguist. **12**, 820–836 (2024)
- [36] Walpole, R.E., et al.: Probability and Statistics for Engineers and Scientists, 9th edn. Pearson (2016)
- [37] Ismael, S.S., et al.: Statistical significance testing in machine learning: A practical guide. Kurdistan J. Appl. Res. (2021)
- [38] Berrar, D.: Performance measures and statistical tests for machine learning. In: Encyclopedia of Machine Learning and Data Science. Springer (2024)
- [39] Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019)
- [40] Bojanowski, P., et al.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
- [41] Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Comput. Intell. **29**(3), 436–465 (2013)

- [42] He, K., et al.: Deep residual learning for image recognition. In: Proc. IEEE CVPR, pp. 770–778 (2016)
- [43] Busso, C., et al.: IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**(4), 335–359 (2008)
- [44] Radford, A., et al.: Robust speech recognition via large-scale weak supervision. In: Proc. ICML, pp. 28492–28518 (2023)
- [45] Picard, R.W.: *Affective Computing*. MIT Press (1997)

APPENDIX

I. List of Publications

1. Pulkit Jain & Sonika Dahiya (2026). A Multimodal Deep Learning System for Enhanced Emotion Detection [Scopus Indexed] [Accepted- ICDAM 2026, Springer LNNS Series]

II. Paper Acceptance Proof



Pulkit Jain <pulkitjain78@gmail.com>

ICDAM 2026: Paper Notification for Paper ID 1434

1 message

ICDAM 2026 <icdam.conf@gmail.com>
To: Pulkit Jain <pulkitjain78@gmail.com>

7 May 2026 at 12:28

7th International Conference on Data Analytics & Management (ICDAM-2026):

Dear Author(s),

Greetings from ICDAM 2026!

We congratulate you that your paper with submission ID 1434 and Paper Title "A Multimodal Deep Learning System for Enhanced Emotion Detection" has been accepted for publication in the Springer LNNS series [Approved]-[Indexing: SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago; All books published in the series are submitted for consideration in Web of Science]. This acceptance means your paper is among the top 20% of the papers received/reviewed. We urge you to complete your registration immediately to secure your spot at this highly anticipated event. **Kindly complete the registration by 10th May 2026. No further Extension.**

Please register as soon as possible and submit the following documents to icdam.conf@gmail.com as soon as possible.

- 1. Final Camera-Ready Copy (CRC) as per the Springer format. (See <https://icdam-conf.com/downloads>)
- 2. Copy of e-receipt of registration fees. (For Registration, see <https://icdam-conf.com/registrations>)
- 3. The final revised copy of your paper must be mailed to the conference mail id.

Note : Standard Paper size -- 10-12 pages. Over length of more than 12 pages, paper charges USD 10 per extra page

The reviewers comments are given at the bottom of this letter, please improve your paper as per the reviewers comments. While preparing the final CRC manuscript, kindly check the following google link of proceedings of the previous International Conference on Data Analytics and Management:

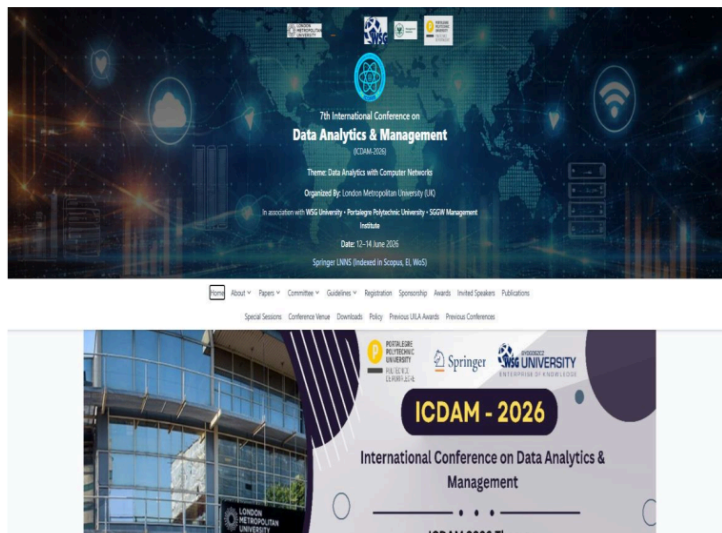
<https://link.springer.com/conference/icdam>

and it is suggested to cite the relevant latest papers matching the area of your current research paper.

The paper prior to submission should be checked for plagiarism from licensed plagiarism softwares like Turnitin/Authenticate etc. The similarity content should not exceed 15%, and AI content must not exceed 20%.

