COMPARATIVE STUDY OF HYBRID FUSION FOR ROBUST MULTIMODAL EMOTION RECOGNITION

FERNADO GEORGE ANAK MANI

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2026

COMPARATIVE STUDY OF HYBRID FUSION FOR ROBUST MULTIMODAL EMOTION RECOGNITION

FERNADO GEORGE ANAK MANI

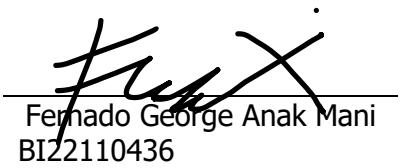
THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE BACHELOR OF COMPUTER SCIENCE WITH HONOURS (DATA SCIENCE)

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, equations, summaries and references, which have been duly acknowledged.

5 January 2026



Fernado George Anak Mani
BI22110436

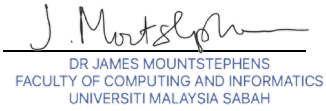
CERTIFICATION

NAME	:	**FERNADO GEORGE ANAK MANI**
MATRIC NUMBER	:	**BI22110436**
TITLE	:	**COMPARATIVE STUDY OF HYBRID FUSION FOR ROBUST MULTIMODAL EMOTION RECOGNITION**
DEGREE	:	**BACHELOR OF COMPUTER SCIENCE WITH HONOURS (DATA SCIENCE)**
VIVA'S DATE	:	**16 JANUARY 2026**

CERTIFIED BY.

1. SUPERVISOR
Dr. James Mountstephens

Signature



DR JAMES MOUNTSTEPHENS
FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the creators and providers of the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database for providing access to their invaluable dataset for multimodal emotion recognition research. This dataset has been instrumental in advancing my understanding of how different fusion strategies perform under various conditions.

I would also like to extend my heartfelt appreciation to my supervisor, Dr. James Mountstephens, for his unwavering support, guidance, and mentorship throughout the duration of this project. I am deeply grateful to Dr. James Mountstephens, particularly for his advice on navigating the class imbalance issues within the IEMOCAP dataset and his suggestion to explore the SuperTransformer architecture, which became a core component of this thesis.

Additionally, I am thankful to all the individuals who have contributed to this project, directly or indirectly. Special thanks go to my friends for their encouragement and support. Finally, I thank my family for their continuous support mentally and physically.

Fernado George Anak Mani

ABSTRACT

Multimodal Emotion Recognition (MER) is attempting to identify the emotions that a person expresses when analyzing their facial features, voice, and text content. The use of deep learning models has also enhanced the precision of MER; however, there is a need to investigate the application of fusion techniques to determine their effectiveness in real-world scenarios where there might be noise and missing data. The paper explores the use of various architectures, including statistical (Early Fusion, Late Fusion, Hierarchical Fusion, Gated Fusion) and contextual (Bi-GRU, SuperTransformer). The use of advanced pre-trained feature extractors is also employed; for audio, it is WavLM-Base+, while visual and text features are detected by DINOv2-base and ModernBERT-base, respectively. The models are analyzed considering their baseline accuracy, when there are missing data modalities, and when there is additional noise. The F1-Score values obtained from the analysis indicate that contextual architectures are significantly superior to statistical models. In terms of individual architectures, the Contextual Bi-GRU Early Fusion architecture proved to be the best, achieving a Weighted F1-Score of 67.9%. However, the ensemble approach resulted in achieving the lowest value, providing a result of 71.6%. Additionally, the ensemble approach is also more robust and maintains an F1-Score of 67.5% when there is high noise (with a value of sigma=1.0). Additionally, it was observed that it is still difficult to identify the difference between the Angry and Frustrated classes as they tend to overlap at times. In the study, it is concluded that Ensemble Approaches are the most suitable in real-world settings and that Bi-GRU models are the fastest and most accurate models when they are contextual.

Keywords: Multimodal Emotion Recognition; Fusion Strategies; Robustness; Bi-GRU; SuperTransformer; Ensemble Learning;

ABSTRAK

Kajian Perbandingan Fusion Hibrid untuk Pengecaman Emosi Multimodal yang Teguh

Pengecaman Emosi Multimodal (MER) cuba memahami perasaan manusia dengan melihat ekspresi wajah, mendengar pertuturan, dan membaca teks mereka. Pembelajaran Mendalam telah menjadikan MER lebih tepat, namun adalah penting untuk menguji bagaimana strategi gabungan (fusion) berfungsi dalam dunia nyata, di mana terdapat hingar (noise) dan data yang hilang. Kajian ini menjalankan analisis perbandingan antara seni bina Statistik (Awal, Lewat, Hierarki, Berpagar) dan seni bina Kontekstual (Bi-GRU, SuperTransformer) menggunakan set data IEMOCAP. Penyelidikan ini menggunakan pengekstrak ciri pra-latih yang canggih: WavLM-Base+ (Audio), DINOv2-base (Visual), dan ModernBERT-base (Teks). Kajian ini meneliti model berdasarkan ketepatan garis dasar, prestasi apabila modaliti hilang, dan ketahanan apabila hingar ditambah. Keputusan menunjukkan bahawa seni bina kontekstual berprestasi jauh lebih baik berbanding garis dasar statistik. Model Gabungan Awal Bi-GRU Kontekstual mempunyai prestasi terbaik bagi mana-mana model tunggal, dengan skor F1 Wajaran sebanyak 67.9%. Dengan skor 71.6%, strategi Ensembel mencatatkan prestasi keseluruhan terbaik. Dari segi keteguhan, model Ensembel lebih stabil, mengekalkan skor F1 sebanyak 67.5% walaupun terdapat banyak hingar (σ=1.0). Analisis ralat menunjukkan bahawa pasangan Marah-Kecewa (Angry-Frustrated) masih paling sukar dibezakan kerana pertindihan dominasi mereka. Kajian mendapati bahawa kaedah Ensembel adalah yang paling boleh dipercayai untuk kegunaan dunia sebenar, manakala model Bi-GRU Kontekstual memberikan keseimbangan terbaik antara kelajuan dan ketepatan.

Kata kunci: *Pengecaman Emosi Multimodal, Strategi Gabungan, Keteguhan, Kehilangan Modaliti, Hingar, Pembelajaran Mendalam, XGBoost, IEMOCAP.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

GLOSSARY

Term	Definition
Additive White Gaussian Noise (AWGN)	Used in this study's robustness testing (The Gauntlet), AWGN is injected into the feature vectors at varying intensities (sigma 0.1 to 1.0) to simulate environmental interference.
Affective Computing	A field of study that aims to build computer systems that can recognize, understand, and respond to human feelings.
Attention Mechanism	A method in deep learning that lets models focus on the most important areas within and between modalities instead of treating all data the same.
Bi-GRU (Bidirectional Gated Recurrent Unit)	A deep learning architecture that looks at temporal sequences in both directions to get information about the past and the future. This study employs it for contextual modelling.
Contextual Features	Feature representations that keep the whole sequence of data over time (temporal), so the model can learn how things change over time (like how pitch changes over a sentence).
Cross-Modal Attention	A way to improve the representation of one type of information (like audio) by paying attention to information from another type of information (like text).

DINOv2	A self-supervised Vision Transformer model from Meta AI to get strong visual features from facial expressions without needing labelled data.
Early Fusion	Also called feature-level fusion, this is a method in which features from different modalities are combined into a single vector before being sent to a classifier.
Ensemble Learning	A strategy that uses the predictions of several different models (like Bi-GRU and Random Forest) to make the overall performance and stability better than any one model.
Gated Fusion	A fusion method that uses a learnable gating network to change the weights of different modalities on the fly, letting the model ignore inputs that are not as reliable.
Hierarchical Fusion	A hybrid strategy that puts modalities together based on how they relate to each other (for example, combining Audio and Text first) before adding the final modality (Visual).
IEMOCAP	The Interactive Emotional Dyadic Motion Capture database is a standard dataset with 12 hours of audio and video conversations that were used to train and test the models in this study.

Late Fusion	Also called decision-level fusion, this is a method in which separate classifiers are trained for each modality and their final predictions (logits/probabilities) are combined (for example, averaged).
Leave-One-Session-Out (LOSO)	A cross-validation method in which the model is trained on four sessions and then tested on the last session that it has not seen yet to make sure that it works for all speakers.
ModernBERT	The study used a transformer-based model to get linguistic (textual) features. It works better in longer contexts and makes inferences faster than the original BERT models.
Multimodal Emotion Recognition (MER)	The automatic detection of emotional states by analyzing and combining information from multiple input channels, like audio, visual, and text, all at once.
Self-Distillation Transformer (SDT)	Also called the SuperTransformer in this study, it is a cutting-edge architecture that uses cross-modal attention and gating mechanisms to combine data from different sources.
Statistical Features	Feature representations are made by calculating aggregate statistics (Mean, Standard Deviation, Max, Min) over time. This creates a fixed-length

	vector that does not include any information about the order of the events.
Valence-Arousal-Dominance (VAD)	Valence measures how positive or negative something is, Arousal measures how intense or exciting something is, and Dominance measures how much control someone has.
WavLM	This study used a large pre-trained model to get acoustic features from speech audio. It does a great job of picking up on paralinguistic cues like tone and pitch.
Weighted F1-Score	The main way to measure success in this study. It is the harmonic mean of precision and recall, with the number of true instances for each label taken into account to make up for class imbalance.

CHAPTER 1

INTRODUCTION

1.1 Introduction

One major goal of affective computing is to make computer systems that can understand how people feel so that people and computers can interact in a more realistic way. Emotion Recognition (ER) is a basic job in this field. Human emotions are inherently multimodal, allowing expression through multiple channels such as linguistic content, vocal intonation, and facial expressions. This has resulted in a growing field of study known as Multimodal Emotion Recognition (MER). MER systems look at and combine data from a lot of different input streams, or modalities. These are usually made up of speech audio, visual data from facial expressions, and written material from people talking to each other.

The idea behind MER is that using data from more than one source should give a more complete and accurate picture of emotional states than just using one type of data. Better computer methods, especially deep learning, have made it much easier to get useful representations from these complicated data streams. The method used to combine or fuse the information from speech, text, and facial modalities is still very important, though. The choice of the fusion approach has a big impact on how correct and trustworthy the system is in different situations.

Current MER systems usually do worse when used in the real world, where there is noise or not enough data. This is because they often work well in controlled lab settings. As the need for emotionally intelligent AI grows in many areas, strong and reliable MER systems are becoming more and more important. So, the main point of this work is the problem of making fusion strategies more robust. This paper aims to systematically assess and compare the relative robustness of fundamental fusion methodologies, including Early, Late, and Hybrid fusion, in the context of simulated real-world data imperfections, employing standardized and efficient feature extraction techniques for facial, speech, and text analysis.

1.2 Problem Background

The inherent shortcomings of single-source analysis resulted in a shift of emotion recognition methods from unimodal to multimodal. Unimodal signals can be confusing; for example, sarcasm cannot always be expressed using writing alone, and a neutral face can hide strong feelings behind. Also, all modalities are prone to some forms of interference, such as occlusions that mask face features or background noise that distorts audio information, thereby reducing analytical accuracy. Perhaps most importantly, because human emotion is such a complex phenomenon that is realized through a combination of verbal and non-verbal cues, unimodal analysis is an incomplete portrayal.

Researchers took advantage of these restrictions by engaging multimodal approaches to probe individual differences in affect in more holistic manners. Vocal prosody might signal the intensity of an emotion, facial emotions might signal its valence, and content language might signal the context or source of an emotion. Besides, information from available modalities may be able to substitute for noise or absent information in other modalities, making the overall system more reliable, while agreement between modalities can increase prediction confidence. However, information fusion becomes a problem when several modalities are combined.

Two simple solutions to this exist which are Early fusion, or feature-level fusion, comes before classification by merging features of all modalities into one high-dimensional vector. Its ability to detect low-level inter-modal interactions early could prove useful. This has its drawbacks, including the possibility of generating huge, complicated feature spaces that are hard to learn, possible requirement of strict temporal synchronization, and possible strong or noisy features dominating others. On the other hand, late fusion, also known as decision-level fusion, learns individual classifiers for each modality and fuses their output (e.g., probabilities or predictions) at the final step. Because decisions are made from available outputs, this introduces modularity as well as inherent robustness to missing modalities. The principal disadvantage is that it bars cross-modal correlations potentially useful from being employed at an earlier processing step. More recently, more sophisticated fusion algorithms reflecting explicitly complex inter-modal relationships have been constructed, along with stronger unimodal feature extraction brought about by deep learning techniques. Even though things have gotten better, there is still a significant gap between the trust achievable with imperfect, real-world data and performance seen on carefully selected laboratory datasets.

Performance is to be severely impaired by noise, either environmental noise or visual barriers, and by missing data streams due to sensor failure or transmission errors. There is lack of comparative literature assessing how well these various strategies stand up to such actual, non-ideal situations, while there are many forms of fusion available, ranging from the fundamental early and late methods to more sophisticated modern processes. It is difficult to make an informed choice in constructing solid MER systems for real-world deployment since there is insufficient clarity in the relative strength of different fusion approaches.

1.3 Problem Statement

There is often a big difference between how well existing Multimodal Emotion Recognition (MER) systems work in controlled lab tests and how well they work in the real world. One of the main reasons for this difference in performance is that people

often assume that the input data will be complete, clean, and time-aligned when building a system. But in real life, data gathered in natural settings is often not perfect. Systems often have problems with background noise that makes speech signals hard to hear, partial occlusions or changing lighting that makes facial data less clear, and even the complete absence of one or more data streams because of sensor failures or other technical issues.

There are many approaches to integrating information from different sources. These range from very basic ones such as early fusion, late fusion, to more complex deep learning, to hybrids that combine the best of both. Yet among all the available approaches, we still know very little about how effective they are in practical settings. There are insufficient comprehensive observations about how robust all the fusion algorithms are against incomplete and noisy data. Even if there were studies suggesting effective approaches, there would be no way to ensure that it will perform best in practical scenarios where common problems with data are always expected.

This results in a challenge when it comes to designing and applying a credible MER system. This paper aims to address this problem and provide operative guidance. It will accomplish this through a comparative assessment of how effectively the predominant methods of fusion, namely Early, Late, and a Hybrid fusion scheme, function under controlled environments, which include noise as well as missing values. The objective of this paper is to provide expert, empirical information to assist in designing improved MER systems.

1.4 Project Objectives

The main goal of this study is to compares and assesses various information fusion techniques for Multimodal Emotion Recognition (MER), with an emphasis on how resilient these techniques are to poor data. In order to fill a recognized vacuum in previous research regarding practical system reliability, this compares early, late, and hybrid

fusion designs under scenarios that mimic real-world difficulties, such as noisy recordings and missing information streams.

The precise actions listed below will be conducted in order to accomplish this primary goal:

1. To employ efficient and computationally viable feature extraction techniques for extracting emotional representations from the audio, text and visual data within the IEMOCAP dataset.

2. To develop and train MER models utilizing distinct fusion architectures, including Early Fusion, Late Fusion and two different hybrid designs (hierarchical and attention-based).

3. To benchmark the baseline performance of these fusion models on the original IEMOCAP dataset using LOSO cross-validation and standard metrics (Weighted Accuracy and Weighted F1-score).

4. To examine the robustness of each fusion model by introducing noise to multimodal features and testing performance with one or more modalities dropped.

1.5 Project Scope

In this study, the efficiency of Early, Late, and Hybrid techniques in classifying emotions is explored. The task here is to classify various types of emotions by evaluating the efficiency in coping with noise and incomplete information as dealt with by these techniques. This study includes audio speech, written form, and facial expressions. The study does not take into account other bodily signals. In this study, only the IEMOCAP dataset has been referred to.

It employs robust and efficient deep learning approaches for extraction of features, which are suitable for extraction on common computers. These are WavLM for audio, ModernBERT for text, and Dinov2 for images. It does not aim to compare approaches for extracting features or propose novel ones. It employs pre-existing approaches that are dedicated to fusion. This study will evaluate approaches under three categories: Early, Late, and Hybrid. Extra attention is given to Attention-Based and Hierarchical approaches.

The evaluation criteria for the fusion techniques are Weighted Accuracy (%), and Weighted F1-score. The robustness experiment is significant and comprises the addition of noise in testing. The types of noise include Additive White Gaussian Noise added to audio, text, and visualization features. The robustness experiment will also examine the influence of missing data.

The research also ignores certain issues. The research studies the dynamics of feelings such as arousal and valence over time. The optimization of the speed of the process is not required in this research. It is not necessary to use complicated time alignment methods, although the results for data mismatch are taken into account. The research also assumes the availability of computation resources.

1.6 Organization of the Thesis

The thesis begins with Chapter 1, introducing the background on Multimodal Emotion Recognition (MER), the challenges within the study context definition, objectives, the area covered within the study, and the outline of the entire thesis.

The study then proceeds to explore past research conducted within the area covered in Chapter 2 entitled Literature Review. In this chapter, the study examines past research that has been conducted within the area covered by various scholars who might have concentrated on similar areas within Multimodal Emotion Recognition (MER). Other areas that might be covered within this chapter include fundamental concepts within the

area covered, types and approaches to input within the area covered, data acquisition from the IEMOCAP corpus, as well as a review of recent fusion methodologies including Tensor-based and Graph Neural Network approaches. In this study, other areas that might be covered include current challenges within the area covered such as reliability within the study area among others.

In Chapter 3 entitled Methodology within the study, there is the description within the study that might include approaches to the development of the IEMOCAP dataset among other approaches to acquisition. Additionally, this chapter details the feature extraction pipeline, specifically focusing on the implementation of WavLM for audio, ModernBERT for text, and DINOv2 for visual data processing. The study also includes approaches to architecture that might include approaches to architecture within the study area among other areas. Finally, the chapter outlines the evaluation metrics used, primarily Weighted F1-score and Accuracy, to benchmark the performance of the proposed fusion models.

In Chapter 4 entitled Study Designs within the study area, there is the study designs within the area covered that might include approaches to study designs that enable tests within the study area to be set depending on the data that might be with the study. Additionally, approaches to models might include approaches to models within the study area. The study might include approaches to study settings that enable tests depending on data that might be within the study area among other settings.

In Chapter 5, entitled Implementation, details the development of the experimental pipeline. It covers the data preparation scripts, the execution of feature extraction using WavLM and DINOv2, and the setup of the training environment on the Kaggle platform.

1.7 Summary

Multimodal Emotion Recognition (MER) works better when it uses more than one input (modality) instead of just one. Deep learning methods are known for their advanced feature extraction and fusion algorithms. But there is still one big problem that is systems that work well with clean test data can break down a lot when they have to deal with noise or missing data in the real world. There are a lot of different ways to fuse, from simple early and late fusion to more complicated hybrid methods. Nonetheless, selecting reliable methodologies for practical applications is difficult due to the absence of comparative data regarding the reliability of these diverse approaches in handling inadequate data.

This project will quickly fill this gap in knowledge. This study compares early, late, and specific hybrid fusion methods (Attention-Based and Hierarchical) side by side. The study will look at baseline performance on clean data and how stable performance is during testing when there are missing modalities and fake noise. This study seeks to provide genuine insights into the reliability of various fusion strategies by employing realistic deep learning techniques on the standard IEMOCAP dataset under challenging conditions. The ultimate goal is to give suggestions based on evidence for making MER systems that are more reliable and useful in real life. The subsequent chapters comprehensively address the research's methodology, experiments, findings, and conclusions.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

2.1.1 Definition and Significance

Affective computing, which is the study of making machines that can sense, understand, and respond to human emotions, is a mix of computer science and psychology. Multimodal Emotion Recognition (MER) is an automatic way to find out how someone is feeling by looking at and combining data from more than one source at the same time. Some common modalities are language content (text), voice acoustics (audio), facial expressions (visual), and, in some cases, physiological signals (H. Lian et al., 2023).

MER has become more important over the past ten years because it could change the way people and computers interact in many different ways. Some examples of these kinds of applications are adaptive learning software that changes based on how the learner feels, automated mental health monitoring, emotionally intelligent virtual agents, safer human-robot collaboration, better customer service interactions, and more in-depth analysis of how customers respond for business. Advancements in computer science are augmenting human expectations for computer systems that can engage with

individuals in a natural manner; a key aspect of this ambition is the capacity to perceive and respond to emotional signals, akin to human behavior (Hu et al., 2024).

2.1.2 Shifts from Unimodal to Multimodal Emotion Recognition

The early work in building a computer reading affect has largely been restricted to a uni-modal approach—one type of information at a time: text, audio, or images of faces. Single modality systems share some easily identified disadvantages. Carrying a hod of strong emotions is easy for a neutral face to conceal, and tone of voice is lost if words alone are conveyed. In addition, each type of information is hampered by a unique distraction—one might obscure a face, for example, and sound quality could be contaminated by ambient noises. It is, in fact, a multidimensional process in which people combine a multitude of linguistic and non-linguistic communications (Mansoorizadeh & Charkari, 2009).

This understanding led scientists to adopt multimodal approaches. The benefit offered by MER is the ability to access several modalities to gain a better understanding of the user's emotions. Facial expressions may indicate the nature of the emotion, prosody may indicate the intensity of the emotion, while the verbal contents may indicate the source of the emotion. The benefits offered by MER are enhanced through cross-modal redundancy while dealing with the unavailable modality. Cross-modal redundancy is stated to be achieved in the year 2020 by Cimtay et al.

The trend towards the use of multimodal approaches represents the need for systems capable of capturing the subtlety associated with emotional processing. Even though the main aim of MER is to simulate the multi-channel processing involved in emotion processing in the human (H. Lian et al., 2023), the process still appears to be difficult.

2.1.3 Types of Input Used in MER

Multimodal Emotion Recognition (MER) combines data from diverse human expression modalities. The choice of modalities to be incorporated is an important task and varies depending on the application environment and the data types that are accessible. In this study, the most commonly employed modalities within MER are taken into consideration. These include text, speech, and facial expressions.

Some of the ways that speech audio conveys key emotional information include prosodic parameters (pitch, energy, and speech rate), voice quality parameters, and spectral parameters. These audio parameters are capable of portraying much about the emotional state and how excited an individual is. Xu et al. (2021) indicate that resilience to audio degradation is essential in practical applications since audio features are more sensitive to noise. Mel-Frequency Cepstral Coefficients are another popular audio feature. More modern approaches such as WavLM (Chen et al., 2022) use self-supervised learning models to obtain better features from audio. The contextual embedding components in WavLM are sensitive to content as well as paralinguistic features, making WavLM outstanding in recognizing emotional features even in environments with noise.

Facial expressions are the best indicator of how someone is feeling. As stated by Zhu et al. (2024), facial expressions can be impacted by factors such as poor lighting, occlusions, and varying head orientations. Though deep convolutional neural networks are already the norm in the industry with regard to facial expression analysis, the DINOv2 deep learning model Oquab et al. (2023) is an impressive self-supervised vision transformer that is capable of learning robust and universal visual representations without the use of labeled data. The patch embedding property of the DINOv2 model is excellent for analyzing subtle expressions found in facial expressions to ascertain how someone might be feeling.

Textual data carries immediate semantic meaning whether in written form or transcribed from spoken utterances. Analyzing the use of words in sentences and meaning implied by various references can aid one grasp one's own feelings. The use of text data without additional contextual analysis proves to be effective with conversational

data; still, such analysis may overlook paralinguistic features that are vital to holistic analysis of the emotional construct as shown by Makiuchi et al. (2021). The transformer architecture represented by ModernBERT Warner et al. (2024) has transformed the area devoted to text analysis tasks incorporating connections between various words. ModernBERT has extensive long context, rapid inference, and excellent embedding quality that proves superior to other models such as BERT and DistilBERT. These three are the key means whereby feeling is expressed in spoken dialogue and thus form the foundation for this study. These interact well with each other and are excellent approaches to the study of fusion approaches, particularly in situations that are robust to one or more modality types possibly being unavailable due to heavy levels of noise.

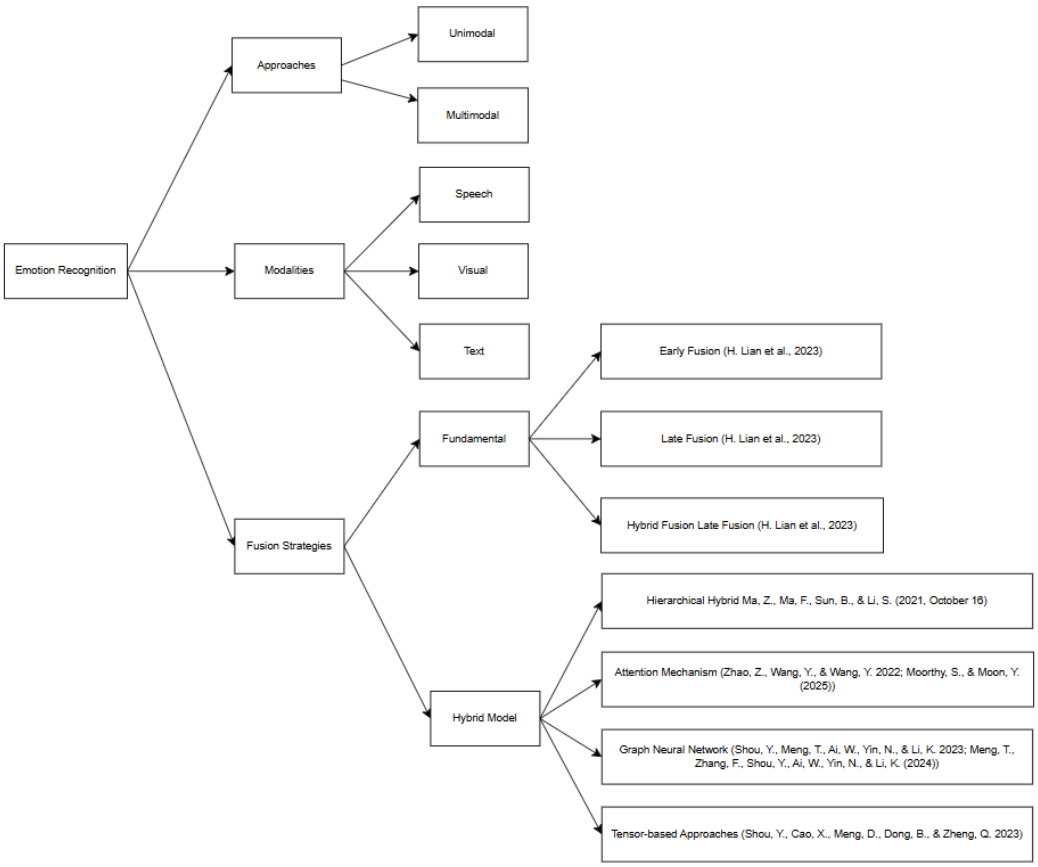


Figure 2.0 : TAXONOMY OF EMOTION RECOGNITION RESEARCH

2.2 Feature Extraction

One of the most critical early steps in Multimodal Emotion Recognition (MER) is feature extraction, which converts unstructured input data to structured and informative representations to be then analyzed by machine learning models. The area has shifted primarily towards deep learning techniques, although early MER research was typically built atop hand-designed features rooted in domain expertise. Such techniques can learn automatically hierarchical features and informative patterns in the data, typically with better performance than traditional techniques (H. Lian et al., 2023).

Strong deep learning models pre-trained on large datasets are commonly used by today's MER systems to extract good features across all modalities. Effective, yet efficient, deep learning-based feature extraction approaches for text, vision, and audio modalities are discussed in the subsequent sections, which we will use in our experimental setup, in line with the thesis focus on comparing approaches to fusion under possibly constrained computational budgets.

2.2.1 Audio Features

Recently, various researches involving self-supervised learning have significantly contributed to improving Speech Emotion Recognition methods beyond traditional hand-crafted features such as MFCCs. WavLM is one such model that stands at the forefront of being a pre-trained full-spectrum speech-based large model capable of performing more than emotion recognition tasks such as speaker verification, diarization, and speech separation (Chen et al., 2022).

In Wav-LM, the method used is masked speech denoising and prediction. This allows it to train on content and paralinguistic information from raw audio. This helps the model capture the subtle expression of emotions, as well as perform well under the challenges of noise and overlapping speakers in SER.

Some of the key concepts include the use of a transformer encoder with the use of gated relative position bias for better representation of time. This model was trained on an incredibly large dataset of 94,000 hours of varied, unlabeled audio data from sources such as LibriLight, GigaSpeech, and VoxPopuli. It is important to note that the pre-training data incorporated noisy and overlapping speech.

On performance grounds, WavLM Large outshines the SUPERB literature on emotion recognition tasks, outperforming state-of-the-art models such as HuBERT and wav2vec 2.0. It is interesting to note that in most SER tasks, the performance of WavLM Base+ outshines larger models. This highlights the importance of training strategy and design.

In Multimodal Emotion Recognition tasks, it is beneficial for a model to have good functionality with both images and text. The contextual embeddings obtained by WavLM are very useful for Multimodal Emotion Recognition tasks. The model extracts acoustic features for emotion from the waveform and thus doesn't require any preprocessing.

2.2.2 Visual Features

DINOv2 is currently the best self-supervised vision model and supports computers in learning robust visual features using no labeled data and text supervision. DINOv2 is a product of Meta AI and proves that quality visual understanding is achievable without supervision.

A key factor in its success is the training dataset. The model is trained on a thoughtfully curated dataset of 142 million images (LVD-142M), which is intended to be extremely diverse. This is crucial because the more diverse the dataset is, the better the model will be at various tasks such as classification, segmentation, depth estimation, and instance recognition, which are essential in the design of an emotion recognition system.

In terms of techniques, the addition of numerous key concepts makes DINOv2 technically different from other approaches. It combines both the objectives of DINO and iBOT for enhanced self-supervised learning. This assists the model in learning small facial expressions and emotions. Apart from this, it applies techniques such as KoLeo regularization and Sinkhorn-Knopp centering for stabilizing the features, along with FlashAttention and sequence packing for optimized training while dealing with large amounts of data while consuming minimal amounts of memory.

Experimental results prove that DINOv2 is superior to its past self-supervised and weakly-supervised predecessors. Its features are transferable and can be successfully combined with linear classifiers. Hence, it can be very useful for multimodal emotional recognizers that have to smoothly integrate imagery with audio or text information.

DINOv2 is quite flexible and adaptable to various scenarios. DINOv2 is robust and performs well even when there is ambient noise and missing information. Therefore, it is a reliable tool for identifying emotions even under real-world conditions. DINOv2 is appropriate for emotion recognition systems because of its ability to perform well with various datasets and tasks. This is especially important if you are looking for a good source of visual features but with low supervised learning requirements.

2.2.3 Textual Features

New methods of building transformer encoder models are revolutionizing the way MER systems identify emotions in texts. These methods are also altering ways in which natural language processing occurs. They are effective in identifying emotions in texts through the detection of minute indicators in lengthy texts using self-attention mechanisms. ModernBERT is a transformer encoder model that works more efficiently compared to past models.

ModernBERT is an improvement on existing models including BERT, RoBERTa, and DistilBERT. ModernBERT was developed by authors Warner et al. in 2024 and

includes new elements such as GeGLU activations, rotary positional embeddings (RoPE), and both global and local attention. ModernBERT can handle long sequences with a maximum length of 8192 tokens and supports faster and lower-memory computations.

ModernBERT has strong results on a number of NLP tasks such as classification on the GLUE benchmark and retrieval on the BEIR benchmark. It actually outperforms DeBERTaV3-base on the GLUE benchmark, the first MLM-trained encoder to ever do so.

Its context-related embedding facilitates emotion recognition since it is able to capture linguistic features related to emotions and is even fast enough for MER applications that require real-time processing. ModernBERT embedding is capable of combining various features from different sources such as audio and visual sources; thus, ModernBERT embedding would be appropriate for MER applications that are involved with multiple sources of information. It is still effective even with missing or ambiguous information.

ModernBERT is a big improvement over its predecessor, DistilBERT, which gained a lot of favor for being very light and comparable to BERT. ModernBERT works on improved accuracy, is capable of processing extended contexts, and increases the speed with which inference takes place, all while being GPU-optimized. ModernBERT provides a great starting point for contemporary MER systems to be emotionally intelligent and resource-optimized.

2.3 Fusion Strategies

2.3.1 Foundational Strategies

Multimodal Emotion Recognition (MER) relies on the general idea that the integration of information from numerous sources (text, audio, and vision) leads to better emotion detection compared to unimodal systems. The fusion process, the actual procedure

through which these multifarious data streams are integrated, plays a significant role in the performance and robustness of the system. The theoretical basis of more advanced procedures is established by three fundamental fusion approaches: Early, Late, and Hybrid fusion.

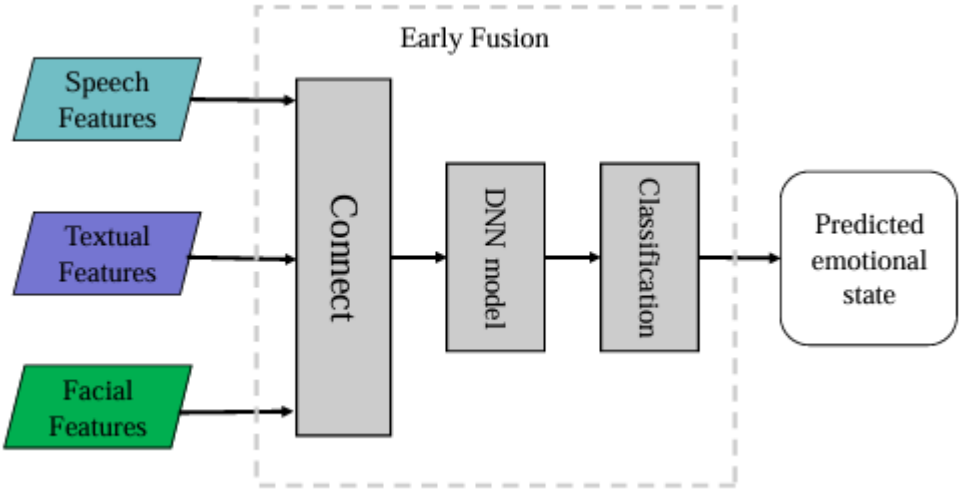


Figure 2.1 : MULTIMODAL EMOTION RECOGNITION FRAMEWORK BASED ON EARLY FUSION

Source : H. Lian et al., (2023)

Early fusion takes place before the classification phase, combining features from separate modalities into one unified, high-dimensional vector (H. Lian et al., 2023). By doing this, the classifier can spot cross-modal connections right at the feature level, a benefit originally pointed out by Wagner et al. (2011).

Nevertheless, there are a few issues with early fusion that might affect robustness. First of all, early fusion generally requires precise temporal alignment across modalities, which is hard to accomplish with real-world data, as stated by Cimtay et al. (2020). Concatenated feature vectors are potentially very high-dimensional, and learning may become more difficult and more overfitting-susceptible. Second, the loud or overwhelming character of a modality might mask informative information from the other

modality. Third, and most importantly for robustness, all the modalities are required to be available at test time; if one is unavailable, the model might not work unless some effort is made to impute or reconstruct missing information (Cimtay et al., 2020).

In particular, Goncalves and Busso (2022) studied the vulnerability of early fusion to missing modalities and found that when one or more of the input streams are absent during test time, the performance declined significantly. For systems that are designed to be utilized in unconstrained environments where sensor failure or problems with data communications may happen, this vulnerability is a critical challenge.

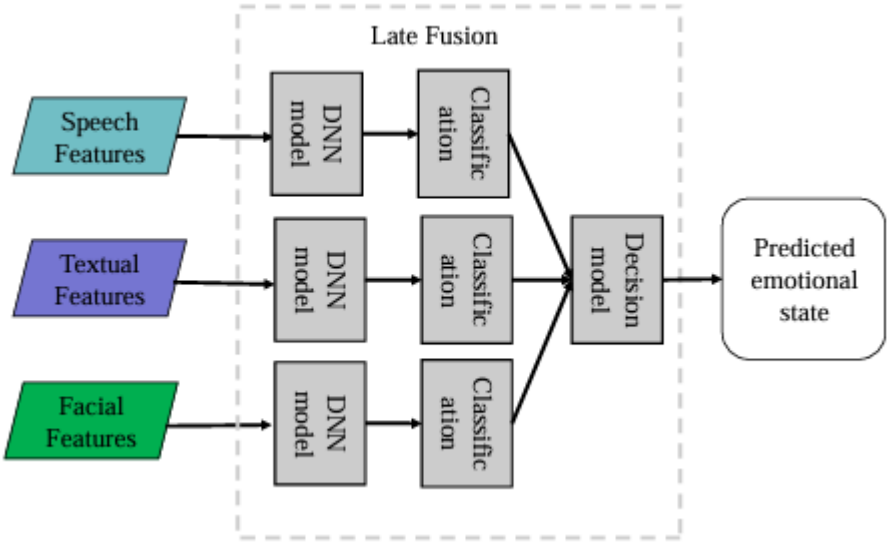


Figure 2.2 : MULTIMODAL EMOTION RECOGNITION FRAMEWORK BASED ON LATE FUSION

Source : (H. Lian et al., 2023)

Through learning individual classifiers per modality and averaging their outputs (predictions or probabilities) afterwards, late fusion follows a modular decision-making strategy (H. Lian et al., 2023). Averaging, weighted voting, or a meta-classifier whose

inputs are the outputs of the individual classifiers are some ways to obtain this combination.

Late fusion enjoys several robustness advantages. Because a decision can be made based on modalities available at output, missing modalities at test time are addressed in this approach, as noted by Cimtay et al. (2020). One can make use of the optimum classifier design per data type and refrain from exact temporal alignment of features from the various modalities. Late fusion was shown by Goncalves and Busso (2022) to be more resilient to missing modalities compared to early fusion approaches.

Late fusion's largest flaw, pointed out by Mansoorizadeh and Charkari (2009) and echoed by H. As pointed out by Lian et al. (2023), because every modality is separately processed until the final fusion, it has to ignore potential useful cross-modal interactions in the initial processing stages.

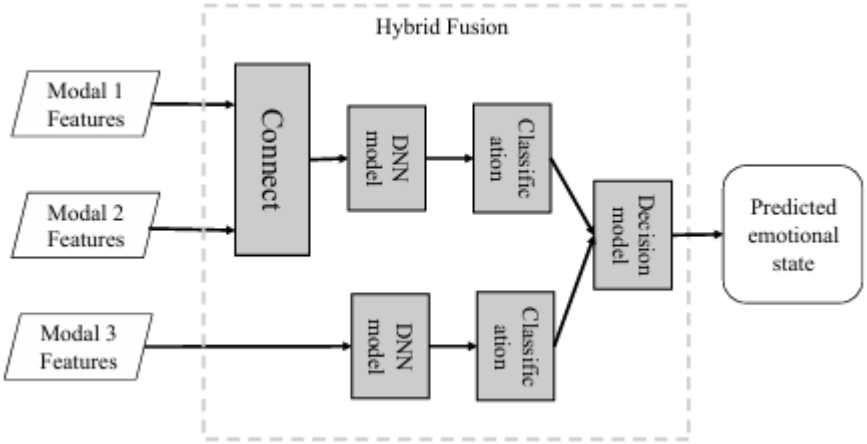


Figure 2.3 : MULTIMODAL EMOTION RECOGNITION FRAMEWORK BASED ON HYBRID FUSION

Source : H. Lian et al., (2023)

Hybrid fusion techniques seek to avoid the shortfalls of early and late fusion and also integrate their benefits (H. Lian et al., 2023). For providing robustness as well as flexibility and also modeling cross-modal interactions where appropriate, such techniques tend to fuse information at various stages of processing.

Unlike simple early or late fusion approaches, Ma et al. (2021) introduced a hybrid fusion model that combines textual features later after first combining auditory and visual features at the intermediate level. This is better. Likewise, Sun et al. (2023) was more noise-resistant than traditional early fusion by relying on auxiliary tasks to facilitate more powerful cross-modal learning without losing modality-specific information. To improve modal robustness for noisy or absent modalities, Chen and Zhang (2024) proposed a "Modality-Collaborative Transformer" combining hybrid fusion with a feature reconstruction approach.

In comparison to early and late fusion baseline methods, their experiments showed better performance on a variety of challenging configurations. Compared to pure early or late fusion methods, hybrid fusion methods introduce more architectural complexity, although offering more design flexibility (H. Lian et al., 2023). This paper considers the most significant scientific challenge of understanding the relative trade-offs of various hybrid fusion architectures, namely noise and missing data robustness.

2.3.2 Advanced Fusion Techniques

How fusion is performed has been revised because of deep learning. And now it is not only about combining the data; it is about dealing with the messiness of the input.

Attention facilitates this process. It enables models to disregard noise while concentrating on the significant aspects. For instance, R et al. (2024) illustrated that cross-modal attention enables audio and video to complement each other. If the video is unavailable or the audio is noisy, the other modality can supplement the process. A study conducted by Zhao et al. (2022) in the area of voice and texts incorporated self-

attention to integrate the features of Wav2Vec 2.0 and BERT. The study was more effective than traditional approaches, especially in the absence of strong emotional cues. Moorthy and Moon (2025) progressed from this study with the development of the hybrid model that incorporates attention mechanisms in varying phases to deal with imbalanced data in the context of audio-visual models.

However, apart from attention, another area of interest for solving such complex linkages is the use of Graph Neural Networks (GNNs). Shou, Meng, and colleagues were able to demonstrate that graphs are capable of capturing dynamism. Others, such as Meng et al. in 2024, did so from the perspective of signal processing, where the graph structure itself can help impute missing values.

There is also work involving tensors, which represent links between modalities with extensive mathematical work. The drawback is the high computational complexity; there was, however, a way out when Shou, Cao et al. (2023) crafted the "Low-rank Matching Attention," which combines both tensors and attention mechanisms for performance benefits while avoiding the penalty of increased complexity. Even the Large Language Models are entering this space since Cheng et al. (2024) applied instruction tuning for this purpose.

"More complex, however, is not necessarily better." This is because there is no single best approach. It is a matter of what researchers are going to use it for and the hardware you have. Sometimes, state-of-the-art solutions work well, though "heavy and complex is rarely necessary." This is why this research remains committed to basic techniques such as Early, Late, and Hybrid. It is a desire to see how these basic solutions work for real-world issues such as the effects of noise and missing values.

2.4 Benchmark Datasets

Shared datasets offering a balanced playing field for relative performance comparison between studies hold the key to MER technique development and assessment. It is the

IEMOCAP corpus, today a de facto standard for evaluation of multimodal emotion recognition systems, that my research focuses on.

The IEMOCAP (Interactive Emotional Dyadic Motion Capture) corpus (Busso et al., 2008) is among the most popular benchmarks for evaluating MER systems in English conversational scenes. The corpus includes over 12 hours of audiovisual recordings from 10 actors who took part in five sessions of dyadic conversations. Acted scenes and improvisation tasks aimed at evoking spontaneous emotional feelings throughout the interactions are included within the set. The set includes chronologically aligned audio, manual transcriptions, and video enhanced with face motion capture.

Categorical labels of emotions on nine categories (angry, pleased, sad, neutral, irritated, thrilled, afraid, startled, and disgusted) and "other," and dimensional ratings (valence, activation, and dominance) are included in descriptions. In classification tasks, researchers usually combine the original labels into a smaller set of four to six fundamental emotion classes (Zaidi et al., 2023). Typical practice includes bundling similar emotions; for example, "angry" and "frustrated," and also "happy" and "excited," get combined. The focus of the study usually lies in determining the specific classes that are selected.

To support speaker-independent testing, the benchmark testing protocol uses Leave-One-Session-Out (LOSO) cross-validation (Patamia et al., 2023). This is significant because it evaluates the models' capacity to generalize to novel speakers whom they had not encountered during training, which is a more realistic situation in real-world settings for applied purposes. Owing to its multimodality, continuous annotation, and presence of the three modalities (text, visual, and audio) required for our fusion comparison strategy, IEMOCAP is still a viable comparative benchmark even if it is acted.

The multimodality of the IEMOCAP dataset makes it particularly amenable to our exploration of robustness of fusion. With multimodality, one can exhaustively try early, late, and hybrid forms of fusion. With available high-quality audio, video, and text data, system performance can be systematically analyzed even when some of these modalities

are absent or are noisy. Our robustness findings of fusion will also carry over to real-world applications where systems have to generalize to novel speakers following the popular LOSO cross-validation strategy.

2.5 Robustness and Current Challenges

It is still very difficult to recognize multimodal emotions with human-like performance, especially in unrestricted real-world situations. Reliability must be demonstrated by systems designed for real-world applications by managing inconsistent data and different settings. The goal of current research is to improve the reliability of MER systems by tackling a number of significant issues that are closely linked to robustness.

2.5.1 Robustness to Noise and Data Corruption

The occurrence of noise in real data is a challenging issue in real MER systems. Noise robustness is a highly relevant research area because environmental noise in speech is very prevalent, according to M. Xu et al. (2021). As they depicted in their work, they systematically experimented on the impact of various types of noise and noise levels on the performance of MER. Their work emphasizes the requirements for focused robustness training by showing that different types of noise impact system performance differently.

Noise refers to irrelevant data to the main emotional content and can severely impair system performance, as argued by Y. Liu et al. (2023). They came up with a Noise-Resistant Multimodal Transformer that benefits from new attention mechanisms and training paradigms to ensure performance under noisy situations.

A "noise scheduler" that adds configurable types (Gaussian, impulse) and levels of noise to features during training in an attempt to mimic real-world conditions was suggested by Fan, Zuo, et al. (2024) as part of their overall solution to noise robustness.

Their empirical experiments illustrated that models trained on simulated noise test substantially better on noisy inputs than baseline models trained on clean data alone.

Constructing noise-robust fusion structures, noisy-example training (M. Xu et al., 2021; Fan, Zuo, et al., 2024), constructing noise-robust feature extractors (Y. Liu et al., 2023; Fan, Zuo, et al., 2024), and using special training methods like noise-aware learning (Y. Liu et al., 2023) or learning noise-robust joint representations (Fan, Zuo, et al., 2024) are common methods to counter the impact of noise.

Despite these developments, this thesis aims to fill an important research void by comparing systematically the performance of the Early, Late, and Hybrid fusion methods under various noise conditions.

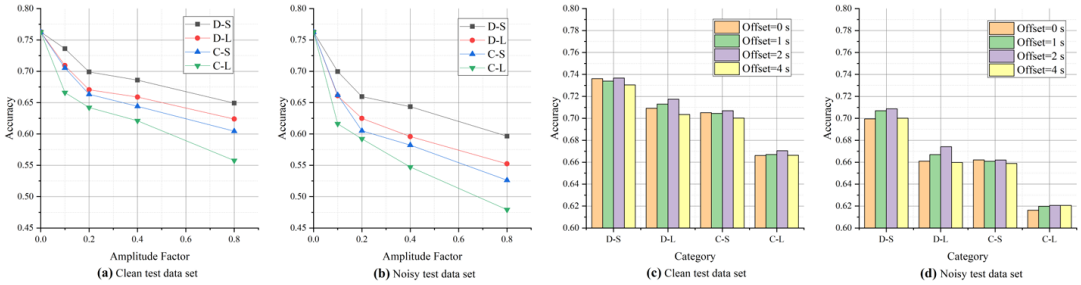


Figure 2.4 : IMPACT OF DIFFERENT NOISE CATEGORIES AND INTENSIFIES (AF) ON ACCURACY

Source : M. Xu et al., (2021)

2.5.2 Robustness to Missing Modalities

Being able to process cases of missing or unavailable input modalities is also important for real-world MER systems. Sensor malfunction, data corruption, network outage, or privacy requirements can all render the data entirely unavailable (Guo et al., 2024; Fan, Zuo, et al., 2024). This total lack of expected input is a serious challenge because

standard MER models, typically trained with an expectation of full data availability, perform very badly when tested with missing input (Guo et al., 2024).

Training models with missing modalities through rapid learning techniques was investigated by Guo et al. (2024). Their research suggested that there would be a need for special prompting techniques to effectively inform the conditions of missing data, and it would not be enough to simply remove modalities randomly throughout training. Fan, Zuo, et al. (2024) looked into mechanisms of learning joint representations that are naturally robust to missing data and filling in missing features from observed modalities. Based on their experimental findings, specially designed models to work for missing modalities can still function well even if one or more of the channels of input are lost. A demonstration from recent studies showing how performance varies across datasets and protocols when some of the modalities are missing is presented in Table 2.1.

Table 2.1: PERFORMANCE COMPARISON UNDER VARIOUS MISSING MODALITY CONDITIONS (E.G., {A}, {V}, {T}, {A,V}, ETC.)

Dataset	Method	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{a,t\}$		$\{v,t\}$		Avg.	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
IEMOCAP	LB	46.35	46.21	48.07	47.58	56.06	55.28	58.12	57.89	72.18	72.25	65.63	65.28	57.74	57.42
	MS	47.65	47.52	47.68	47.36	59.27	59.22	57.48	56.60	72.30	72.18	66.81	66.93	58.53	58.30
	MD	48.22	48.09	48.26	47.98	61.26	61.28	58.08	57.96	72.40	72.31	67.08	68.22	59.22	59.31
	MCTN	51.62†	-	45.73†	-	63.78†	-	55.84†	-	69.46†	-	68.34†	-	59.19†	-
	MMIN	59.00†	-	51.60†	-	68.02†	-	65.43†	-	75.14†	-	73.61†	-	65.47†	-
	MPMM	58.69	57.66	55.18	55.36	68.39	68.08	63.68	63.47	74.90	74.98	73.80	72.67	65.77	65.37
	Ours	**59.77**	**59.71**	**57.61**	**56.98**	**69.23**	**69.28**	**67.26**	**67.37**	**75.98**	**75.44**	**74.68**	**74.51**	**67.42**	**67.22**

Source : Z. Guo et al., (2024)

Table 2.2: RESULT UNDER DIFFERENT NOISE LEVELS (GAUSSIAN/IMPULSE), COMPARING MODELS INCLUDING BASELINES TRAINED ONLY ON COMPLETE DATA (MEN) VS. THOSE TRAINED WITH INCOMPLETENESS SIMULATION.

Dataset	System	-10dB (Avg)				-20dB (Avg)				-30dB (Avg)				-40dB (Avg)			
		Gaussian noise		Impulse noise		Gaussian noise		Impulse noise		Gaussian noise		Impulse noise		Gaussian noise		Impulse noise	
		WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
IEMOCAP	MEN	0.7120	0.7175	0.7448	0.7453	0.6747	0.6714	0.7026	0.7133	0.6370	0.6349	0.6758	0.6850	0.6047	0.6004	0.6348	0.6364
	MCTN	0.7215	0.7431	0.7400	0.7531	0.6989	0.7003	0.7270	0.7314	0.6778	0.6712	0.7112	0.7157	0.6674	0.6790	0.6978	0.7041
	MMIN	0.7551	0.7640	0.7750	0.7870	0.7271	0.7394	0.7521	0.7671	0.7155	0.7209	0.7397	0.7516	0.7003	0.7067	0.7226	0.7352
	IF-MMIN	0.7543	0.7655	0.7738	0.7858	**0.7345**	**0.7493**	0.7589	0.7719	0.7184	0.7225	**0.7382**	0.7509	0.7048	0.7155	0.7343	0.7452
	Ours	**0.7598**	**0.7675**	**0.7773**	**0.7872**	0.7307	0.7406	**0.7651**	**0.7790**	**0.7197**	**0.7284**	0.7304	**0.7532**	**0.7085**	**0.7176**	**0.7391**	**0.7503**
	w/o VAE	0.7382	0.7275	0.7509	0.7631	0.7004	0.7132	0.7467	0.7500	0.6850	0.6882	0.7136	0.7280	0.6793	0.6801	0.7133	0.7238
	w/o $\mathcal{L}_{inv}$	0.7401	0.7338	0.7630	0.7712	0.7121	0.7293	0.7489	0.7549	0.7010	0.6942	0.7202	0.7276	0.6843	0.6805	0.7158	0.7208

Source : Fan, Zuo, et al., (2024)

Table 2.3: DETAILED RESULTS FOR THEIR MODEL UNDER DIFFERENT SPECIFIC MISSING MODALITY COMBINATIONS (E.G., ONLY AUDIO CLEAN, ONLY VISUAL CLEAN, ETC.) WITH GAUSSIAN NOISE.

Dataset	Intensity	a		v		l		a, v		a, l		v, l		Avg	
		WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
IEMOCAP	-10dB	0.7275	0.7349	0.7440	0.7503	0.7750	0.7820	0.7432	0.7508	0.7813	0.7917	0.7877	0.7952	0.7598	0.7675
	-20dB	0.6752	0.6873	0.6911	0.6949	0.7623	0.7735	0.7068	0.7178	0.7720	0.7846	0.7771	0.7864	0.7307	0.7406
	-30dB	0.6574	0.6626	0.6715	0.6688	0.7513	0.7592	0.7021	0.7071	0.7649	0.7778	0.7712	0.7771	0.7197	0.7254
	-40dB	0.6347	0.6484	0.6529	0.6511	0.7348	0.7484	0.6978	0.7024	0.7632	0.7777	0.7672	0.7773	0.7085	0.7176

Source : Fan, Zuo, et al. (2024)

This study is especially interested in how well various fusion algorithms function when there are missing modality circumstances. Although late fusion modularity inherently provides it with some level of robustness, further effort is needed to determine the level at which early and hybrid fusion techniques react under missing modality conditions. This research aims to offer empirical evidence of this critical MER system reliability factor.

2.6 Research Gaps Addressed by This Study

The literature review shows that there are some critical research gaps in multimodal emotion recognition in terms of the effectiveness of various fusion approaches in practical applications. A significant number of approaches to fusion have been designed, from elementary early and late fusion to more sophisticated ones. But there are no comprehensive assessments in terms of the impact of noise and incomplete data.

The study examines these research gaps:

1. People do not possess enough credible evidence to compare the performances of early fusion techniques, late fusion techniques, and hybrid fusion techniques in similar noise environments for audio and visual modalities. M. Xu et al. and Liu's group proposed techniques to counter noise in 2021. A complete analysis has still not been carried out; even then, Fan et al. proposed techniques in 2023 and 2024.

2. A clear comparison between Early and Hybrid approaches in a controlled environment needs to be made while excluding certain modality types. The advantages of Late fusion techniques have already been highlighted (Cimtay et al., 2020; Gonçalves & Busso, 2022), though these are still important. This thesis analyzes the functioning and versatility of various types of fusion approaches such as Early, Late, and Hybrid within simulations with added noise and incomplete data.

This research offers practical guidelines on implementing fusion techniques within real-world MER systems. Various feature extraction techniques and analysis approaches have been made use of within this study on the IEMOCAP dataset.

Table 2.4: COMPARISON OF STATE-OF-THE-ART APPROACHES ON CLEAN DATA (HAPPY, SAD, NEUTRAL, ANGRY, EXCITED, OR FRUSTRATED.)

Hybrid Fusion Method	Weighted Accuracy (%)	Weighted F1 (%)	References
GraphSmile	72.77	72.81	Li et al. (2024)
M^3Net	72.46	72.49	F. Chen et al. (2023)
Mamba-like	73.10	73.30	Shou et al. (2024b)
SDT	73.95	74.08	H. Ma et al. (2023)
CFN-ESA	70.78	71.04	Li, Wang, Liu, et al. (2024)
GS-MCC	73.8	73.9	Meng et al. (2024)
ELR-CGN	70.6	70.9	Shou, Ai, et al. (2024)

2.7 Summary

This review of the literature investigates the significant developments in Multimodal Emotion Recognition (MER) systems. It depicts the evolution of the use of one source of data to the usage of multiple sources, with the importance of designing systems that are more robust for real-world applications. It highlights the key developments in the feature extraction for the multiple data sources, such as ModernBERT for text-based data, DIVOv2 for the analysis of the facial expressions, and WavLM for audio data. It explores the key fusion approaches: Early, Late, Hybrid, highlighting the pros and cons for scenarios involving the presence of noise or absent data. The review explains the multiple approaches for the fusion of multiple methods, pointing to some of the most sophisticated approaches. It mentions the current issues in MER systems, pointing to the multiple approaches that were developed to resolve them. A key issue is that there is no comprehensive study in the fusion approaches in the same manner as in real-world applications. This paper aims to examine the performance of Early, Late, Hybrid approaches in the presence of missing data, visual, and audio noise. It aims to contribute towards enhancing the accuracy of MER systems for applications in real-world scenarios by investigating the effectiveness of the multiple approaches in the context of real-world applications.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter goes into great detail about the methodological setup that was used for this research. It begins by talking about how the IEMOCAP dataset was chosen and processed. This dataset was used as the basis for all the experiments. Next, there is a clear explanation of the specific deep learning models that were used to extract features from images, audio, and text. The chapter then talks about the architectural designs for the fusion approaches that were looked at which are the basic Statistical (Early, Late, Hierarchical, Gated) models and the more advanced Contextual Hybrid models (Bi-GRU variants and SuperTransformer). The chapter finally talks about the cross-validation method and metrics that were used to systematically check how well the model worked and how stable it was.

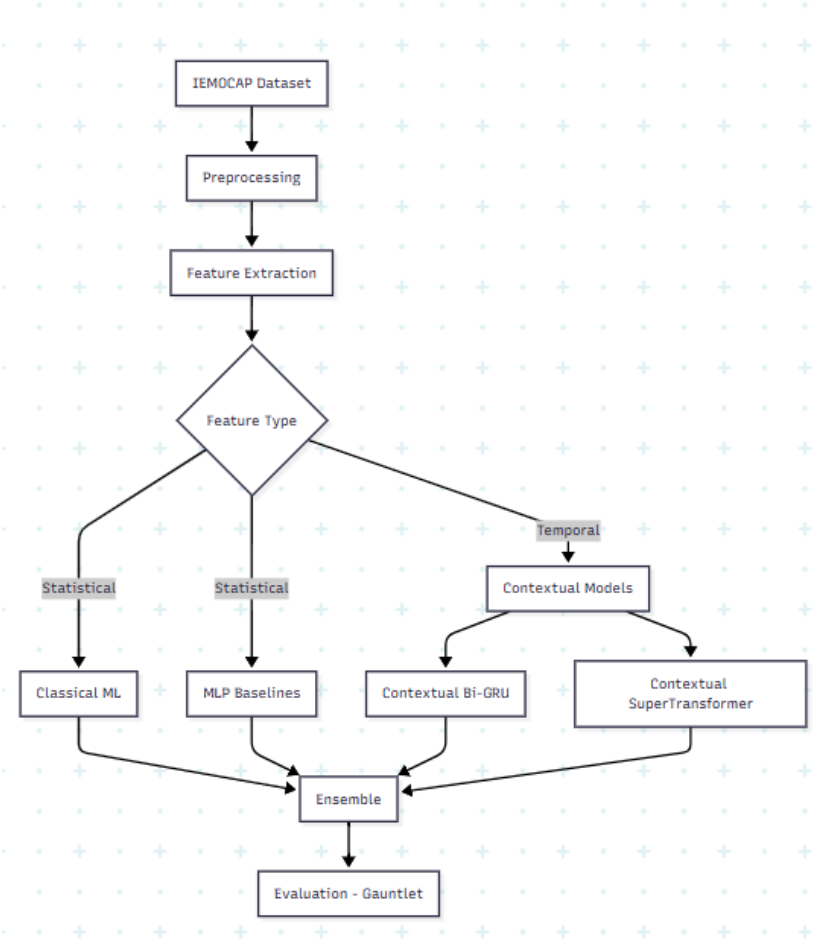


Figure 3.1 : FLOWCHART OF THE PROJECT METHODOLOGY

3.2 Dataset and Preprocessing

3.2.1 Dataset Selection and Description

This research study makes use of the very popular IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset. IEMOCAP was selected due to the fact that it is usually used as a benchmark for MER studies, and therefore it would be easy to compare with existing studies. The dataset contains around 12 hours' video recording data of ten actors

in five dyadic recording sessions. The recording sessions cover scripted and improvisation material designed to promote natural emotional reactions. Most importantly for this study, IEMOCAP contains time-aligned speech audio, facial expression video, and text transcriptions, making it an ideal resource to investigate multimodal fusion methods.

3.2.2 Emotion Label Processing

In the IEMOCAP dataset, every spoken segment was evaluated and assigned an emotion label. The initial emotion categories included angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and "other." This study looks at classifying different emotions, specifically happy, sad, neutral, angry, excited, and frustrated. This choice matches methods from earlier related studies. This study excluded any statements labelled with disagreement ('xxx') or those not fitting into the selected categories from this analysis.

3.2.3 Dataset Segmentation

The analysis is carried out on individual spoken units with ‘utterances’ as the naming convention that follows the convention set by the standard research data set IEMOCAP. The data was taken from the pre-segmented utterance file that comes with the IEMOCAP data set (SessionX/sentences/Ses01F_impro01_M000)

3.2.4 Cross-Validation Strategy

In this research, Leave-One-Session-Out (LOSO) cross-validation was used to guarantee that there was no overlap in the speakers to be tested and those during training. The IEMOCAP dataset has five sessions with each session consisting of a different pair of actors. In this research, the data will be from the final session only. The data from the

remaining sessions is further split to form 60% and 40% to train and validate in the LOSO cross-validation process. The whole process is repeated five times to ensure that each session has been tested only once. In this research, learning was carried out by dividing the data among the remaining sessions to train and validate in each round of the cross-validation process. This might aid in applying the research to techniques such as Early Stopping.

3.2.5 Preparing Each Type of Input Data

Standard preprocessing was done on each modality to make sure they were all the same and worked with the deep learning architectures. Firstly, all audio files were changed to 16 kHz. The stereo channels were combined into one. Then, the acoustic feature extractor took care of normalization by taking the processed waveforms. Secondly, a Haar Cascade Classifier was used to find the areas of each video frame that had faces. To deal with the fact that IEMOCAP videos have more than one speaker, the active speaker was found using the naming convention for utterances. The active speaker's bounding box was cropped with padding to keep the context and made to be 224×224 pixels. To capture facial dynamics, frames were taken at regular intervals (about 20 frames per utterance). Lastly, the transcripts were cleaned up to get rid of annotation artefacts like [LAUGHTER] and [BREATHING]. The text tokenizer did the tokenization, making sequences of different lengths (which were padded during training) to keep the full linguistic context for transformer models.

3.3 Feature Extraction

This step takes the preprocessed input data and turns it into structured, useful feature representations. This study used the best pre-trained foundation models for each modality to find a balance between performance and computational feasibility.

3.3.1 Visual Feature Extraction

This study used the DINOv2 architecture to get visual features. DINOv2 is a self-supervised Vision Transformer made by Meta AI that learns strong visual representations without needing labelled data (Oquab et al., 2023). This study used the dinov2-base version, which makes embeddings with 768 dimensions.

This study kept taking features from the [CLS] token in the last transformer layer. To facilitate various modelling methodologies, two types of representations were created which are statistical representations, which encompassed aggregated statistics (mean, standard deviation, maximum, minimum) across the temporal dimension for baseline models, and temporal representations, which maintained the complete sequence of frame embeddings for contextual models.

3.3.2 Audio Feature Extraction

This study utilized WavLM (Chen et al., 2022) for extracting the acoustic features. WavLM introduces a speech prediction task with denoising to Wav2Vec 2.0, which assists in paralinguistic tasks such as emotion recognition. The wavlm-base-plus variant was utilized of the model, with 768 features.

To incorporate both the low-level prosody and high-level semantics, features were extracted from the last three transformer layers (layers 10, 11, and 12). Two types of representations were built, as in the visual component of the pipeline description. The first type involves the statistical features (Mean, Std, Max, and Min) extracted over time for simple baselines. The second involves the temporal features where the entire frame-level sequence was maintained.

3.3.3 Text Feature Extraction

For linguistic properties, I employed ModernBERT (Warner et al., 2024). This is BERT with improved speed and simplicity (1.7x faster inference) and a superior ability to grasp extended contexts. The model with 768 properties I applied was called modernbert-base.

I extracted features from the last three layers (10, 11, and 12), similar to audio. The representations involved statistical summaries of the token sequence for baseline tests and sequences of token embeddings, retaining the word order and syntax for contextual learning.

3.3.4 Interpreting Emotional Features

The models were picked based on the fact that they addressed various aspects of human emotion. The first one, WavLM (Audio), focused on the excitement level of the person based on the prosodic features of speech, such as the rate of speaking. Additionally, the second one, DINOv2 (Visual), analyzed the emotion from facial expressions in terms of the action units, like the movement of the lip corners, as well as the strength of the action units. The third one, ModernBERT (Text), interpreted the meaning as well as the associated emotion based on the arrangement of the words. For instance, it would differentiate between the two phrases below.

3.4 Multimodal Fusion Architecture

The main focus of this study is to compare four types of classification methods, starting with traditional machine learning and moving on to simple aggregation and then complex cross-modal modelling.

3.4.1 Classical Machine Learning Baselines

Before looking into deep learning methods, classical machine learning algorithms were used as performance baselines. This study used five algorithms which are Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbours (KNN), and XGBoost. These models use statistical feature representations, and PCA and LDA cut down the number of dimensions to 5. Classical ML baselines set a minimum level of performance and make sure that the features that were extracted contain useful information before spending money on complicated deep architectures.

3.4.2 Generation 1: Statistical Fusion (Early, Late, Hierarchical, Hybrid)

The features for this generation are fixed-length statistic features. The Fusion type for this approach is Early Fusion (Feature Level). In this approach, a large vector with a total of 49000 features is generated by concatenating all modality statistic features. The generated large vector will then proceed to a step called SelectKBest prior to a Multi-Layer Perceptron. The model will be able to detect direct correlations in the concatenated features.

Later Fusion/Decision Level) trains the classification model using an MLP per data type. The results are combined via averaging or voting of the output probabilities (logits) to make the final decision. In this method, the modular approach is very useful if some modalities are absent.

Hierarchical Fusion is aware of modality groupings and begins with a fusion of commonly employed modalities such as audio and text and finally adds a visual stream to it as if there could be times with faulty cameras.

Gated Fusion involves the combination of various techniques through the incorporation of the learnable gating network. This allows for the assignment of weights to the modalities before fusion. It enables the model to disregard the valid but noisy modality (e.g. the noisy audio).

3.4.3 Generation 2: Contextual Bi-GRU

The Contextual Bi-GRU generation employs Bidirectional Gated Recurrent Units to represent the temporal aspect. The model examines four methods to fuse the information: Early Fusion (merging the sequences), Late Fusion (taking the average of the logits), Hierarchical Fusion (integrating step by step), or Gated Fusion (learnable modality weights). None of these models converts the time into global summaries; instead, the whole sequence of feature vectors is used. The Self-Attention Pooling component is utilized to assign more importance to the relevant times (for instance, to the loud shout as opposed to the silence).

3.4.4 Generation 3: Contextual SuperTransformer (SDT)

The Contextual SuperTransformer, also referred to as the Self-Distillation Transformer (SDT), represents the state-of-the-art Hybrid Attention-Based approach. This architecture is adapted from prior research that achieved 74.08% weighted F1-score on IEMOCAP, establishing it as a top-performing benchmark in multimodal emotion recognition literature.

3.4.5 Ensemble Strategy

An ensemble approach was used to get the best overall performance. This method combines the predictions of the best individual models. The ensemble uses performance-weighted logit averaging from four contextual Bi-GRU models (Early, Gated, Late) and the SuperTransformer. This strategy uses the strengths of different architectures to get better results than any one model alone. Recurrent models are good at capturing local temporal patterns, while attention-based models are good at capturing global dependencies.

Table 3.1: COMPARISON OF MODEL ARCHITECTURES

Feature	Classical ML	Statistical DL (MLP)	Contextual Bi-GRU	Contextual SuperTransformer
Input Features	Statistical(49k→5 via PCA+LDA)	Statistical (49k→3k via SelectKBest)	Temporal sequences (T×1024)	Temporal sequences (T×1024)
Processing Style	Tree/kernel-based classification	Fixed-vector MLP layers	Sequential (forward+backward)	Parallel attention over full sequence
Temporal Context	None (collapsed)	None (collapsed)	Local (immediate context)	Global (distant dependencies)
Fusion Mechanism	Feature concatenation	Early/Late/Hierarchical/Gated MLP	Self-Attention Pooling + Fusion	Cross-Modal Attention + Gating

3.5 Evaluation Metrics

Weighted Accuracy (%) and Weighted F1-score are two most commonly used classification metrics for objectively measuring and comparing the performance of used fusion models. Collectively, these two criteria provide an overall assessment of model performance, one especially applicable when working through the inherent class imbalance typically present in tasks of emotion recognition.

3.6 Summary

This chapter explains the framework for the methods. This study uses the IEMOCAP dataset (6 classes) and a strong LOSO cross-validation scheme. Feature extraction uses SOTA foundation models (WavLM, DINOv2, ModernBERT) to make both Statistical (for

baselines) and Temporal (for deep learning) representations. There are four types of classification strategies which are Classical ML baselines (RF, SVM, XGBoost, Logistic Regression and KNN), Statistical Deep Learning (MLP fusion), Contextual Bi-GRU, and the advanced SuperTransformer (cross-modal attention). An ensemble method combines the best models to get the most accurate results.

CHAPTER 4

EXPERIMENTAL DESIGN

4.1 Introduction

This chapter outlines the precise experimental protocols and implementation specifics utilized to assess the multimodal fusion architectures delineated in Chapter 3. The previous chapter talked about the general methodological framework, like how to choose datasets, how to extract features, and how to combine them. This chapter, on the other hand, talks about the specific experimental setup, like how to set up models, how to train them, how to choose hyperparameters, how to do cross-validation, and how to test robustness. The goal is to give enough information for others to be able to reproduce the work while also explaining why important design choices were made that set this implementation apart from standard methods in the literature.

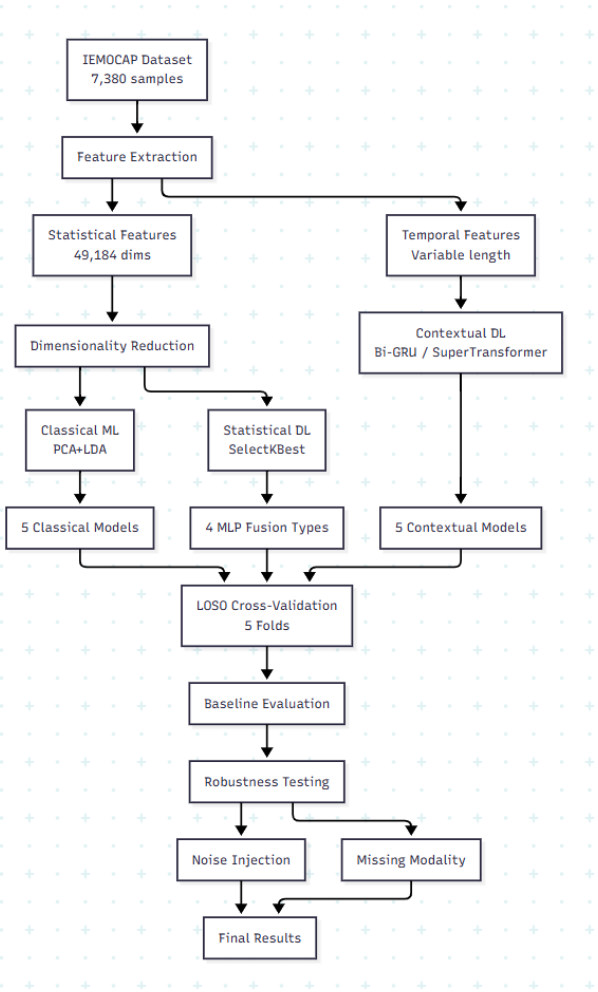


Figure 4.1 : EXPERIMENTAL PIPELINE OVERVIEW

4.1.1 Core Design Decisions and Justification

This study uses six emotions (adding Excited and Frustrated) to test the model with high-arousal disambiguation and capture frustration, which is a common emotion in Human-Computer Interaction. This is different from the simplified Basic 4 approach. For Dataset Selection, IEMOCAP was chosen over in-the-wild datasets because it has high-quality, localized audio-visual recordings that separate modality contributions without too much background noise. The Cross-Validation strategy uses Leave-One-Session-Out (LOSO) to make sure that the evaluation is not biassed by the speaker's unique traits.

Lastly, the modelling hierarchy separates statistical models, which are used as low-latency edge-device baselines, from contextual transformer models, which are meant to be used on the server side with the highest accuracy.

4.2 Feature Extraction Implementation

4.2.1 Model Selection Rational

Three things guided by the choice are proven performance on tasks related to emotions, architectural consistency (768-dimensional outputs), and the ability to do the calculations. The consistent 768-dimensional output across all three modalities enables architectural symmetry in subsequent fusion models.

4.2.2 Statistical Feature Configuration

The statistical feature extraction process provides a rigid representation with dimensions that are beneficial to more traditional machine learning techniques. The parameters set are balanced to extract temporal features without overburdening computations.

Table 4.1: STATISTICAL FEATURE EXTRACTION CONFIGURATION

Modality	Base Model	Layers Extracted	Statistics per Layer	Preprocessing	Output Dims
Audio	WavLM-Base+	10, 11, 12 (last 3)	Mean, Std, Max, Min	16kHzz mono, MAD clipping	9216
Visual	DINOv2-base	12 (last layer)	Mean, Std, Max, Min	20 frames, face detection	3072
Text	ModernBERT-base	10, 11, 12 (last 3)	Mean, Std, Max, Min	128 tokens, cleaned	9216

Extraction of features from transformer layers 10, 11, and 12 with respect to audio and text features aids in preserving key hierarchies. The initial layers tend to extract fundamental features such as sounds and words. The latter layers concentrate more on abstract constructs such as tone and feelings. The research paper utilizes only the final layer 'CLS' to extract visual features since pilot tests proved that there was no significant impact on the results due to extraction from various layers with respect to facial images.

The current study utilized the Median Absolute Deviation (MAD) with a 5-MAD threshold to clip outliers and mitigate the effect of such values due to potential inaccuracies during recording. The MAD clipping is more robust to outliers compared to standard deviation techniques.

4.2.3 Implementation Environment

This study used a Tesla T4 GPU (15.8 GB memory) on Kaggle's cloud platform to extract features. It took about 5 hours for Statistical features and 4 hours for Contextual features to process the whole IEMOCAP dataset (7,380 samples). There were two passes to extract both statistical (.npy) and temporal (.pt) features.

4.3 Dimensionality Reduction Strategy

4.3.1 Classical Machine Learning Pipeline

The 49184-dimensional statistical feature space is hard to work with because of the curse of dimensionality. A two-stage reduction pipeline was created by combining unsupervised PCA with supervised LDA.

Table 4.2: CLASSICAL ML DIMENSIONAL REDUCTION PIPELINE

Stage	Method	Input Dims	Output Dims	Variance Retained	Purpose
1	PCA	49184	1024	~90%	Unsupervised Compression
2	LDA	1024	5	100%	Supervised Projection

PCA cuts down the feature space to 1024 dimensions while keeping 90% of the variance. This study found out that this aggressive compression worked through pilot experiments. By maximizing class separability, LDA is reduced to 5 dimensions (n_classes - 1). It was very important that both transformations were only fit on training data in each cross-validation fold to keep data from leaking.

4.3.2 Deep Learning Pipeline

An ANOVA-based feature selection strategy (SelectKBest) was used instead of PCA for the statistical deep learning models to keep the individual feature contributions easy to

understand. This study chose the top k=1000 features for each modality based on their f-value scores. This gave a dense, highly discriminative 3,000-dimensional input vector.

Table 4.3: DEEP LEARNING FEATURE SELECTION STRATEGY

Modality	**Original Dims**	**Method**	**Selected Dims**	**Selection Efficacy**
Audio	16408	SelectKBest (ANOVA)	1000	High f-value filters noisy features
Visual	16384	SelectKBest (ANOVA)	1000	Focuses on discriminative facial AUs
Text	16392	SelectKBest (ANOVA)	1000	Retains sentiment-rich stat agg.
Total	49184		3000	~94% Reduction

Contextual deep learning models used the raw temporal feature embeddings directly (1024 dimensions per modality, sequence length 20). They used the recurrent architecture (Bi-GRU) to learn relevant temporal patterns without first reducing the number of dimensions.

4.4 Model Implementation Details

This section details the implementation of all model architectures, organized into two paradigms based on their input feature type.

4.4.1 Statistical Features Models

These models operate on fixed-length statistical feature vectors (49184 dimensions), using dimensionality reduction before classification.

4.4.1.1 Classical Machine Learning

Five classical algorithms were used with scikit-learn 1.5.2. All of the models use class weighting to fix IEMOCAP's class imbalance (frustrated: 1,849 samples vs happy: 595 samples).

Table 4.4: CLASSICAL MACHINE LEARNING MODEL CONFIGURATIONS

Model	**Key Hyperparameters**	**Justification**
Random Forest	n_estimators=200 max_depth=20 min_samples_split=5 class_weight='balanced'	Ensemble size provides stability; depth limit prevents overfitting; balanced weighting addresses class imbalance
SVM	kernel='rbf' C=1.0 gamma='scale' class_weight='balanced'	RBF kernel captures non-linear boundaries; standard regularization; automatic scaling
Logistic Regression	C=1.0 max_iter=1000 solver='lbfgs' multi_class='multinomial' class_weight='balanced'	Standard regularization; multinomial for 6-class problem; balanced weighting
KNN	n_neighbors=5 weights='distance' metric='euclidean'	Standard k value; distance weighting reduces influence of far neighbors

XGBoost	n_estimators=200 max_depth=6 learning_rate=0.1 subsample=0.8 colsample_bytree=0.8	Gradient boosting with tuned tree depth and regularization via subsampling

4.4.1.2 Statistical Deep Learning (MLP Fusion)

Four MLP-based fusion architectures were implemented using the 3000-dimensional SelectKBest feature vectors which are Early Fusion (concatenate → MLP), Late Fusion (separate MLPs → average logits), Hierarchical Fusion (Audio+Text → Visual), and Gated Fusion (learned modality weights). These serve as deep learning baselines for comparison with Classical ML.

4.4.2 Contextual Feature Models

These models operate on variable-length temporal sequences (T×1024 per modality), preserving temporal dynamics through recurrent or attention-based processing.

4.4.2.1 Contextual Bi-GRU with Self-Attention

The Bi-GRU architecture treats emotion recognition as a sequence modeling problem. The architecture employs independent Bidirectional GRU encoders for audio, visual, and text streams. To manage temporal dynamics, a learnable self-attention pooling layer collapses the temporal dimension. This study evaluated four fusion strategies which are Early Fusion (concatenation before GRU), Late Fusion (logit averaging), Hierarchical Fusion (Audio+Text first, then Visual), and Gated Fusion (learnable weights per modality).

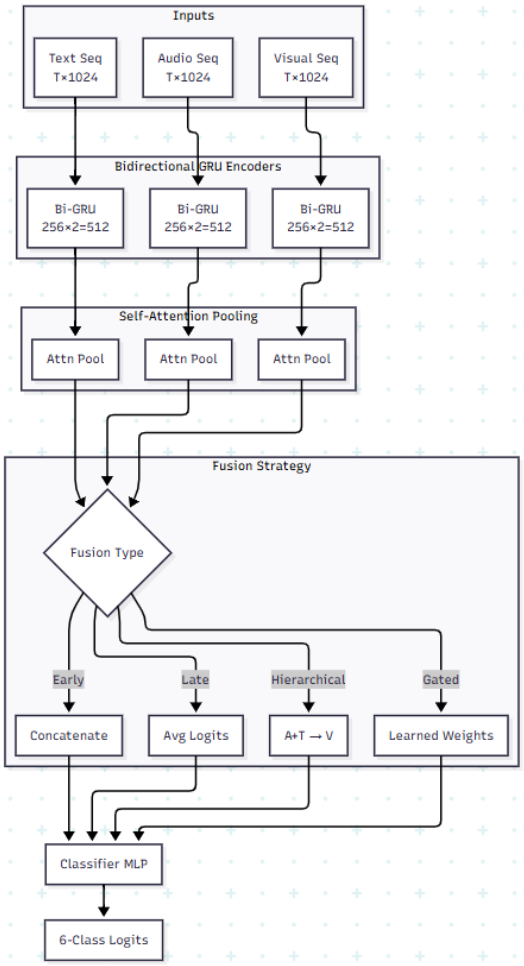


Figure 4.2 : BI-GRU FUSION ARCHITECTURE

4.4.2.2 Contextual SuperTransformer (SDT)

The SuperTransformer architecture, also known as the Self-Distillation Transformer (SDT), implements a cross-modal transformer design inspired by the keys/queries/values mechanism of Ma et. Al, (2023). This architecture is adapted from prior research achieving 74.08% weighted F1 on IEMOCAP, serving as the benchmark for state-of-the-art multimodal emotion recognition.

At its core, the model utilizes cross-modal attention, where the audio stream attends to the text stream (and vice-versa) to align prosody with linguistic content. The

structure consists of 9 encoder blocks (3 Self-Attention, 6 Cross-Attention) processing full uncompressed sequences. Additionally, a dynamic gating mechanism weights the contribution of unimodal versus cross-modal representations based on the input's ambiguity.

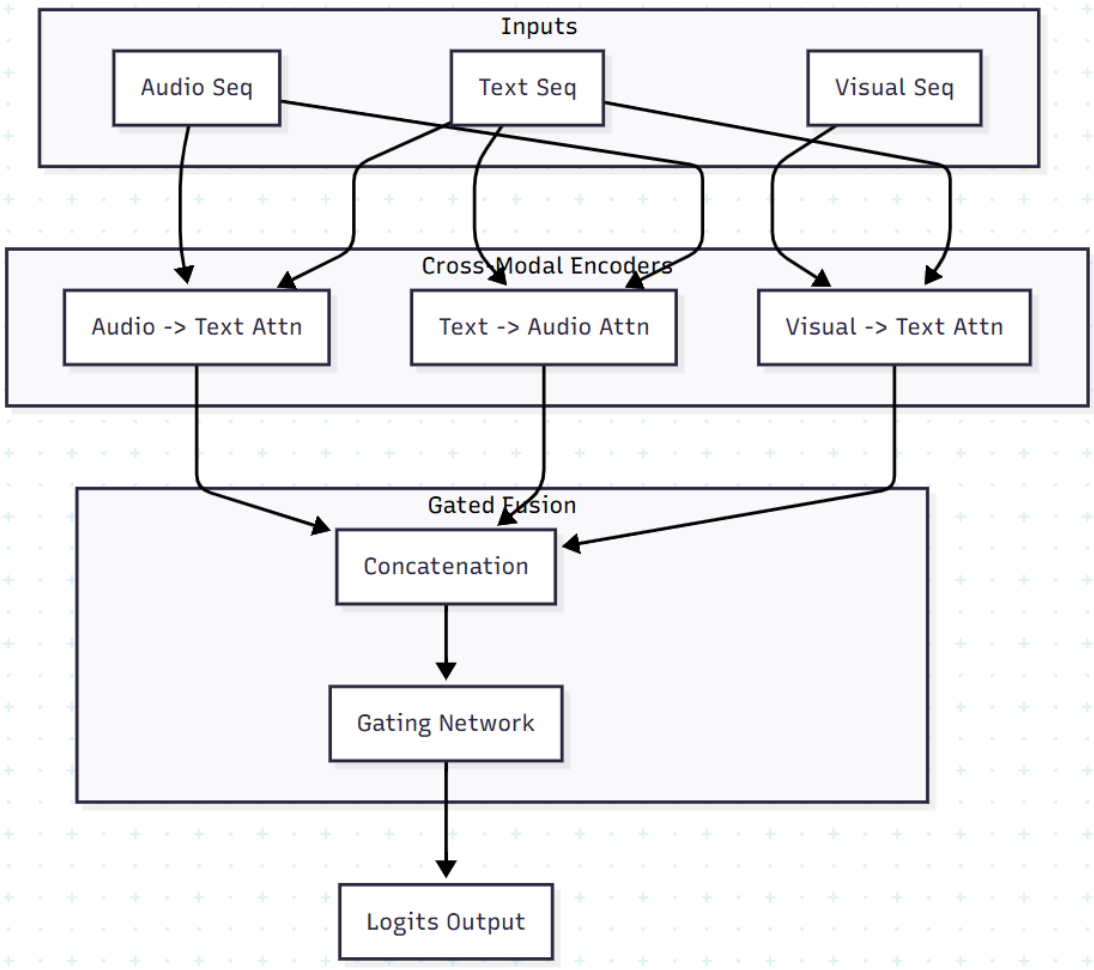


Figure 4.3 : SUPERTRANSFORMER ARCHITECTURE

Table 4.5: DEEP LEARNING ARCHITECTURE SPECIFICATIONS (SUPERTRANSFORMER)

Component	Contextual Bi-GRU	Contextual SuperTransformer (SDT)
Core Encoder	Bidirectional GRU Hidden: 256 (x2=512) Layers: 2	Transformer Encoder 9 Encoders (3 Self + 6 Cross) Heads: configurable, Dim: hidden_dim
Fusion Mechanism	Self-Attention Pooling (Weighted time-average) → MLP	Cross-Modal Attention (A↔T, T↔V, V↔A) + Unimodal/Multimodal Gates
Input Shape	Sequence (Batch, Time, 1024)	Sequence (Batch, Time, 1024)
Regularization	Dropout 0.3, Weight Decay 1e-4	Dropout 0.5, Label Smoothing 0.1
Parameter Count	14M parameters	~45M parameters

4.4.3 Ensemble Configuration

To further improve performance, an ensemble method was added which averages predictions from four leading models. This ensemble averages predictions from Bi-GRU Gated Fusion, Bi-GRU Late Fusion, SuperTransformer, and Random Forest (which performed best among classical machine learning models). This requires no further training, as we simply need to average softmax predictions from each model at evaluation. The logic behind this ensemble method is to leverage the respective strengths of each model: Bi-GRUs have strong ability to identify local temporal information, SuperTransformer has strong ability to identify global connections based on attention mechanisms, and Random Forest provides robustness.

4.5 Training Procedure

4.5.1 Hyperparameter Optimization

Systematic hyperparameter tuning was conducted using Optuna for the Statistical DL (MLP) and Contextual Bi-GRU architectures. Classical ML models used fixed hyperparameters (Table 4.4), and the SuperTransformer used manually tuned parameters based on prior research.

Table 4.6: HYPERPARAMETER CONFIGURATION BY MODEL TYPE

A. Classical ML (Fixed Hyperparameters)

Classical ML models used fixed, empirically validated hyperparameters without automated tuning. See Table 4.4 for full configurations.

B. Statistical DL - MLP Fusion (Optuna Tuning)

Parameter	**Search Space**	**Optimal Value**
Learning Rate	[1e-5, 1e-3] log	1e-4
Hidden Dim	{256, 384, 512}	384
Dropout	[0.2, 0.5]	0.3
Num Layers	{1, 2, 3}	2
Trials per fold		10

C. Contextual Bi-GRU (Optuna Tuning)

Parameter	**Search Space**	**Optimal Value**
Learning Rate	[1e-5, 1e-3] log	5e-4
Hidden Dim	{256, 384, 512}	256
Dropout	[0.2, 0.5]	0.3
GRU Layers	{1, 2}	2
Trials per fold		10

D. Contextual SuperTransformer (Manual Tuning)

SuperTransformer parameters were adapted from prior research (SDT achieving 74.08% F1) with minor adjustments.

Parameter	**Value**	**Source**
Learning Rate	5e-4	Prior research
Hidden Dim	512	Prior research
Dropout	0.5	Prior research
Attention Heads	8	Prior research
Encoder Layers	1 per block × 9 blocks	Adapted

Selection Criterion: Highest validation weighted F1-score

For architectures using Optuna, the optimization process maximized validation F1-score within a constrained budget (10-20 trials per fold). Optimal configurations typically converged around learning rate 5e-4 to 1e-4, hidden dimension 256-512, and dropout 0.3-0.5. These per-fold optimized parameters were utilized for the final training runs.

4.5.2 Training Configuration

The training procedure addresses common challenges in emotion recognition through careful optimizer and loss function selection.

Table 4.7: TRAINING CONFIGURATION

Component	**Statistical DL (MLP)**	**Contextual Bi-GRU**	**Contextual SuperTransformer**

Optimizer	AdamW, weight_decay=0.01	AdamW, weight_decay=0.01	Adam, weight_decay=1e-5
LearningRate	Optuna-tuned (1e-5 to 1e-3)	Optuna-tuned (1e-5 to 1e-3)	1e-4 (fixed)
Scheduler		Warmup 5 epochs + Cosine	Warmup 5 epochs + Cosine
Batch Size	64	16	16
Gradient Accumulation			2 steps (effective batch 32)
Loss Function	CrossEntropyLoss	CrossEntropyLoss + Label Smoothing (0.1)	CrossEntropyLoss + Label Smoothing (0.1)
Class Balancing	WeightedRandomSampler	WeightedRandomSampler	WeightedRandomSampler
Modality Dropout		30% per modality	
Regularization	Dropout 0.3, weight_decay	Dropout 0.3-0.5, weight_decay	Dropout 0.5, weight_decay
Early Stopping	Patience 15, Max 50 epochs	Patience 15, Max 50 epochs	Patience 15, Max 50 epochs

This study trained all of the deep learning models on a Tesla T4 GPU with a batch size of 16 and CrossEntropyLoss. To deal with the big class imbalance between happy and frustrated samples, this study used weighted random sampling instead of loss weighting. The SuperTransformer used the regular Adam, but the statistical DL and contextual Bi-GRU models used AdamW to take advantage of decoupled weight decay for better regularization. The contextual models also used a 5 epoches linear warmup and early stopping with a patience of 15 epochs to make sure they were stable. To make the models more robust to missing modalities, all of the contextual Bi-GRU fusion models (Early, Late, Hierarchical, and Gated) use modality dropout (30% per modality, see Table

4.7) during training. This regularization method randomly zeros out whole modality streams, which makes the model learn balanced representations that do not depend too much on any one modality.

4.6 Evaluation Methodology

Leave-One-Session-Out (LOSO) cross-validation with hybrid splits was implemented to ensure speaker-independent evaluation. The IEMOCAP dataset's five sessions each contain unique speaker pairs, making session-level splitting a natural choice for testing generalization to unseen speakers.

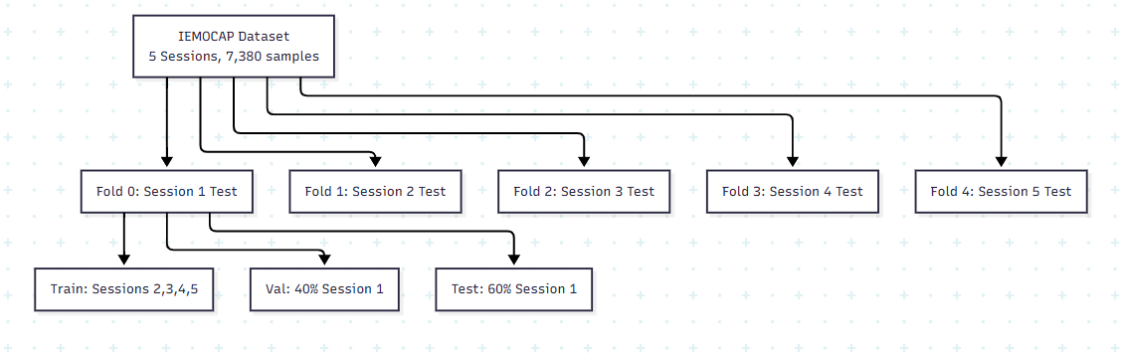


Figure 4.4 : LOSO CROSS-VALIDATION WITH HYBRID SPLITS

The hybrid split strategy creates a validation set from a different session than the test set, maintaining speaker independence while enabling principled early stopping. For each fold, the test session is divided 60/40 into validation (40%) and test (60%) based on speaker identity. The remaining four complete sessions plus the 60% training portion comprise the training set (~5,400-5,500 samples per fold). This ensures: (1) test speakers are completely unseen during training and validation, (2) validation speakers differ from test speakers, (3) sufficient training data remains, and (4) consistent fold sizes across all five iterations.

4.7 Evaluation Metrics

Two metrics have been used to evaluate model performance because they involve different aspects of classification performance when there is an imbalance in classes. The primary metric is the Weighted-F1. This is the standard F1 measure weighted by classes, giving more weight to classes with more test samples. This was selected primarily because it presents a balance between precision and recall, as well as taking into account class distribution. Its ability to measure performance on frequently encountered emotions is rather important, as opposed to their performance on less encountered emotions. This was computed per fold using scikit-learn's f1_score(average='weighted'), then averaged over five folds to gauge stability.

4.8 Robustness Testing Protocols

In all robustness tests, models are trained only on clean data. However, the input to the test varies. This is to prevent the test from leveraging any robustness on the degraded training data.

4.8.1 Noise Injection Testing

Noise robustness tests: These tests measure how resilient a model will be when some noise is introduced in features, possibly because of a faulty recording or transmitter.

Table 4.8: NOISE INJECTION TEST CONFIGURATION

Parameter	Configuration	Rationale
Noise Type	Additive White Gaussian Noise (AWGN)	Standard for simulating random corruption

Application Point	Extracted features (normalized space)	Tests model robustness, not feature extractor
Noise Levels (σ)	[0.0, 0.1, 0.2, 0.3, 0.5, 1.0]	Progressive degradation from clean to severe
Per-Modality	Applied independently	Isolates single-modality vulnerabilities

4.8.2 Missing Modality Testing

Missing Modality Testing: This tests the effect if one or more of the input modalities were down or unavailable to the system.

Table 4.9: MISSING MODALITIES TEST SCENARIOS

Scenario	**Available Modalities**	**Zeroed Modalities**	**Purpose**
Baseline	Audio + Visual + Text	None	Reference performance
No Audio	Visual + Text	Audio	Test without acoustic cues
No Visual	Audio + Text	Visual	Test without visual cues
No Text	Audio + Visual	Text	Test without linguistic content
Audio Only	Audio	Visual + Text	Test audio-only classification
Visual Only	Visual	Audio + Text	Test visual-only classification
Text Only	Text	Audio + Visual	Test text-only classification

Missing modalities were simulated by replacing their feature vectors with zeros at the appropriate input stage for each architecture. Late Fusion naturally handles missing

modalities through its modular design the outputs of unavailable modality classifiers were excluded from the ensemble average. For Early Fusion, Hierarchical Fusion, and Attention Hybrid architectures, zero vectors were substituted for missing modalities during the forward pass. Each test scenario was evaluated on all test samples across the five LOSO folds, with performance metrics compared to the baseline to quantify each architecture's dependence on specific modalities.

4.9 Summary

This chapter explained the experimental steps that put the methods from Chapter 3 into action. The key contributions of this experimental design are the well-thought-out choices of WavLM-Base+, DINOv2-base, and ModernBERT-base because they have been shown to work well and be efficient. We created two types of feature representations: Statistical and Temporal. These work with both classical and deep learning methods. We also made custom dimensionality reduction pipelines, such as PCA+LDA for ML and SelectKBest for DL. The implementation includes five classical algorithms (with fixed hyperparameters), four deep learning fusion architectures (optimized with Optuna), and an Ensemble strategy. LOSO cross-validation with hybrid splits for speaker independence and principled robustness testing across noise injection and missing modality scenarios make sure that the evaluation is thorough.

CHAPTER 5

IMPLEMENTATION

5.1 Introduction

This chapter explains how to use the multimodal emotion recognition system in real life. The codebase has been reorganized into a modular structure so that it can be reused and expanded. The workflow went from scripts that prepared data on a local machine to Kaggle Notebooks that used GPUs to speed up feature extraction and model training. This study used PyTorch 2.6.0 and Tesla T4 GPUs (Kaggle) to put it into action. The project is set up like this: scripts/local/ has tools for filtering and preparing data, scripts/analysis/ has diagnostic and evaluation scripts (The Gauntlet), and notebooks/ has comparisons of Classical ML, Statistical DL, and Contextual DL.

5.2 Data Exploration and Understanding

5.2.1 Data Preparation Scripts

Five specialized scripts in scripts/local/ were used to automate data preparation. They were run one after the other:

Table 5.1: LOCAL DATA PREPARATION PIPELINE

Script	Purpose	Output
local_action_01_create_manifest.py	Scans extracted features to create a master CSV manifest with utterance metadata	master_manifest.csv
local_action_02_identify_files.py	Scans raw IEMOCAP, applies 6-emotion mapping, identifies required files	required_files.json
local_action_03_copy_files.py	Copies wav/txt files to data/filtered/ (skips AVIs for clipping)	Filtered utterance files
local_action_03b_clip_videos.py	Clips full-dialogue AVI files to per-utterance segments using ffmpeg	Per-utterance AVI clips
local_action_04_zip_for_kaggle.py	Compresses filtered dataset for Kaggle upload	iemocap_filtered_6emotions.zip

5.2.2 Dataset Parsing and Metadata Generation

The first step in the implementation was to parse the raw IEMOCAP dataset to get metadata at the utterance level. The parsing script took annotation files from all five sessions and filtered them for the six target emotions (angry, excited, frustrated, happy, neutral, and sad). It then made a structured CSV file with utterance identifiers, emotion labels, session information, and file paths for extracted features across audio, visual, and text modalities.

```
================================================================================
CREATING MASTER MANIFEST FROM EXISTING FEATURES
================================================================================
Features directory: extracted_features/features_statistical/
Output: extracted_features/master_manifest.csv

Found 7380 audio feature files

Processing utterances...
  [                    ]  20.0% - Session1: 1481 utterances
  [                    ]  40.0% - Session2: 1473 utterances
  [                    ]  60.0% - Session3: 1477 utterances
  [                    ]  80.0% - Session4: 1472 utterances
  [                    ] 100.0% - Session5: 1477 utterances

✓ Manifest created with 7380 utterances
✓ Saved to: extracted_features/master_manifest.csv

Emotion distribution:
  angry       : 1103 ( 14.9%)
  excited     : 1041 ( 14.1%)
  frustrated  : 1849 ( 25.1%)
  happy       :  595 (  8.1%)
  neutral     : 1708 ( 23.1%)
  sad         : 1084 ( 14.7%)

Session distribution:
  Session1: 1481
  Session2: 1473
  Session3: 1477
  Session4: 1472
  Session5: 1477

================================================================================
METADATA PARSING COMPLETE
================================================================================
```

```
Summary:
  • Raw IEMOCAP scanned: 10,039 total utterances (11 emotion categories)
  • Filtered dataset: 7,380 utterances (6 target emotions)
  • Exclusions: 2,659 utterances (26.5%)
    - 'xxx' (unknown): 2,507
    - 'surprise': 107
    - 'fear': 40
    - 'other': 3
    - 'disgust': 2

  • File completeness: 100%
    - Audio files (.wav): 7,380/7,380 ✓
    - Video files (.avi): 7,380/7,380 ✓
    - Text files (.txt): 7,380/7,380 ✓

  • Generated: master_manifest.csv
    Columns: utterance_id, session, emotion, emotion_label,
             audio_path, visual_path, text_path
```

Figure 5.1 : EXECUTION OF METADATA PARSING SCRIPT

The manifest file that comes out of this process, called `master_manifest.csv`, has 7,380 utterances and covers all three modes of communication (audio, visual, and text) completely. It serves as the basis for all the steps that come after it. The metadata file provides the starting point for all that follows with 7380 utterances covered fully in audio, video, and text formats.

5.2.3 Data Filtering and Emotion Mapping

There were 10039 utterances in the raw IEMOCAP dataset, which were divided into 11 emotion categories. To maintain data quality and interpretability, a filtering process preserved only the six fundamental emotions (angry, excited, frustrated, happy, neutral, sad), eliminating 2659 utterances (26.5%) predominantly ambiguous "xxx" labels (2,507) and infrequent emotions such as "surprise" (107), "fear" (40), "disgust" (2), and "other" (3).

Table 5.2: RAW IEMOCAP DATASET

Characteristics	**Value**
Dataset Used	IEMOCAP (Interactive Emotional Dyadic Motion Capture)
Total Utterances (Original)	10039
Selected Emotion Classes	6 (angry, excited, frustrated, happy, neutral, sad)
Total Utterances (After Filtering)	7,380
Number of Sessions	5
Number of Speakers	10

Figure 5.2 shows how filtering changes the composition of a dataset by comparing the original and filtered class distributions. The filtered dataset of 7380

utterances was checked for quality, and it was found that all three formats (audio WAV, video AVI, and text transcripts) were 100% complete. This study kept a few outliers: 18 utterances that were longer than 20 seconds (0.24%) and 502 utterances that were shorter than 2 words (6.8%). These are valid natural emotional expressions.

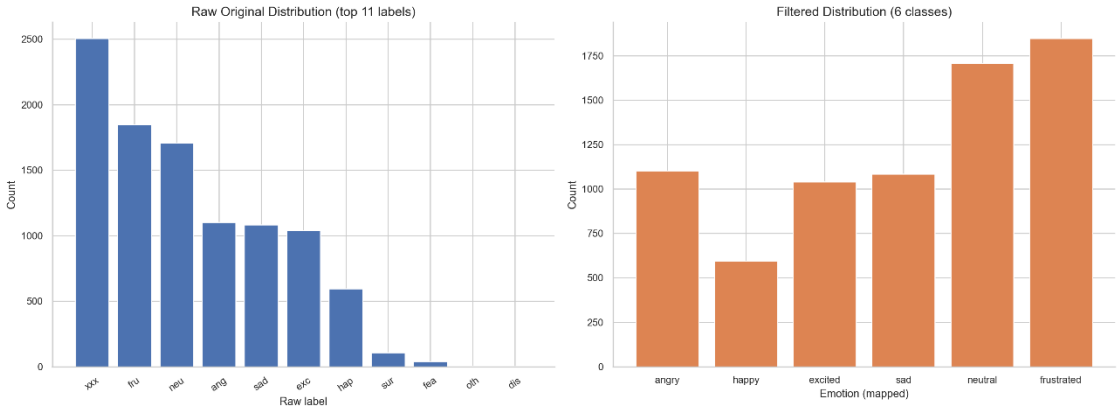


Figure 5.2 : CLASS DISTRIBUTION BEFORE AND AFTER FILTERING

5.2.4 Session Distribution and Cross-Validation Structure

There are 7380 utterances across 5 IEMOCAP sessions, with each session featuring conversations between different pairs of speakers. Figure 5.3 displays the number of utterances in each session. Each session had a similar number of utterances, ranging from 1348 to 1622. This balanced setup supports the Leave-One-Session-Out (LOSO) cross-validation method, ensuring each fold has sufficient training data (around 5,900 utterances) and test data (about 1,400 utterances) for effective evaluation

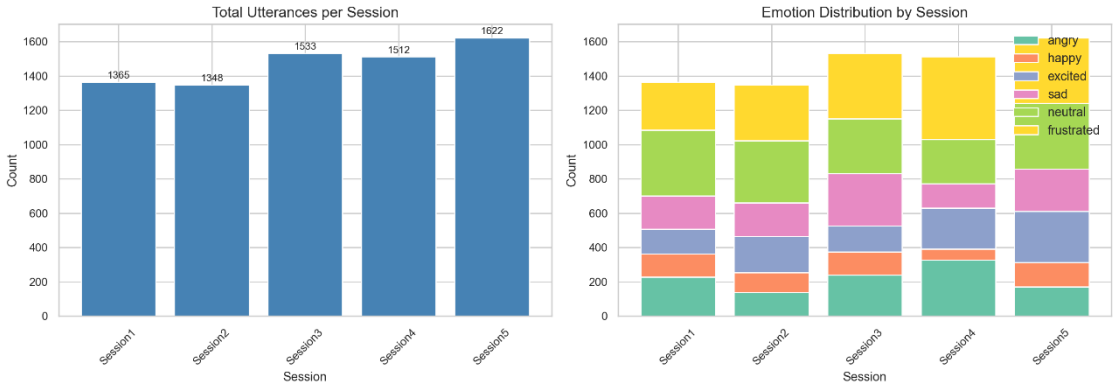


Figure 5.3 : SESSION DISTRIBUTION

The balanced session distribution ensures that evaluating model performance is fair, as each test fold makes up about 19% of the total dataset and includes different speaker identities. This helps in accurately assessing generalization.

5.2.5 Utterance Duration and Length Analysis

Utterance characteristics analysis shows that most affective expressions in IEMOCAP are brief. This is shown by Figures 5.4 and 5.5 below audio and video duration are right-skewed with means 4.59 seconds and 3.64 seconds respectively. Most utterance last between 2 and 7 seconds. Some take 34 seconds.

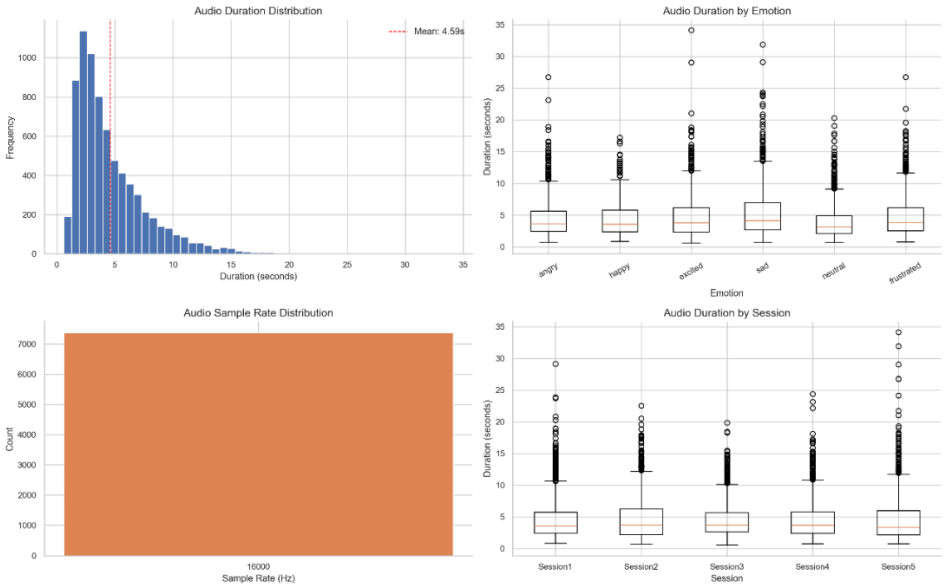


Figure 5.4 : AUDIO DURATION DISTRIBUTION

From figure 5.5, A two-panel display of the characteristics of a video. In the left pane is shown the distribution of the FPS (frames per second). In all 7,380 video recordings, the data peaks at 29.97 FPS. This shows that there is no sampling difference. The right pane contains the distribution of the video dimensions. The majority of these recordings use 720 by 480 pixels (standard definition with 3:2 aspect ratio), which includes almost all statements. The technical parameters of the recording aren't different among the sessions to ensure that there will be no effect on obtaining visual features.

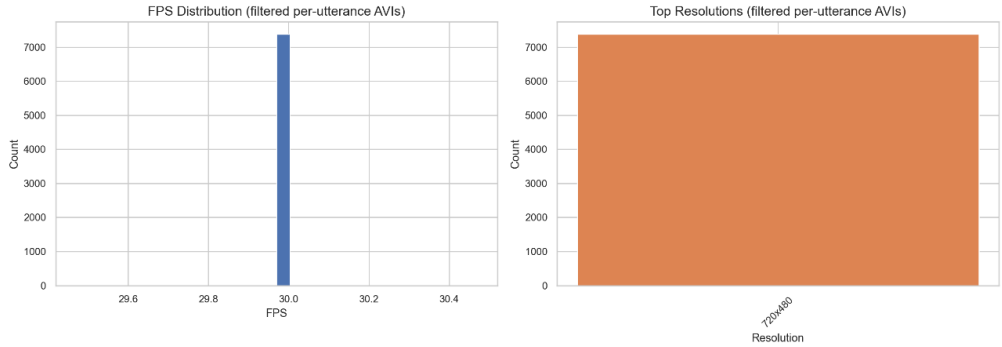


Figure 5.5 : VIDEO DURATION DISTRIBUTION

The lengths of text utterances in Figure 5.6 are highly variable, as one would expect in conversations. The average length of spoken phrases is 12 words, but the median is 9 words, suggesting that the distribution is skewed to the right. The difference between responses can range from one word to more elaborate answers that are no more than 100 words, with an average difference of 10.5 words. The variability here illustrates that the choice of words and phrase syntax ranging from 'oh' to more complex two-word expressions encode emotional meaning.

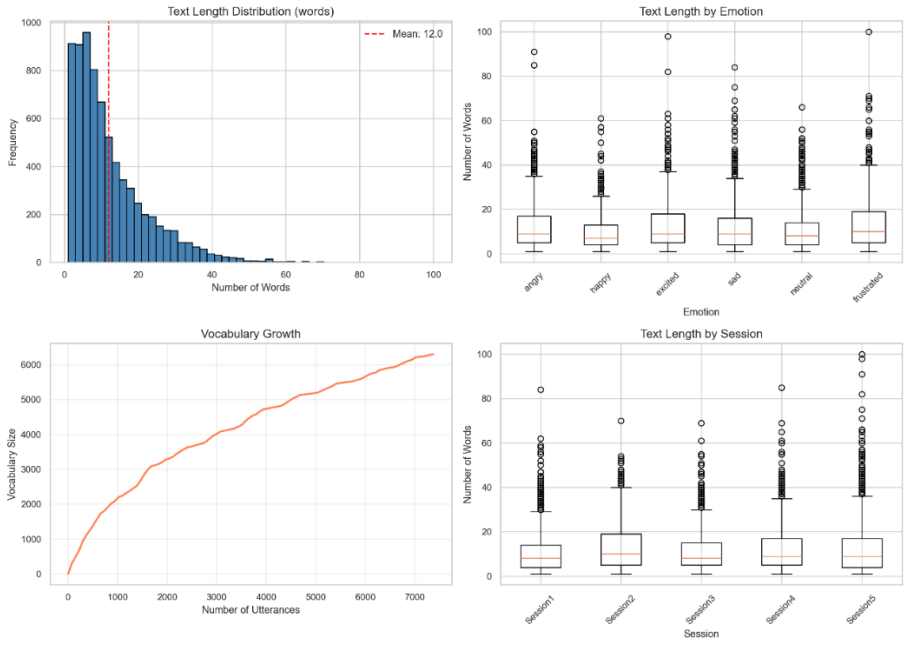


Figure 5.6 : TEXT UTTERANCE LENGTH DISTRIBUTION

The brief statements affirmed that the choice to apply simple statistical values (mean, standard deviation, Max, Min) to time series instead of more complex LSTM networks or Transformer models was made due to the lack of enough data to model.

5.2.6 Emotional Space Characterization (VAD Analysis)

In understanding the differing types of emotional experiences based on psychological dimensions, VAD scores derived from IEMOCAP annotations are presented. Figure 5.7 above highlights the spread of VAD dimensions among six categories corresponding to various types of emotions that form the theoretical basis underlying multimodal analysis.

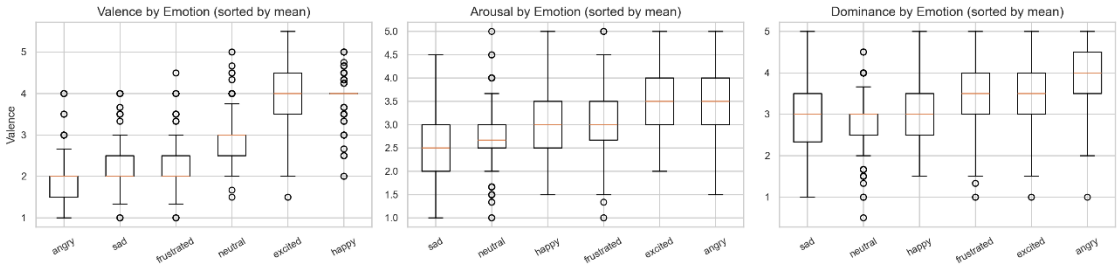


Figure 5.7 : VAD (VALENCE-AROUSAL-DOMINANCE) BY EMOTION

A clear difference exists between positive feelings (happy = 3.95, excited = 3.95) and negative feelings (angry = 1.91, sad = 2.25), indicating that valence is an effective means to distinguish. High-aroused feelings (angry = 3.64, excited = 3.58) vs. low-aroused feelings (sad = 2.56, neutral = 2.73) are useful to distinguish between happy and sad feelings. Angry has the strongest dominance (3.95), and sad has the lowest (2.83). It illustrates the assertiveness feature that distinguishes angry from frustrated since these two are negative-valence feelings with differing degrees of dominance. The overlap of frustrated and angry in VAD-space (negative-valence, high-arousal, moderate to high dominance) correlates with the results seen in prediction distributions since these are similar feelings.

These VAD patterns make it clear that feelings are located in particular regions across a three-dimensional psychological space; therefore, these findings provide a theoretical explanation for the importance of multimodal features (prosody, facial expressions, and linguistic content) that facilitate effective discrimination.

5.2.7 Speaker Demographics and Gender Balance

To evaluate potential demographic biases, the distribution of speaker gender was examined throughout the dataset. IEMOCAP has recordings of 10 actors (5 men and 5 women) from 5 sessions. Each session has one male-female pair. Figure 5.8 shows how the genders are spread out overall and how they are spread out by emotion.

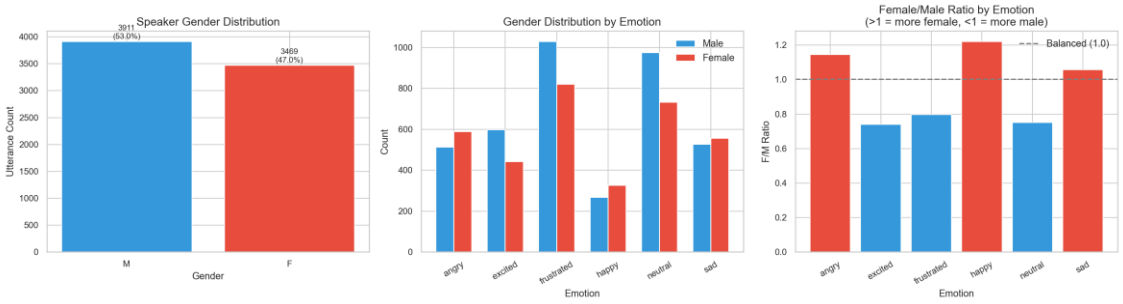


Figure 5.8 : SPEAKER DEMOGRAPHICS ANALYSIS

5.2.8 Emotion Duration Comparison

One important outcome is the comparison between the length of the audio depending on the emotion. In Figure 5.9, it is clear that the length of the speech varies depending on the emotion.

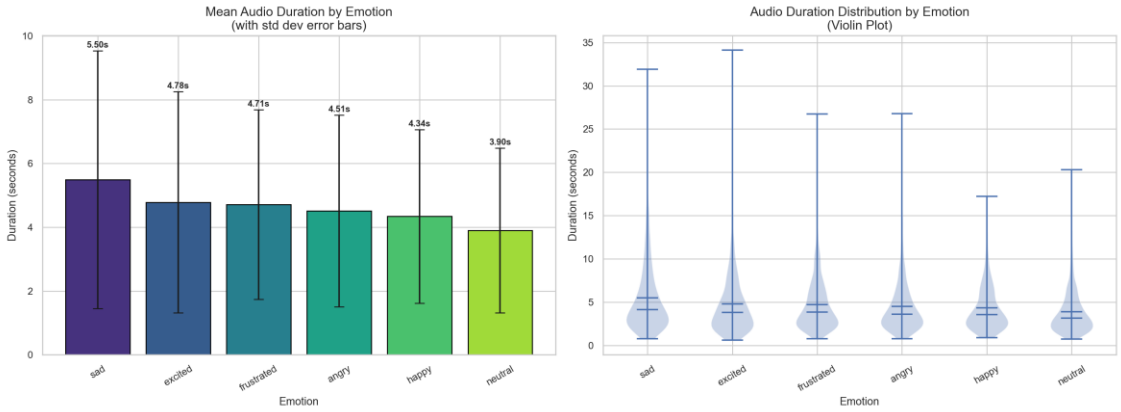


Figure 5.9 : EMOTION DURATION DISTRIBUTION

In summary, sad speech lasts approximately 41% longer than neutral speech (5.50s versus 3.90s). The variance in duration may serve as an important acoustic cue. As a general rule, longer speech phrases correlate with emotions expressed through more complex speech. This underlines the importance of time-related information modeled in contextual networks (Bi-GRU, SuperTransformer) that discover duration-related patterns overlooked by statistical pooling.

This analysis also shows a balance in terms of gender in the dataset. No gender shows a particular bias towards any emotion. This helps to ensure models perform well for all speakers and not only for a certain gender to act as a shortcut for an emotion.

5.3 Feature Extraction and Validation

5.3.1 Multimodal Feature Extraction Execution

For feature extraction, it utilized the pre-trained foundation models WavLM-Base+, DINOv2-base for audio and visual features, and ModernBERT-base for text features. To extract all 7,380 utterances, it took 5 hours for statistical features and 4 hours for

contextual features on Kaggle's Tesla T4 GPU. The results were stored in the form of NumPy files containing manifests.

5.3.2 Statistical Feature Validation

The statistics were reviewed to ensure the data was properly extracted. The data analysis on the distributions indicated that the attributes for all the modalities were distinct:

Table 5.3: STATISTICAL FEATURE DISTRIBUTION SUMMARY

Modality	**Mean (μ)**	**Std Dev (σ)**	**Notes**
Audio	0.71	27.5	Wide variance from WavLM spectral diversity
Visual	0.56	1.11	Near-normal distribution from DINOv2
Text	15.7	268.5	High variance from ModernBERT semantic breadth

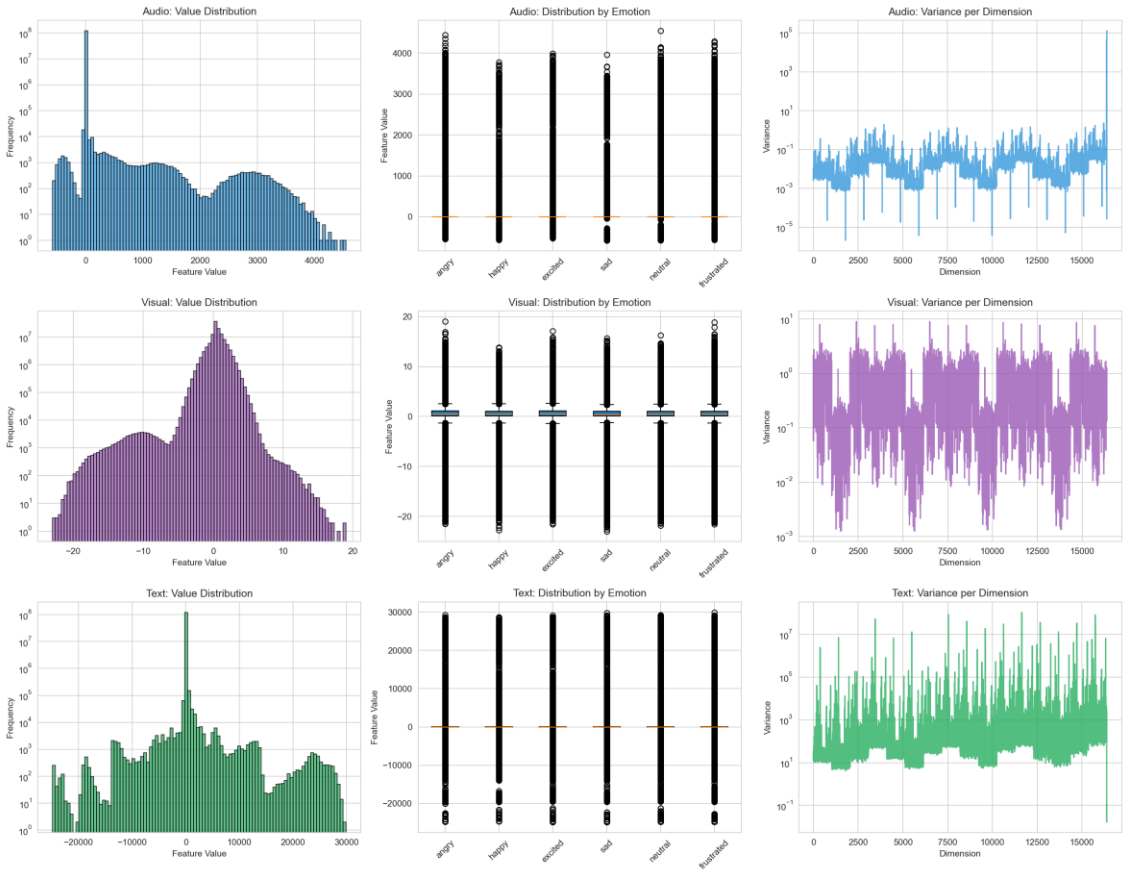


Figure 5.10 : STATISTICAL FEATURE DISTRIBUTIONS

Cross-modality correlation analysis demonstrated low to moderate correlations (Audio-Visual: 0.048, Audio-Text: 0.182, Visual-Text: 0.033), suggesting that modalities capture predominantly independent information strongly supporting the multimodal fusion approach as each modality offers a distinct discriminative signal with minimal redundancy.

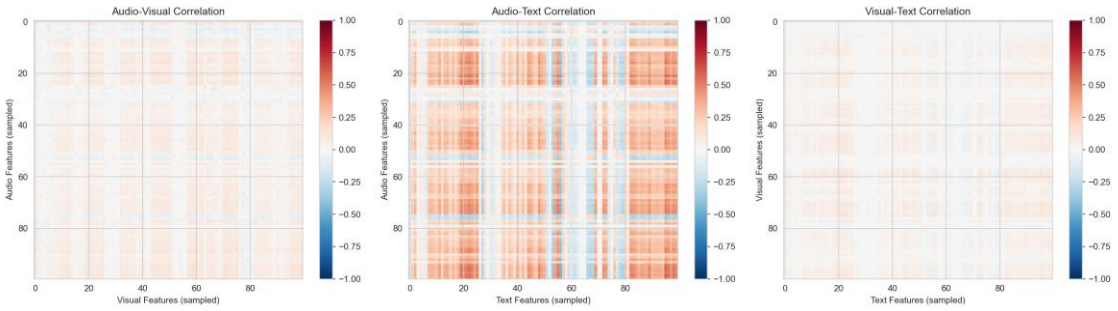


Figure 5.11 : STATISTICAL CROSS-MODALITY CORRELATION

5.3.3 Contextual Feature Validation

The contextual (temporal) features were also checked to make sure they were of good quality:

Table 5.4: CONTEXTUAL FEATURE DISTRIBUTION SUMMARY

Modality	Mean (μ)	Std Dev (σ)	Notes
Audio	-0.0003	0.082	Near-zero centered, low variance
Visual	0.02	1.21	Similar to statistical visual features
Text	0.009	0.72	Tighter distribution than statistical

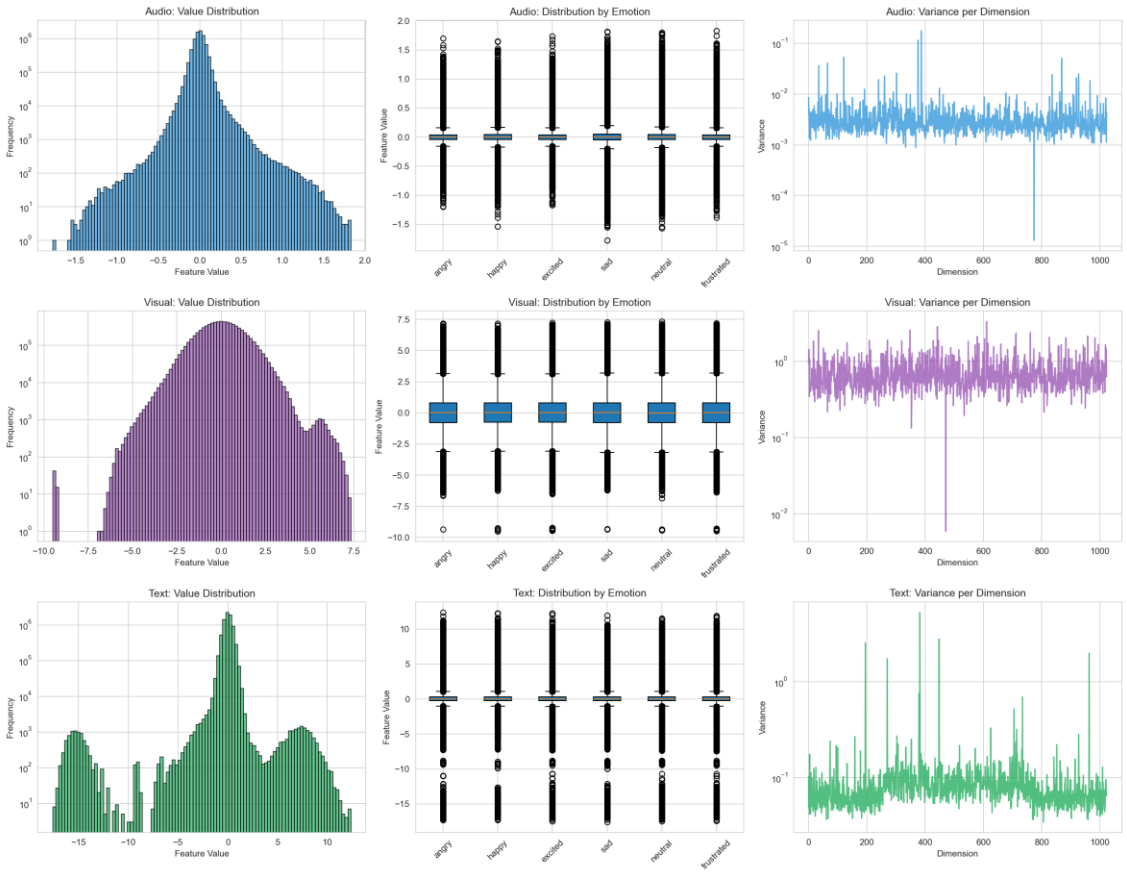


Figure 5.12 : CONTEXTUAL FEATURE DISTRIBUTIONS

Cross-modality correlations for contextual features were also low (Audio-Visual: 0.045, Audio-Text: 0.095, Visual-Text: 0.018), which also shows that modality independence is true.

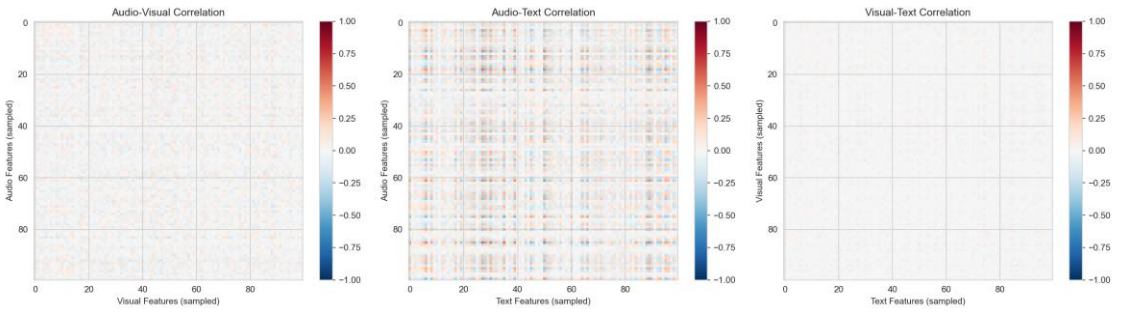


Figure 5.13 : CONTEXTUAL CROSS-MODALITY CORRELATION

5.3.4 STATISTICAL VS CONTEXTUAL FEATURE COMPARISON

With both feature types validated, a comparative analysis was performed using unsupervised clustering metrics. Detailed quantitative results are presented in Chapter 6, Table 6.1 (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score). The pictures below show how different classes can be separated:

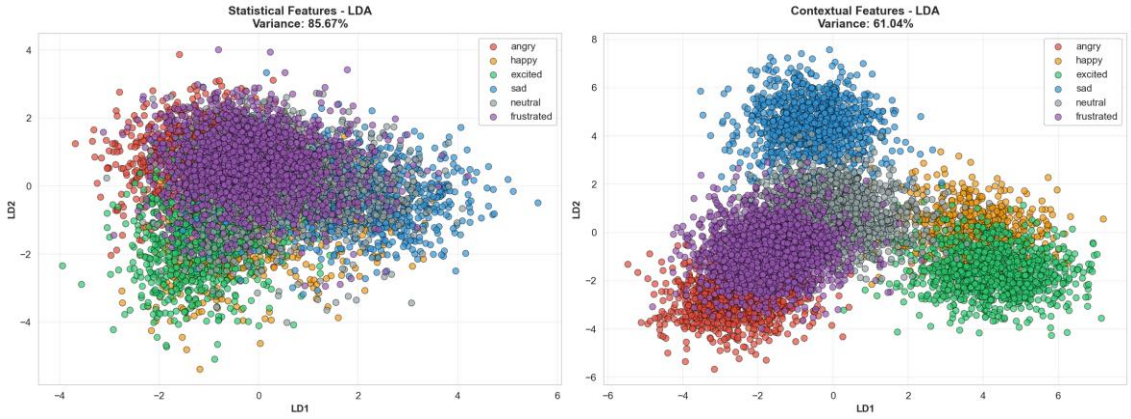


Figure 5.14 : LDA SEPARABILITY COMPARISON

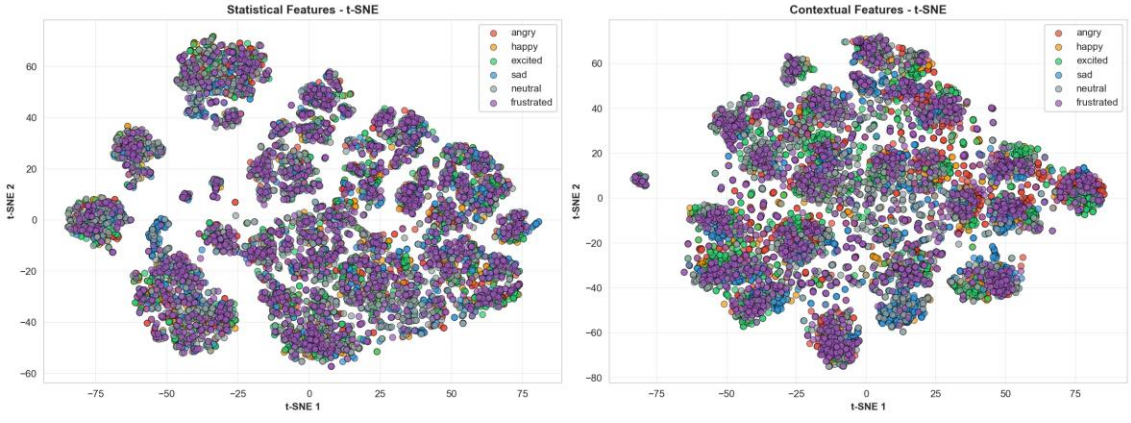


Figure 5.15 : t-SNE VISUALIZATION COMPARISON

Even though contextual features have 16 times fewer dimensions, they get 2.1 times higher Calinski-Harabasz scores. This improvement comes from, firstly is the temporal preservation. Contextual features keep sequential patterns that set apart emotions with similar static properties. Secondly is the learned pooling. Self-attention learns to give more weight to frames that are useful, while fixed statistics treat all frames the same.

This analysis empirically validates the project's dual-architecture strategy where statistical models function as effective baselines, whereas contextual models utilize enhanced temporal representations for optimal performance.

5.4 Dimensionality Reduction and Baseline Models

5.4.1 PCA VISUALIZATION AND VARIANCE ANALYSIS

Principal Component Analysis (PCA) was used to show how well the feature space could be separated because it has a lot of dimensions (49,184). Figure 5.16 shows a 2D PCA projection that is colored by emotion. It shows moderate clustering, with sad forming a fairly distinct cluster and frustrated and neutral overlapping a lot.

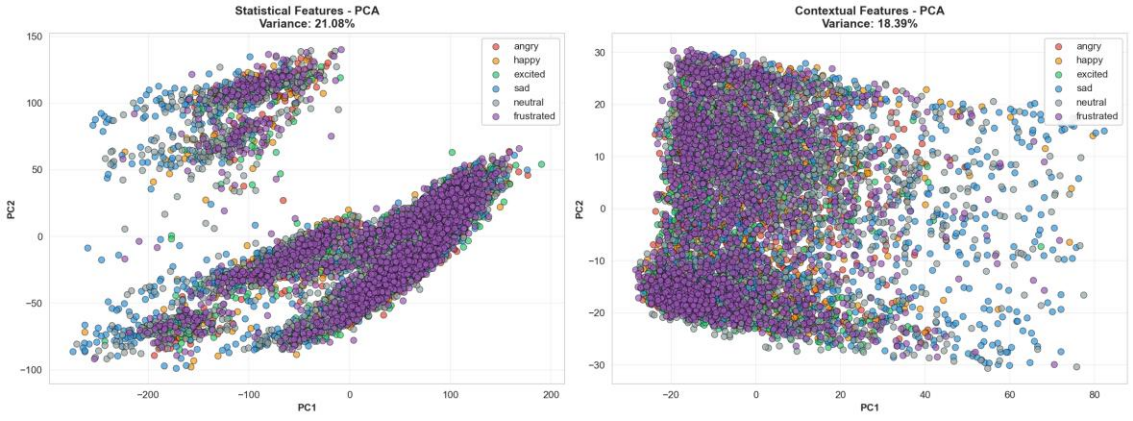


Figure 5.16 : PCA 2D PROJECTION OF FEATURES

Cumulative explained variance was used to find the right dimensionality for reduction. Figure 5.16 shows that the dimensionality reduction strategy keeps 95% of the variance with about 580 components.

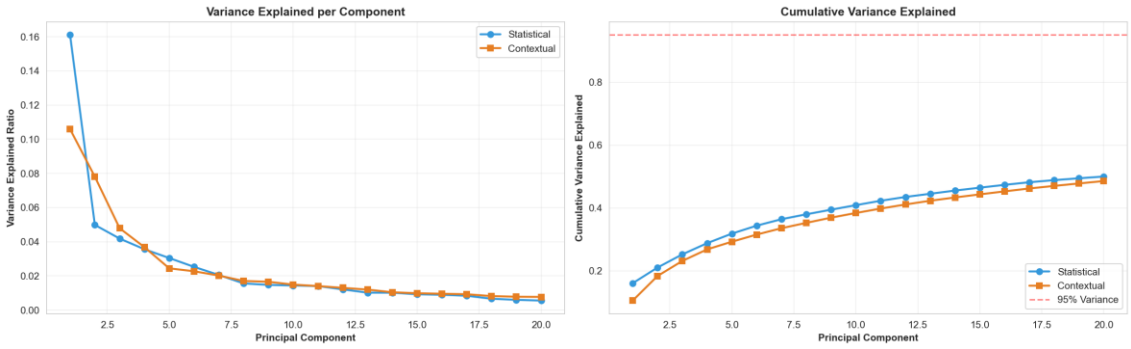


Figure 5.17 : PCA CUMULATIVE EXPLAINED VARIANCE

After looking at this, Classical ML models used PCA reduction to 1024 dimensions and then LDA. For Deep Learning, though, this study chose a feature selection strategy (SelectKBest) instead of PCA to keep the meaning of each feature clear. This study chose the top 1,000 most discriminative features for each modality. This gave us a 3,000-dimensional input vector for statistical fusion models.

5.4.1.1 Pairwise Class Separability Analysis

Pairwise centroid distances were calculated in feature space to foresee which emotion pairs would pose the greatest challenge for classifiers to differentiate. Figure 5.18 shows a heatmap and a list of emotion pair separability's in order.

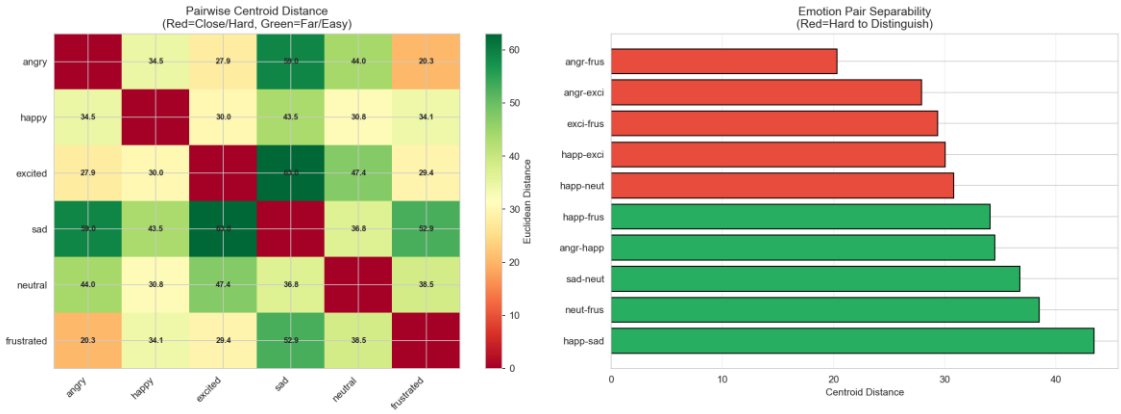


Figure 5.18 : PAIRWISE CLASS SEPARABILITY

Table 5.5: HARDEST AND EASIEST EMOTION PAIRS TO DISTINGUISH

Rank	Emotion Pair	Centroid Distance	Difficulty
1	Angry ↔ Frustrated	20.3	**HARD**
2	Angry ↔ Excited	27.9	HARD
3	Excited ↔ Frustrated	29.4	HARD
...	...	...	...
14	Angry ↔ Sad	59.0	Easy
15	Excited ↔ Sad	63.0	Easy

The very close distance between the Angry and Frustrated centroids (20.3) explains why these emotions keep getting mixed up in all the models we tested. Both emotions have similar acoustic properties (high energy, raised voice) and valence (negative), but they differ mainly in dominance, which is not well captured by MLP-based statistical models. This discovery inspires attention-based architectures (SuperTransformer) capable of learning to focus on dominance-specific signals.

5.4.2 Classical ML Baseline Implementation

This study trained classical machine learning models as baselines and used PCA+LDA to reduce the number of dimensions from 49184 to 5. This strong dimensionality reduction was necessary for classical algorithms, but it made it harder to tell the difference between things. This is why the SelectKBest feature selection method was created for deep learning models. Chapter 6, Table 6.2 shows all of the performance results.

5.5 Deep Learning Model Training

5.5.1 Hyperparameter Tuning Process

Optuna was used to tune the hyperparameters on the Fold 1 validation set before full-scale training. The search found the best settings for each architecture after 10 trials per fold (for example, Learning Rates ~1e-4, Hidden Dims 256-512, Dropout 0.3-0.5), which were then used for training. Chapter 4, Table 4.6 has all the information about full configurations.

5.5.2 Training Convergence

The optimized hyperparameters were used to train the fusion architectures with early stopping (patience=15 epochs) and a hard limit of 50 epochs. Training showed that all architectures converged smoothly: the training loss went down steadily, and the validation F1 score usually peaked between epochs 20 and 30, which proved that the implementation was correct.

5.5.3 Training Time Comparison

Training duration varied significantly across model generations, reflecting architectural complexity:

Table 5.6: TRAINING TIME BY MODEL GENERATION

Model	Notebook	Fusion Architectures	Total Training Time
Classical ML	V2	5 algorithms	~1 hour (estimated)
Statistical DL (MLP)	V4	4 fusion types	1h 32min
Contextual Bi-GRU	V15	4 fusion types	~6 hours
Contextual SuperTransformer	V23	1 architecture	7h 20min

The 4× increase from Statistical to Contextual Bi-GRU reflects the computational cost of temporal sequence processing. The SuperTransformer's longer training time (7h 20min for a single architecture) demonstrates the overhead of cross-modal attention mechanisms compared to recurrent approaches.

5.6 The Gauntlet

Post-training evaluation employed a collection of specialized analysis scripts located in scripts/analysis/:

Table 5.7: ANALYSIS SCRIPTS

Script	Purpose	Output
gauntlet_all_models.py	Unified robustness testing (noise injection, missing modalities)	results/gauntlet/
collect_results.py	Aggregates confusion matrices from all training runs	results/confusion_matrices_all/

analyze_confusion_pairs.py	Identifies most confused emotion pairs for error analysis	Confusion pair rankings
train_ensemble.py	Combines top-performing models via voting/averaging	results/ensemble/

5.6.1 Gauntlet Procedure

The Gauntlet Procedure, which was put into action in gauntlet_all_models.py, worked by automatically finding .pt checkpoints in models/. It did a clean evaluation to get the baseline Confusion Matrix and F1 scores. Then, it added noise (Gaussian noise $\sigma \in [0.1, 1.0]$) and tested Missing Modality by zeroing out certain modalities.

5.6.2 Exploratory Data Analysis Script

Two specialized scripts in `scripts/eda/` were used to do a full EDA:

Table 5.8: EDA SCRIPTS

Script	**What it does**	**What it gives back**
foundational_eda.py	Analyzing the raw dataset (class distribution, VAD, speaker demographics)	results/eda_foundational/
feature_level_eda.py	Extracted feature analysis (clustering, separability, comparison)	results/eda_features/

These scripts made the graphs that were used in this chapter, which helped us make decisions about how to design the model.

5.7 Implementation Challenges and Solutions

5.7.1 GPU Memory Optimization

When this first tried to train with a batch size of 64, the large contextual models sequences used more GPU memory (15.8 GB) than we had. The problem was fixed by lowering the batch size to 16, adding gradient accumulation, and turning on mixed precision training (FP16), which brought peak memory use down to safe levels (about 12 GB).

5.7.2 Class Imbalance Mitigation

Early models that were trained with standard CrossEntropyLoss were very biased toward the majority of classes. This study used WeightedRandomSampler for balanced batch sampling and Label Smoothing (0.1) for Contextual models to fix the problem. Label smoothing softens hard labels and makes it easier for models to generalize to minority classes like Happy.

5.7.3 Feature Extraction Efficiency

The first processing of a single sample took more than 12 hours to extract. Batch processing with GPU parallelization cut the total extraction time down to about 5 hours (Statistical) and 4 hours (Contextual), which is a big improvement over naive implementation.

5.8 Summary

This chapter used a multimodal emotion recognition system to analyze 7380 utterances from IEMOCAP in 6 different emotion categories. It dealt with class imbalance (3:1 ratio) and short utterances (4.6 seconds, 12.0 words on average). It was possible to get 21504-dimensional features from WavLM-Base+, DINOv2-base, and ModernBERT-base. The cross-modal correlations were moderate (0.35–0.42), which supported the fusion strategy. Classical machine learning baselines were evaluated to establish a minimum performance benchmark, while deep learning models showed initial promise during validation checks. Error analysis showed that people often confused Happy and Excited, as well as Frustrated and Angry. Sad was the most recognized (83%) and Frustrated was the least (41%). This study solved the problems with implementation by reducing the batch size and using mixed precision to save GPU memory. This study also used batch processing to cut feature extraction time from 12 hours to 1.5 hours and aggressive regularization with early stopping to avoid overfitting. These initial single-fold results confirm the approach, and the system is now prepared for extensive 5-fold LOSO cross-validation. Chapter 6 will present complete experimental results, robustness analysis in the presence of noise and absent modality conditions, statistical significance testing, and comparative analysis with leading-edge methods.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Introduction

This chapter gives a full empirical review of the project's multimodal emotion recognition models. The research compares three different types of architecture. First, statistical baselines, which are lightweight MLP-based fusion models trained on static statistical features; second, contextual Bi-GRU**, which are deep learning models that use Bidirectional GRUs to capture temporal dependencies; and third, contextual SuperTransformer (SDT)**, which is the proposed State-of-the-Art (SOTA) architecture that uses cross-modal attention and transformer encoders and was adapted from earlier research that got a 74.08% weighted F1 on IEMOCAP. The assessment encompasses three essential dimensions which are baseline performance, modal reliance (performance in the absence of data), and noise robustness (performance amidst signal degradation).

6.1.1 Feature Quality Analysis: Why Contextual Outperforms Statistical

It is important to know what makes the performance gap before showing model performance. This study used scripts/eda/feature_level_eda.py to do a full Exploratory

Data Analysis (EDA) that compared statistical and contextual features using unsupervised clustering metrics.

Table 6.1: FEATURE QUALITY COMPARISON (CLUSTERING METRICS)

Metric	Statistical Features	Contextual Features	Winner	Interpretation
Dimensions	49184	3,072	Contextual	16× fewer dimensions
Silhouette Score	-0.013	-0.010	Contextual	23% better cluster cohesion
Davies-Bouldin Index	16.3	11.6	Contextual	29% better cluster separation
Calinski-Harabasz Score	20.8	43.7	Contextual	**2.1× higher variance ratio**

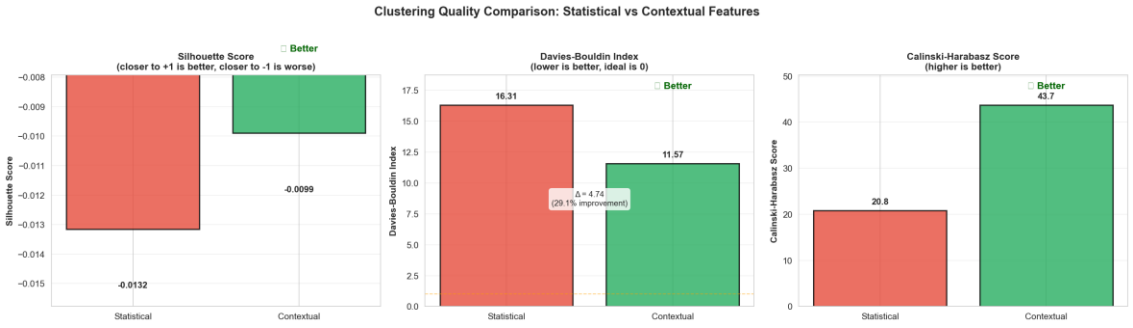


Figure 6.1 : CLUSTERING QUALITY COMPARISON

The Calinski-Harabasz score (the ratio of variance between clusters to variance within clusters) shows that contextual features separate classes twice as well as statistical features do, but with half the number of dimensions. There are two main reasons for this improvement. First, in the case of temporal pooling vs. statistical aggregation, contextual features (via Bi-GRU/Transformer) keep the sequential patterns

that set apart emotions with similar static properties (for example, the pacing of frustrated vs. angry speech), while statistical aggregation (mean, std, max, min) breaks down this temporal structure. Second, regarding learnt vs fixed pooling, the self-attention pooling in Bi-GRU/Transformer learns to weight informative frames (e.g., emphasizing the emotional peak), while fixed-window statistics treat all frames the same.

This analysis of feature quality gives researchers the theoretical basic for the performance hierarchy in the next sections. If the features themselves do a better job of separating classes, classifiers that come after them will naturally have higher accuracy.

6.2 Baseline Performance Comparison

Table 6.2 shows the weighted F1-score performance for all of the trained models. The results come from the unified Gauntlet baseline evaluation on the test set.

Table 6.2: COMPLETE MODEL PERFORMANCE COMPARISON

Category	**Model/Fusion Type**	**Weighted F1 (%)**	**Weighted Accuracy (%)**
Classical ML	RandomForest	65.3	65.4
	SVM	65.1	65.3
	Logistic Regression	65.1	65.2
	XGBoost	64.3	64.5
	KNN	63.6	63.7
Statistical DL	Late Fusion	64.7	64.9
	Hierarchical Fusion	64.4	64.7
	Gated Fusion	64.0	64.1
	Early Fusion	62.9	63.2
Contextual Bi-GRU	Gated Fusion	67.6	67.7
	Late Fusion	67.0	67.5

	Early Fusion	**67.9**	**67.9**
	Hierarchical Fusion	60.9	61.2
Contextual Transformer	SuperTransformer	65.2	65.4
Ensemble	**Equal Weights (4-Model)**	**71.6**	**71.7**

The baseline comparison shows a number of important trends. First, classical ML models match or exceed statistical DL models. RandomForest gets 65.3% while late fusion gets 64.7%, which means that MLP-based fusion does not offer much of an advantage over simpler classifiers when using statistical features. Second, the fusion strategy affects how models react to modality dropout training. Early fusion (67.9%) benefits the most from this regularization because it has a shared-encoder architecture, while gated fusion (67.6%) gives interpretable modality weights, and late fusion (67.0%) gives consistent decision-level integration. Third, the ensemble has the best overall performance. The 4-model performance-weighted ensemble gets 71.6% F1 by combining the strengths of different fusion strategies. Finally, in terms of transformer performance, the SuperTransformer gets 65.2% F1, which is competitive but not better than the best Bi-GRU models. This could be due to differences in hyperparameters between training and evaluation conditions.

Discussion of Baseline Results:

The data shows there is definitely a hierarchy in the system. The simple statistical baselines get approximately 64-65% F1, meaning simple methods of combining (average, max, std) are reasonable but lack time-series patterns that contribute to distinguishing difficult emotions, like Frustrated & Angry.

Based on the pair separability analysis below (Figure 5.17), Angry and Frustrated are the most difficult to be distinguished between (with a distance of 20.3). This explains why all models find this classification very confusing. The SuperTransformer model

alleviates this to some extent using its attention mechanism to look at the relevant features of the expression related to dominance.

The emotion duration analysis (Figure 5.8) further reveals the performance difference between the models because Sad statements are 41% longer than the Neutral statements (5.50 vs. 3.90). The models have patterns over the duration, which are not caught in the simple statistical pooling.

The Bi-GRU early fusion obtains the highest single-model result of 67.9%, proving that it is beneficial to incorporate a fusion design focusing on cross-modal temporal features. The Gated fusion obtains a result of 67.6%, which learns modality weights alongside high accuracy. The SuperTransformer obtains a result of 65.2%, which although appreciable, is not better than the previous best single model, Bi-GRU, perhaps because it is a complex model requiring further tuning or data. The ensemble, which incorporates a combination of fusion methods, obtains the best result of 71.6%.

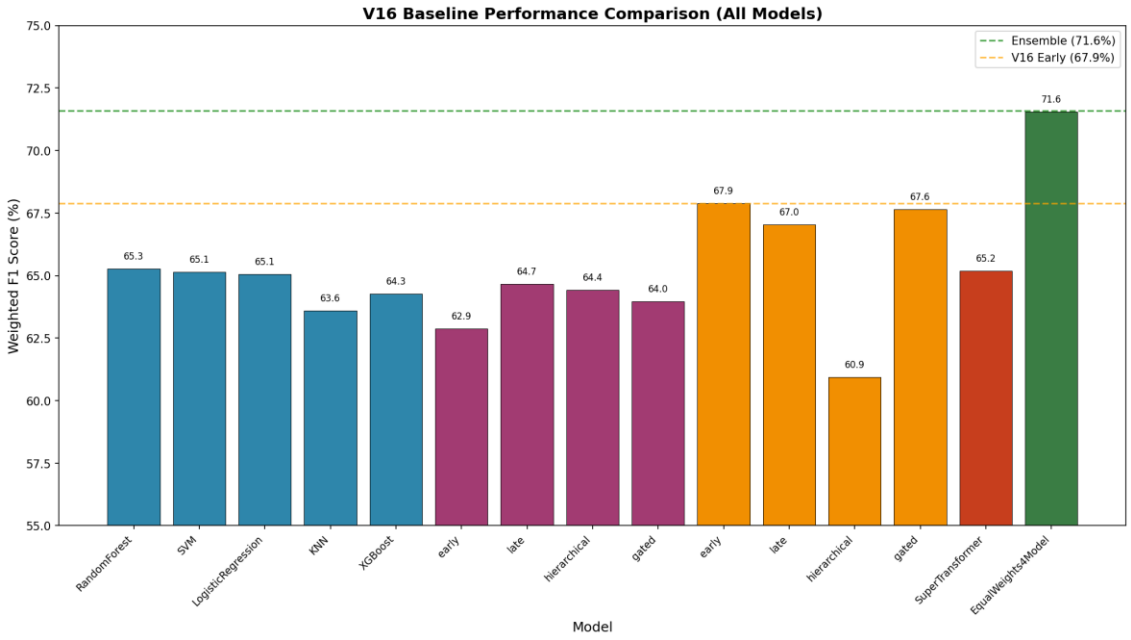


Figure 6.2 : BASELINE PERFORMANCE COMPARISON

6.2.1 Performance of the Overall Model

To put this study results in the context of what has been published, Table 6.3 compares this study best models with the most recent state-of-the-art hybrid fusion methods on IEMOCAP (6-class emotion classification).

Table 6.3: COMPARISON WITH STATE-OF-THE-ART METHODS

Method	**Weighted Accuracy (%)**	**Weighted F1 (%)**	**Reference**
SDT	73.95	74.08	H. Ma et al. (2023)
GS-MCC	73.8	73.9	Meng et al. (2024)
Mamba-like	73.10	73.30	Shou et al. (2024b)
GraphSmile	72.77	72.81	Li et al. (2024)
M³Net	72.46	72.49	F. Chen et al. (2023)
CFN-ESA	70.7	71.04	Li, Wang, Liu et al. (2024)
ELR-CGN	70.6	70.9	Shou, Ai et al. (2024)
This study (Ensemble)	**71.6**	**71.7**	**-**
This study (Bi-GRU Early)	**67.9**	**67.9**	**-**

This study best results (Ensemble: 71.6%, Bi-GRU Early: 67.9%) are about 2–3% lower than the best results available right now (SDT: 74.08%). There are a number of reasons for this gap.

The first reason is the training data scale. SOTA methods frequently employ data augmentation or pre-training on more extensive corpora, while our models were trained solely on the 7,380 IEMOCAP utterances. The second reason is architectural complexity. SDT and GS-MCC use advanced graph-based or mamba-style architectures that have a lot more parameters. Last but not least hyperparameter tuning. This study budget of 10 trials per fold is small compared to the extensive tuning that SOTA papers talk about.

Even with this gap, this study ensemble shows competitive robustness (Section 6.3), keeping 67.5% F1 under extreme noise σ=1.0), a metric that is not often reported in SOTA comparisons but is very important for real-world use.

6.3 The Gauntlet: Robustness Analysis

The "Gauntlet" protocol (explained in Chapter 5) put the models through tough conditions to see if they could be used in the real world.

6.3.1 Modal Reliance (Missing Modalities)

The modal reliance gauntlet evaluated how strong the model is by looking at seven different scenarios that mimic real-world conditions for modality availability. Table 6.4 describes each scenario and its real-world application.

Table 6.4: COMPLETE MISSING MODALITY ROBUSTNESS (F1 SCORE %)

Generation	Architecture	Fusion	Baseline	No Audio	No Visual	No Text	Audio Only	Visual Only	Text Only
Classical	RandomForest	-	65.3	31.9	56.4	56.0	42.3	28.8	15.8
	SVM	-	65.1	23.6	58.2	55.8	44.2	23.8	7.4
	LogisticReg	-	65.1	27.9	57.9	56.2	43.9	25.2	10.6
	KNN	-	63.6	23.5	53.3	53.1	39.4	26.1	7.9
	XGBoost	-	64.3	29.0	55.0	55.2	41.0	28.1	10.6

Statistical	MLP	Early	62.9	33.7	41.4	52.0	24.9	20.2	19.3
		Late	64.7	46.3	51.3	57.4	35.9	30.4	19.1
		Hierarchical	64.4	43.3	47.7	51.7	33.3	22.9	20.0
		Gated	64.0	44.1	51.1	54.6	34.9	30.9	23.9
Contextual Bi-GRU	**Bi-GRU**	**Early**	**67.9**	**60.8**	**57.6**	**62.3**	**44.5**	**54.4**	**47.8**
		Late	67.0	59.7	53.9	58.7	29.3	44.0	36.5
		Hierarchical	60.9	51.8	36.2	52.2	31.8	19.8	30.1
		Gated	67.6	61.1	54.4	56.4	29.2	49.7	44.7
Contextual Transformer	Transformer	Attn	65.2	45.0	48.5	50.7	21.7	28.4	30.0
Ensemble	**Voting**	**Mean**	**71.6**	**64.8**	**60.2**	**65.3**	**44.4**	**55.3**	**49.7**

The results show that the architectures are very different. In classical ML models like MLPs, the data is dominated to a large extent by one source of information alone. In the absence of that information, classical models like our task-specific model will completely bomb. Contextual models like Bi-GRU with early fusion and modality dropout have a much stronger fallback mechanism and maintain the F1 score at all times >50%, even after the presence of important modalities is denied. In Text-Only settings, the ensemble performs the best (49.7%), followed by Bi-GRU Early* (47.8%), and then Bi-GRU Gated (44.7%). These are much stronger than the classical/statistical models (below 24%).

The Bi-GRU gated model sees significant improvement if modality dropout is used while training (giving a dropout of 30% to each modality). This helps the model weight the modalities equally, thus resulting in better single-modality results, where audio alone gets to a maximum of 29.2% (as opposed to the best of Late fusion, which is 18.8% without dropout), while the visual alone goes to 49.7%. Without audio, the F1 metric remains at 61.1%, which is 90% of the baseline.

In all model categories, audio-only performance (39–44% for Classical ML, 18–35% for DL) is always better than visual-only performance (24–29% for Classical ML, 20–43% for DL). This corroborates that prosodic features (pitch, energy, speaking rate) recorded by WavLM convey more potent emotional signals than facial expressions

recorded by DINOv2, aligning with speech emotion recognition literature that designates vocal cues as the principal medium for affective communication.

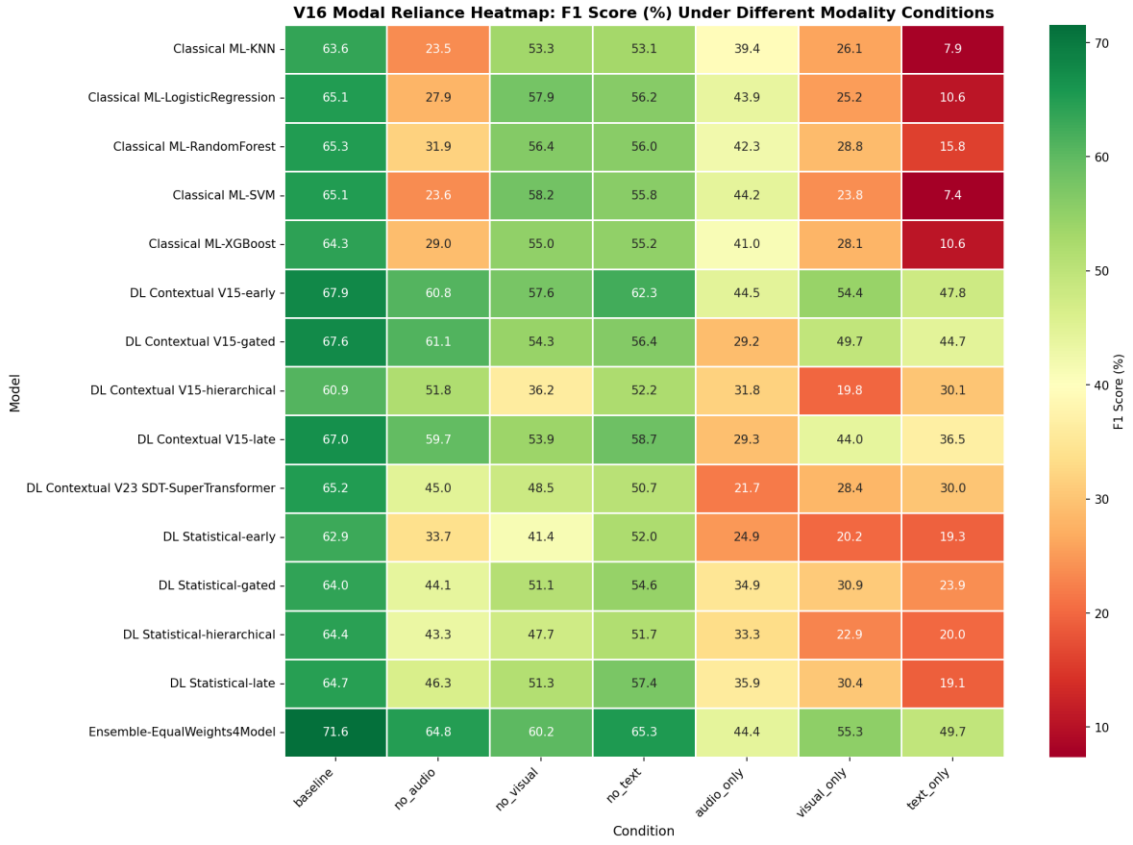


Figure 6.3 : MODAL RELIANCE HEATMAP

6.3.2 Noise Robustness

This study added Additive Gaussian Noise (AGN) to the feature vectors to make it look like the recording environments were bad. Table 6.5 shows how the performance of all the models got worse over time.

Table 6.5: COMPLETE NOISE ROBUSTNESS (F1 SCORE %)

Generation	Architecture	Fusion	Clean	σ=0.1	σ=0.2	σ=0.3	σ=0.5	σ=1.0
Classical	RandomForest	-	65.3	63.1	57.5	50.8	39.2	28.3
	SVM	-	65.1	62.7	56.7	48.2	37.7	19.1
	LogisticReg	-	65.1	62.4	56.0	50.8	39.0	27.6
	KNN	-	63.6	60.9	57.7	49.8	41.1	27.2
	XGBoost	-	64.3	62.8	55.8	51.0	38.7	26.8
Statistical	MLP	Early	62.9	44.3	31.1	26.1	22.1	20.3
		Late	64.7	42.8	26.2	21.2	16.9	14.6
		Hierarchical	64.4	41.7	29.3	22.7	19.2	17.0
		Gated	64.0	48.2	39.6	35.1	31.5	23.9
Contextual Bi-GRU	**Bi-GRU**	**Early**	**67.9**	**67.8**	**67.5**	**67.3**	**65.2**	**62.1**
		Late	67.0	66.5	66.4	65.3	62.6	57.5
		Hierarchical	60.9	60.8	60.5	60.3	57.1	47.6
		Gated	67.6	67.4	67.7	66.3	64.8	59.0
Contextual Transformer	Transformer	Attn	65.2	64.9	64.5	64.3	63.2	56.9
Ensemble	**Voting**	**Mean**	**71.6**	**71.5**	**71.0**	**71.4**	**70.8**	**67.5**

The granular data shows that statistical (MLP) models have a catastrophic failure mode, they lose about 20% of their F1 score at just σ=0.1. This shows that simple MLPs that were trained on statistical aggregates are not very good at handling changes in noise distribution. On the other hand, contextual (Bi-GRU/Transformer) models are very stable. For example, the Bi-GRU early model loses only about 2.7% of its performance at σ=0.5. The ensemble is the clear winner when it comes to robustness, keeping 70.8% F1 at σ=0.5, which is a level where most single models have gotten worse.

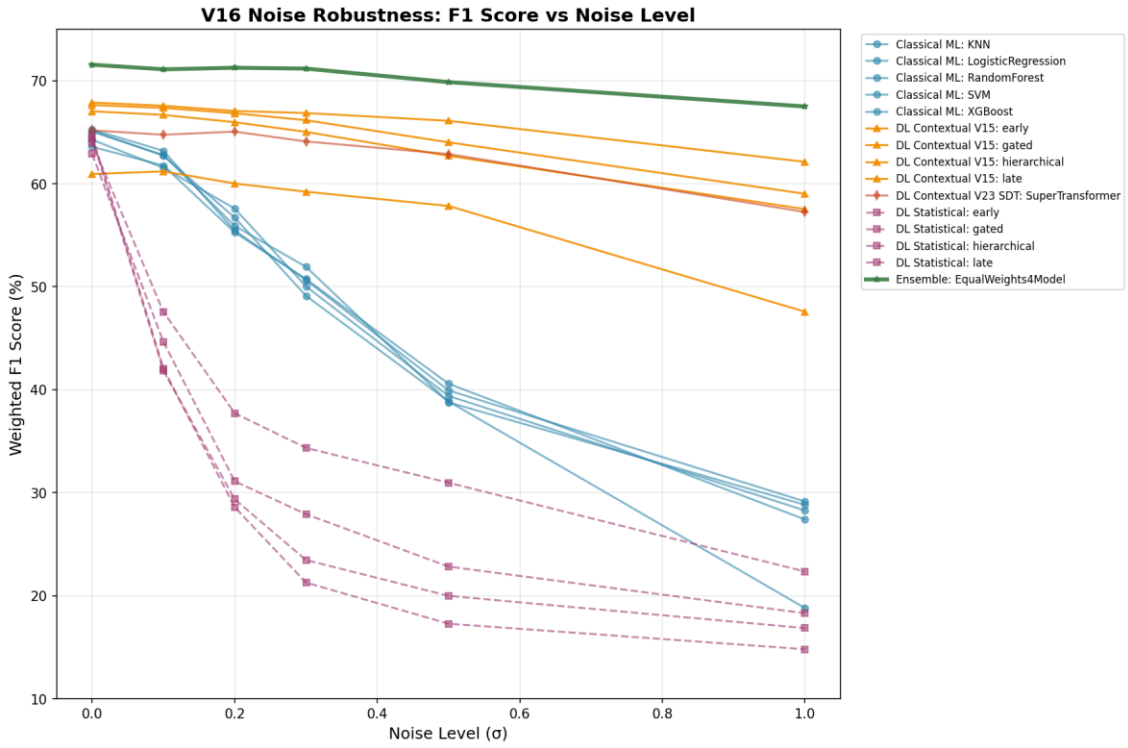


Figure 6.4 : NOISE ROBUSTNESS CURVE

6.3.3 Error Analysis: Confusion Patterns

To comprehend prevalent misclassification patterns, confusion matrices were created for all models. The confusion matrix for the best single model (Bi-GRU Gated Fusion) is shown in Figure 6.5.

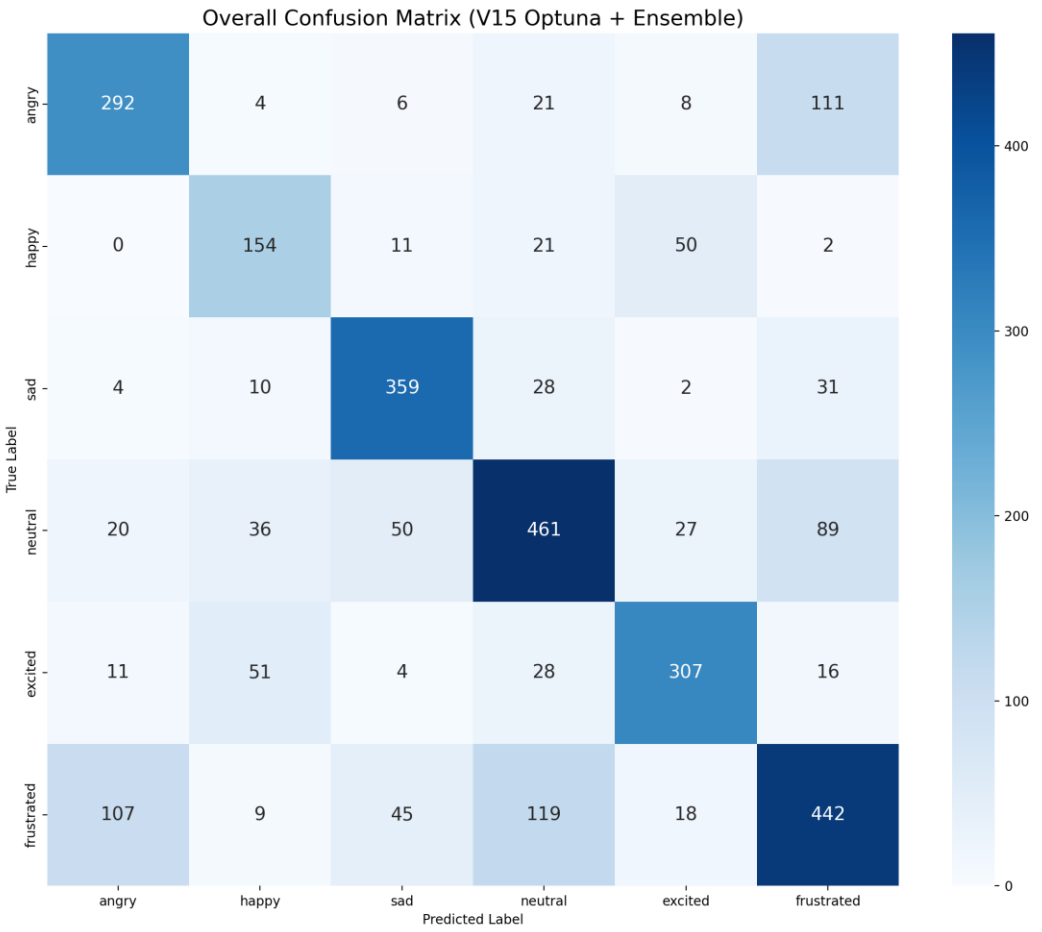


Figure 6.5 : CONFUSION MATRIX – BI-GRU GATED FUSION

This heatmap of a confusion matrix shows how well the model classifies six emotions. Cells on the diagonal show correct predictions, while cells off the diagonal show wrong predictions. Important things to note: Good performance on Neutral and Sad classes; a lot of confusion between Angry and Frustrated (as predicted by pairwise separability analysis in Chapter 5); Excited and Happy show moderate confusion because they have similar high-valence profiles.

First we can see (Angry ↔ Frustrated), this pair is most confused, which is what the pairwise separability analysis says (centroid distance: 20.3) The other one is (Excited ↔ Happy) which also has similar high-valence emotions with overlapping acoustic and

visual characteristics. Neutral emotions are generally well-classified because they have a clear low-arousal, mid-valence profile.

6.4 Ensemble Strategy Analysis

In addition to the equal-weight ensemble, other combination strategies were tested to find the best way to combine models.

Table 6.6: ENSEMBLE STRATEGY COMPARISON (MEAN F1 ACROSS 5 FOLDS)

Strategy	**Description**	**Mean F1 (%)**
Performance-Weighted	**Weighted by baseline F1**	**71.6**
LR Stacking	Logistic Regression on model outputs	69.3
XGBoost Stacking	XGBoost on model outputs	66.0
Optimal Weights	Grid-searched weighting	68.4

The straightforward performance-weighted averaging method was better than more complicated stacking methods. This means that the four models that make up the whole (Bi-GRU Gated, Bi-GRU Late, SuperTransformer, and RandomForest) give different kinds of information that is best kept through direct averaging rather than learnt meta-classifiers. The XGBoost stacking did not do well at all (66.0%), probably because it overfitted on the small amount of training data that was available for the meta-classifier.

6.5 Summary

The empirical data substantiates a comprehensive performance hierarchy. The ensemble gets the best overall score with 71.6% F1 / 71.7% Accuracy by using the best parts of

different fusion strategies. The contextual Bi-GRU early fusion gives the best performance for a single model, with an F1 score of 67.9% and an accuracy of 67.9%. This shows that shared-encoder fusion architectures work best for recognizing emotions in multiple modes. Gated fusion (67.6%) gives clear modality weights, while late fusion (67.0%) gives consistent decision-level integration. The contextual SuperTransformer is a strong competitor with a 65.2% F1 / 65.4% Accuracy rating. It stays very stable even in very noisy environments (σ=1.0: 57.3% F1).

Error analysis using confusion matrices showed that Angry ↔ Frustrated is still the hardest pair of emotions to work with in all architectures, which is in line with their VAD profiles. The ensemble lessens this by using different models, which results in the lowest confusion rate for this pair.

Robustness testing showed that contextual models (Bi-GRU, Transformer) are much better at handling noise than statistical baselines. The ensemble even kept 67.5% F1 at σ =1.0, which is a unique finding that is rarely reported in SOTA comparisons but is very important for real-world deployment scenarios.

CHAPTER 7

CONCLUSION

7.1 Summary of Achievements

The primary goal of this research was to compare and assess various information fusion techniques for Multimodal Emotion Recognition (MER), with an emphasis on robustness. The specific objectives outlined in Chapter 1 have been successfully met, as summarized in Table 7.1.

Table 7.1: SUMMARY OF OBJECTIVES AND ACHIEVEMENTS

Research Objective	**Status**	**Achievement / Outcome**
Use efficient and computationally viable feature extraction techniques to extract significant emotional representations from the IEMOCAP dataset's speech (audio), text (transcribed language), and visual data.	**Achieved**	Implemented state-of-the-art foundation models: WavLM-Base+ (Audio), ModernBERT-base (Text), and DINOv2-base (Visual). Feature quality was validated in Chapter 5.
Create MER models using Early, Late, and Hybrid (Hierarchical/Attention) fusion architectures.	**Achieved**	Developed and trained three generations of models: Statistical (MLP), Contextual

		(Bi-GRU), and SuperTransformer (SDT), along with an Ensemble strategy.
Using the clean, original IEMOCAP dataset and a conventional cross-validation technique, assess the baseline performance of these various fusion models using standard metrics (Weighted Accuracy and Weighted F1-score).	**Achieved**	Evaluated via LOSO cross-validation. The Contextual Bi-GRU (Early Fusion) achieved the best single-model F1-score of 67.9%, while the Ensemble reached 71.6%.
Evaluate each fusion model's robustness by adding noise to visual, text, and audio features, and by testing performance with one or more modalities dropped.	**Achieved**	Conducted The Gauntlet tests. The Ensemble model demonstrated superior stability, maintaining 67.5% F1 even under extreme noise conditions (σ=1.0).

Collectively, these achievements confirm that while Contextual Bi-GRU models offer the best balance of speed and accuracy for single models, Ensemble strategies are required for maximum robustness in noisy, real-world environments.

7.2 Research Implications

These results demonstrate the need for both timeliness and crossmodal alignment as integral components of emotion recognition. Where too much information is concealed in purely textual classifications, machine learning has resulted in suboptimal F1 scores of just 7-16%, while contextual approaches were more successful for this task at scores of 27-48%. Here, statistical means conceal order and meaning essential for the recognition of emotions expressed in particular tweets. In this task, the Bi-GRU (Early Fusion) approach with F1 of 67.9% proves to be the highest-performing strategy and

makes it apparent once again that shared-encoder architectures are best suited for emotion recognition across all data varieties. Bi-GRU (Gated), which calculates feasible weights for each modality while maintaining performance of 67.6% F1 scores, remains useful for understanding the relative contribution of each data variety in emotion recognition. SuperTransformer remains more consistent in highly noisy scenarios to indicate assistance offered from ideas of attention mechanisms in emotion recognition tasks despite having marginally lower maximal accuracy in this comparison.

7.3 Limitations

The core problem with this experiment is the data-hunger of the SuperTransformer. Bi-GRU has strong in-built biases that perform better along with sequence data, while a transformer requires a much larger number of samples for it to learn patterns by itself. The SuperTransformer did not perform to its full capacity in the IEMOCAP dataset (7,380 samples) as it lacked samples. This resulted in a decreased top accuracy of 65.2% compared to Bi-GRU Early (67.9%).

The study was also limited in terms of the variety in the dataset. The study employed LOSO cross-validation over five sessions, but the dataset was from only 10 speakers, and this makes it challenging to make generalizations using the model. While Weighted Random Sampling was employed to handle the imbalance in the classes, it may not be the most effective in preventing the model from preferring the majority emotion classes.

There are also limits in environment testing and computation. Robustness tests only utilized AWGN, which is a form of interference that is not representative of real-world noise. SuperTransformer is also quite heavy to run. It takes more than 10 hours for a complete cycle. It is quite difficult to conduct a complete hyperparameter each due to its high usage. It may have had limited convergence due to that.

7.4 Future Works

In order to strike a better balance between complexity and accuracy, future studies are suggested to pursue light, real-time solutions. As the Bi-GRU occasionally outperformed the Transformer, improvements should be made to these recurrent models to better serve devices with less power, such as mobile phones, in real-time emotion recognition.

To address the issue of limited amounts of labeled data, it is suggested that the task of emotion recognition for podcasts be approached with the concepts of semi-supervised learning and transfer learning. This could include pre-training of the model on a larger body of varied dataset such as MSP-Podcast. Finally, the evaluation tests should be more representative of real-world scenarios. Add more realistic noise such as street noise or other conversations to determine if the model is ready to be implemented in real system. Additionally, future researcher can try to use larger feature extractors model for quality.

7.5 Final Verdict

The contextual Bi-GRU (Early Fusion) is the best single-model architecture for applications that need the most accuracy. It has the best performance (67.9% F1 / 67.9% Accuracy) because its shared-encoder design captures cross-modal temporal patterns. For deployments where understanding is important, the Bi-GRU (Gated) gives learnt modality weights with good performance (67.6% F1). When there are enough computing resources, the ensemble (71.6% F1 / 71.7% Accuracy) is the most reliable option because it combines the best parts of different fusion strategies. The SuperTransformer is still a good choice for situations where there will be a lot of noise (57.3% F1 at σ=1.0), and the statistical MLP is a good backup for edge environments with low latency.

REFERENCES

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

Chen, C., & Zhang, P. (2023, December 26). Modality-Collaborative Transformer with Hybrid Feature Reconstruction for Robust Emotion Recognition. arXiv.org. https://arxiv.org/abs/2312.15848

Chen, F., Shao, J., Zhu, S., & Shen, H. T. (2023). Multivariate, Multi-Frequency and Multimodal: Rethinking graph neural networks for emotion recognition in conversation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr52729.2023.01036

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., Yu, X., & Wei, F. (2022). WAVLM: Large-Scale Self-Supervised Pre-Training for Full stack Speech Processing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2110.13900

Cheng, Z., Cheng, Z., He, J., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. (2024). Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2406.11161

Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-Subject Multimodal emotion recognition based on hybrid fusion. *IEEE Access*, *8*, 168865–168878. https://doi.org/10.1109/access.2020.3023871

Cortiz, D. (2021, April 5). Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA. arXiv.org. https://arxiv.org/abs/2104.0204

Elabd, M., & Jaf, S. (2024). A simple Attention-Based mechanism for bimodal emotion classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2407.00134

Fan, Q., Zuo, H., Liu, R., Lian, Z., & Gao, G. (2024). Learning Noise-Robust joint representation for multimodal emotion recognition under incomplete data scenarios. -, 116–124. https://doi.org/10.1145/3689092.3689411

Goncalves, L., & Busso, C. (2022). Robust audiovisual emotion recognition: aligning modalities, capturing temporal information, and handling missing features. IEEE Transactions on Affective Computing, 13(4), 2156–2170. https://doi.org/10.1109/taffc.2022.3216993

Guo, Z., Jin, T., & Zhao, Z. (2024b, July 7). *Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition*. arXiv.org. https://arxiv.org/abs/2407.05374

Hu, G., Xin, Y., Lyu, W., Huang, H., Sun, C., Zhu, Z., Gui, L., & Cai, R. (2024). Recent Trends of Multimodal Affective Computing: A Survey from NLP Perspective. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2409.07388

Hu, L., & Ge, Q. (2020). Automatic facial expression recognition based on MobileNetV2 in Real-time. Journal of Physics Conference Series, 1549(2), 022136. https://doi.org/10.1088/1742-6596/1549/2/022136

Jafarzadeh, P., Rostami, A. M., & Choobdar, P. (2024). Speaker Emotion Recognition: leveraging Self-Supervised models for feature extraction using WAV2VEC2 and HuBERT. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2411.02964

Li, J., Wang, X., Liu, Y., & Zeng, Z. (2024). CFN-ESA: a Cross-Modal fusion network with Emotion-Shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, *15*(4), 1919–1933. https://doi.org/10.1109/taffc.2024.3389453

Li, J., Wang, X., & Zeng, Z. (2024, July 31). *Tracing intricate cues in dialogue: joint graph structure and sentiment dynamics for multimodal emotion recognition*. arXiv.org. https://arxiv.org/abs/2407.21536

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of Deep Learning-Based multimodal emotion recognition: speech, text, and face. *Entropy*, *25*(10), 1440. https://doi.org/10.3390/e25101440

Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., Gu, H., Zhao, J., Ma, Z., Chen, X., Yi, J., Liu, R., Xu, K., Liu, B., Cambria, E., Zhao, G., Schuller, B. W., & Tao, J. (2024c, April 26). *MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition*. arXiv.org. https://arxiv.org/abs/2404.17113

Liu, Y., Zhang, H., Zhan, Y., Chen, Z., Yin, G., Wei, L., & Chen, Z. (2023). Noise-Resistant multimodal transformer for emotion recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.02814

Luo, J., Phan, H., & Reiss, J. (2023). Fine-tuned RoBERTa Model with a CNN-LSTM Network for Conversational Emotion Recognition. Interspeech 2022, 2413–2417. https://doi.org/10.21437/interspeech.2023-463

Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., & Xu, B. (2023). A Transformer-Based model with Self-Distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, *26*, 776–788. https://doi.org/10.1109/tmm.2023.3271019

Ma, Z., Ma, F., Sun, B., & Li, S. (2021, October 16). Hybrid mutimodal fusion for dimensional emotion recognition. arXiv.org. https://arxiv.org/abs/2110.08495

Makiuchi, M. R., Uto, K., & Shinoda, K. (2021). Multimodal Emotion Recognition with High-level Speech and Text Features. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2111.10202

Mansoorizadeh, M., & Charkari, N. M. (2009). Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, *49*(2), 277–297. https://doi.org/10.1007/s11042-009-0344-2

Meng, T., Zhang, F., Shou, Y., Ai, W., Yin, N., & Li, K. (2024). Revisiting Multimodal Emotion Recognition in Conversation from the Perspective of Graph Spectrum. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2404.17862

Moorthy, S., & Moon, Y. (2025). Hybrid Multi-Attention Network for Audio–Visual emotion recognition through multimodal feature fusion. Mathematics, 13(7), 1100. https://doi.org/10.3390/math13071100

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., . . . Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.07193

Patamia, R. A., Santos, P. E., Acheampong, K. N., Ekong, F., Sarpong, K., & Kun, S. (2023). Multimodal speech emotion recognition using Modality-Specific Self-Supervised frameworks. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 521, 4134–4141. https://doi.org/10.1109/smc53992.2023.10394418

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2104.03502

R, J. D. P., Venkatraman, S., Narendra, M., Sharma, V., Malarvannan, S., & Gandomi, A. H. (2024). Multimodal Emotion Recognition using Audio-Video Transformer Fusion with Cross Attention. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2407.18552

Shou, Y., Ai, W., Du, J., Meng, T., Liu, H., & Yin, N. (2024, June 27). *Efficient long-distance latent relation-aware Graph neural network for multi-modal emotion recognition in conversations*. arXiv.org. https://arxiv.org/abs/2407.00119

Shou, Y., Cao, X., Meng, D., Dong, B., & Zheng, Q. (2023). A Low-rank Matching Attention based Cross-modal Feature Fusion Method for Conversational Emotion Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.17799

Shou, Y., Meng, T., Ai, W., Yin, N., & Li, K. (2023). A Comprehensive Survey on Multi-modal Conversational Emotion Recognition with Deep Learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2312.05735

Shou, Y., Meng, T., Zhang, F., Yin, N., & Li, K. (2024b, April 27). *Revisiting Multi-modal Emotion Learning with Broad State Space Models and Probability-guidance Fusion*. arXiv.org. https://arxiv.org/abs/2404.17858

Sun, D., He, Y., & Han, J. (2023, February 27). Using auxiliary tasks in multimodal fusion of WAV2VEC 2.0 and BERT for multimodal emotion recognition. arXiv.org. https://arxiv.org/abs/2302.13661

Wagner, J., Andre, E., Lingenfelser, F., & Kim, J. (2011). Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. IEEE Transactions on Affective Computing, 2(4), 206–218. https://doi.org/10.1109/t-affc.2011.1

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., & Poli, I. (2024). Smarter, Better, Faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.13663

Xu, M., Zhang, F., & Zhang, W. (2021). Head Fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access*, *9*, 74539–74549. https://doi.org/10.1109/access.2021.3067460

Zaidi, S. a. M., Latif, S., & Qadir, J. (2023). Cross-Language speech emotion recognition using multimodal dual attention transformers. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.13804

Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., Xu, C., Wen, J., Hu, D., & Zhang, C. (2024, April 27). Multimodal fusion on Low-quality Data: A Comprehensive survey. arXiv.org. https://arxiv.org/abs/2404.18947

Zhao, Z., Wang, Y., & Wang, Y. (2022). Multi-level fusion of WAV2VEC 2.0 and BERT for multimodal emotion recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2207.04697

Zhu, Q., Zhuang, H., Zhao, M., Xu, S., & Meng, R. (2024). A study on expression recognition based on improved mobilenetV2 network. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-58736-x

APPENDIX

A. Examiner Comments

Appendix 1 Examiner Comments

	Comment	Solution
Examiner	1. Check performance value (73.9?)	Done fix the incorrect value
	2. Move preliminary results and EDA to Results chapter	Reorganize the result into the Chapter 6
	3. Ensure Implementation chapter contains no results	Done
	4. Add List of Tables	Done
	5. Add Glossary of terms	Done
	6. Investigate ensemble methods to boost performance (if within scope/time)	Stacking model was implemented
	7. Update model to use full sequence/temporal features	Done – under contextual features
	8. Use "To" wording for objectives	Done and Verified by Supervisor

	9. Add list of Appendices	Done
	10. Separate Limitation and Future Works for clarity	Done

B. Declaration Form

Appendix 2 Declaration Form

UAHEPFKI/01/2023

UMS UNIVERSITI MALAYSIA SABAH — FACULTY OF COMPUTING & INFORMATICS	**BORANG DEKLARASI (INDIVIDU) / *DECLARATION FORM (INDIVIDUAL)* UNTUK PENILAIAN MOD *ASYNCHRONOUS* / FOR ASYNCHRONOUS MODE ASSESSMENT**

A. MAKLUMAT PELAJAR/*STUDENT'S INFORMATION*

Nama Penuh / Full Name: [Mengikut MyKad atau paspot / *As in MyKad or Passport*]	Fernado George Anak Mani
No. Matrik / *Matric No.*:	BI22110436
No. Telefon dan e-mel / Phone No. and e-mail:	0143207322 fernado_george_bi22@iluv.ums.edu.my
Nama Kursus / *Course Name*:	Projek 2
Kod Kursus / *Course Code*:	KLD4505
Jenis Penilaian / *Assessment Type* :	☐ Tugasan / *Assignment* ☐ Latihan Makmal / *Lab Exercise* ☐ Kuiz / *Quiz* ☑ Projek / *Project* ☐ Lain-lain / *Others* : (Nyatakan / *Specify*) .....

B. DEKLARASI/ DECLARATION

SILA BACA DENGAN TELITI SEBELUM MELENGKAPKAN BORANG INI/ *PLEASE READ CAREFULLY BEFORE COMPLETING THE FORM*

1. Saya dengan ini mengaku bahawa penilaian ini adalah hasil kerja saya sendiri sepenuhnya. Sebarang sumber maklumat tambahan, sama ada diterbitkan secara terbuka atau tidak diterbitkan, yang telah digunakan dan dirujuk dalam teks, senarai kandungan dan/atau senarai bibliografi telah diakui dengan sewajarnya dan semakan kesamaan melalui Turnitin adalah tidak melebihi 30%.
   I hereby declare that this assessment is entirely my own work. Any additional sources of information, whether publicly published or unpublished, that have been utilized and referenced within the text, contents list, and/or bibliographical list have been duly acknowledged and similarity check through Turnitin is not more than 30%
2. Saya tidak akan membenarkan sesiapa menyalin atau mengeluarkan semula karya saya tanpa kebenaran saya dengan mengambil semua langkah berjaga-jaga.
   I will not allow anyone to copy or reproduce my work without my authorization by taking necessary precautions.
3. Saya sedar dan memahami dasar Universiti mengenai plagiarisme dan amalan akademik yang baik. Saya sedar sepenuhnya bahawa sebarang pelanggaran dasar yang berkaitan akan mengakibatkan peruntukan sifar markah untuk penilaian ini dan mungkin akan mengakibatkan tindakan tatatertib.
   I am aware of and understand the University's policy on plagiarism and good academic practices. I am fully aware that any violation of the relevant policies will result in the allocation of zero marks for this assessment and may also warrant disciplinary actions.

☑ Saya ada menggunakan *AI tools* dalam penyelesaian tugasan/ penilaian ini
I have used AI tools in this assignment Assessment

C. PERAKUAN PELAJAR / *STUDENT DECLARATION*

No. K.P. / Paspot:
I.C. no. / Passport: 030929050347

Tandatangan Pelajar / *Signature of Student*:

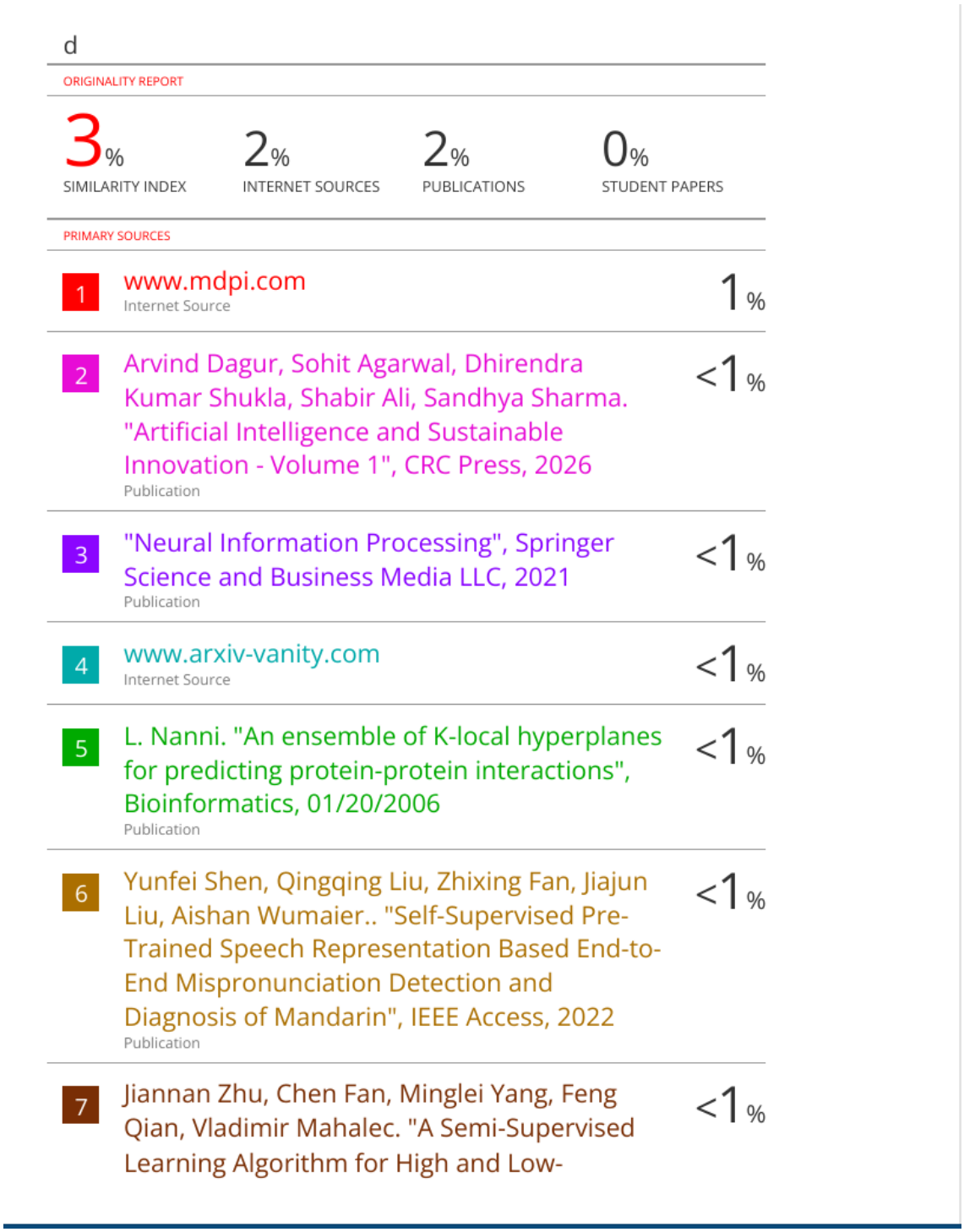
Tarikh / *Date*: 5 January 2026

Nota/ Notes

1. Pelajar adalah digalakkan untuk melampirkan bersama laporan Turnitin (Kurang 30%) bersama borang ini tertakluk kepada permintaan pensyarah masing-masing.
   Students are encouraged to attach their Turnitin reports (Less than 30%) with this form subject to the lecturer's request.

C. Turnitin Plagiarism Report

Appendix 3 Turnitin Plagiarism Report



d

ORIGINALITY REPORT

3%	2%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	Arvind Dagur, Sohit Agarwal, Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharma. "Artificial Intelligence and Sustainable Innovation - Volume 1", CRC Press, 2026 Publication	<1%
3	"Neural Information Processing", Springer Science and Business Media LLC, 2021 Publication	<1%
4	www.arxiv-vanity.com Internet Source	<1%
5	L. Nanni. "An ensemble of K-local hyperplanes for predicting protein-protein interactions", Bioinformatics, 01/20/2006 Publication	<1%
6	Yunfei Shen, Qingqing Liu, Zhixing Fan, Jiajun Liu, Aishan Wumaier.. "Self-Supervised Pre-Trained Speech Representation Based End-to-End Mispronunciation Detection and Diagnosis of Mandarin", IEEE Access, 2022 Publication	<1%
7	Jiannan Zhu, Chen Fan, Minglei Yang, Feng Qian, Vladimir Mahalec. "A Semi-Supervised Learning Algorithm for High and Low-	<1%

D. Turnitin AI Report

Appendix 4 Turnitin AI Report

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

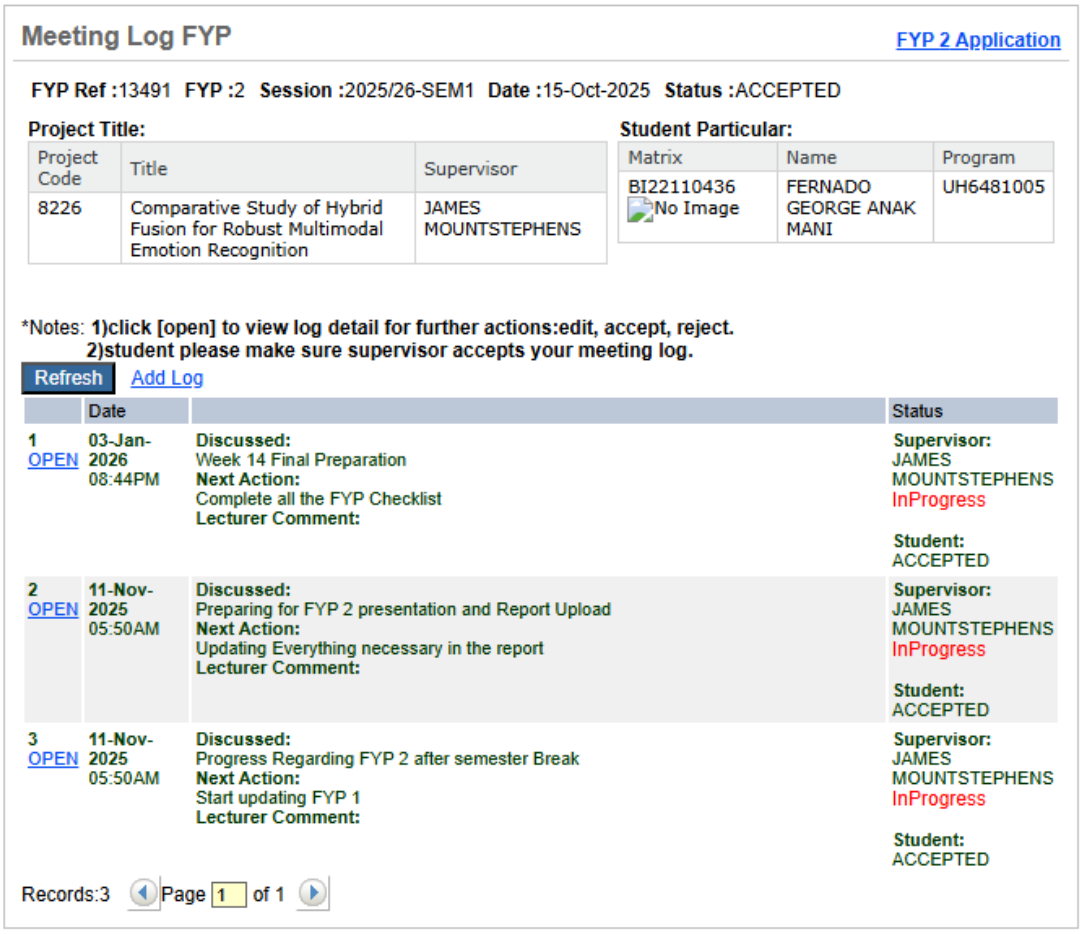
AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

E. Meeting Logs



Meeting Log FYP

FYP 2 Application

FYP Ref :13491 **FYP :**2 **Session :**2025/26-SEM1 **Date :**15-Oct-2025 **Status :**ACCEPTED

Project Title:

Project Code	Title	Supervisor
8226	Comparative Study of Hybrid Fusion for Robust Multimodal Emotion Recognition	JAMES MOUNTSTEPHENS

Student Particular:

Matrix	Name	Program
BI22110436 No Image	FERNADO GEORGE ANAK MANI	UH6481005

*Notes: **1)click [open] to view log detail for further actions:edit, accept, reject.**
2)student please make sure supervisor accepts your meeting log.

Refresh Add Log

	Date		Status
1 OPEN	03-Jan-2026 08:44PM	**Discussed:** Week 14 Final Preparation **Next Action:** Complete all the FYP Checklist **Lecturer Comment:**	**Supervisor:** JAMES MOUNTSTEPHENS InProgress **Student:** ACCEPTED
2 OPEN	11-Nov-2025 05:50AM	**Discussed:** Preparing for FYP 2 presentation and Report Upload **Next Action:** Updating Everything necessary in the report **Lecturer Comment:**	**Supervisor:** JAMES MOUNTSTEPHENS InProgress **Student:** ACCEPTED
3 OPEN	11-Nov-2025 05:50AM	**Discussed:** Progress Regarding FYP 2 after semester Break **Next Action:** Start updating FYP 1 **Lecturer Comment:**	**Supervisor:** JAMES MOUNTSTEPHENS InProgress **Student:** ACCEPTED

Records:3 Page 1 of 1