Pay registration fees via online portal:

III. Indexing of Conference Proof



IV. Conference Paper Registration Receipt

Ui educon



Payment Receipt Transaction Reference: pay_SnhJAYjnOdGHCl
This is a payment receipt for your transaction on ICDAM 2026

AMOUNT PAID **₹ 17,150.00**

ISSUED TO
pulkitjainji761@gmail.com
+918225922761

PAID ON
10 May 2026

Category
Research Scholar/Student

DESCRIPTION	UNIT PRICE	QTY	AMOUNT
Convenience Fee	₹ 250.00	1	₹ 250.00
Amount	₹ 16,900.00	1	₹ 16,900.00
Total			₹ 17,150.00
Amount Paid			₹ 17,150.00

ORIGINALITY REPORT

9%

SIMILARITY INDEX

7%

INTERNET SOURCES

4%

PUBLICATIONS

%

STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	1 %
2	nitdelhi.ac.in Internet Source	1 %
3	v1.overleaf.com Internet Source	1 %
4	www.mdpi.com Internet Source	1 %
5	etd.aau.edu.et Internet Source	<1 %
6	ajitelectricals.com Internet Source	<1 %
7	ethesis.nitrkl.ac.in Internet Source	<1 %
8	"Computational Vision and Bio-Inspired Computing", Springer Science and Business Media LLC, 2020 Publication	<1 %
9	repository.tudelft.nl Internet Source	<1 %
10	files.core.ac.uk Internet Source	<1 %
11	ntnuopen.ntnu.no Internet Source	<1 %
12	psasir.upm.edu.my	

Internet Source

<1 %

13

www.grin.com

Internet Source

<1 %

14

Manoj Singh, Subhash Panwar, Sanju Choudhary. "BiERU-CapsNet: Bidirectional Emotional Recurrent Unit with Capsule Network, a Novel Approach for Textual Sentiment Analysis", SN Computer Science, 2026

Publication

<1 %

15

objectstorage.ap-dcc-gazipur-1.oraclecloud15.com

Internet Source

<1 %

16

"Proceedings of International Conference on Computing Systems and Intelligent Applications", Springer Science and Business Media LLC, 2025

Publication

<1 %

17

cors.archive.org

Internet Source

<1 %

18

carleton.scholaris.ca

Internet Source

<1 %

19

du.diva-portal.org

Internet Source

<1 %

20

pmc.ncbi.nlm.nih.gov

Internet Source

<1 %

21

phuakl.tripod.com

Internet Source

<1 %

22

unsworks.unsw.edu.au

Internet Source

<1 %

23

ebin.pub

Internet Source

<1 %

24

www.theseus.fi

Internet Source

<1 %

25

"Advances in Artificial Intelligence - IBERAMIA 2018", Springer Science and Business Media LLC, 2018

Publication

<1 %

26

"Advances in Intelligent Computing Techniques and Applications II", Springer Science and Business Media LLC, 2026

Publication

<1 %

27

"Neural Information Processing", Springer Science and Business Media LLC, 2026

Publication

<1 %

28

Basma Ammour, Toufik Bouden, Larbi Boubchir. "Face-iris multi-modal biometric system using multi-resolution Log-Gabor filter with spectral regression kernel discriminant analysis", IET Biometrics, 2018

Publication

<1 %

29

Loay Hatem, Ahmed Omar, Abdelmgeid A. Ali, Heba Mamdouh Farghaly. "Tackling toxicity in Arabic social media through advanced detection techniques", Scientific Reports, 2025

Publication

<1 %

30

Tanusree Nath, Vedika Gupta, Manjari Gupta, Rajesh Sharma. "Decoding Multimodal Text Analytics: Tasks, Datasets, Fusion Models, and Future Frontiers", WIREs Data Mining and Knowledge Discovery, 2026

Publication

<1 %

31	Xiangshi Ren, Chen Wang. "Human-Engaged Computing", CRC Press, 2026 Publication	<1 %
32	era.ed.ac.uk Internet Source	<1 %
33	iarjset.com Internet Source	<1 %
34	www.atlantis-press.com Internet Source	<1 %
35	www.bloodraynebetrayal.com Internet Source	<1 %
36	Şafak Kılıç. "Densenet201+ with multi-scale attention and deep feature engineering for automated Kellgren–Lawrence grading of knee osteoarthritis", PeerJ Computer Science, 2025 Publication	<1 %
37	"Computational Intelligent Systems", Springer Science and Business Media LLC, 2026 Publication	<1 %
38	Intelligent Systems Reference Library, 2016. Publication	<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography On