MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ MOHAMED KHIDER - BISKRA
FACULTÉ DES SCIENCES ET DE LA TECHNOLOGIE
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE

Thèse de Doctorat

En vue de l'obtention du diplôme de docteur LMD en Automatique

Reconnaissance et Analyse du Comportements Humains Individuels et Collectif par la Combinaison de l'Information de l'Image et du Signal Parole

Réalisé par: **Zineddine Sarhani Kahhoul**

Soutenue publiquement le: 16/12/2025

Devant le jury composé de:

Zine Eddine Baarir	Professeur	Université de Biskra	Président
Nadjiba Terki	Professeur	Université de Biskra	Rapporteur
Abdelkrim Ouafi	Professeur	Université de Biskra	Examinateur
Abdelhamid Messaoudi	MCA	HNS RE2SD, Batna	Examinateur

Année Universitaire 2025/2026

MOHAMED KHIDER UNIVERSITY - BISKRA
FACULTY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING
DIVISION OF AUTOMATIC
Ref: .....

جامعة محمد خيضر - بسكرة
كلية العلوم والتكنولوجيا
قسم الهندسة الكهربائية
شعبة الآلية
المرجع: .....

Thesis title:

Recognition and Analysis of Individual and Collective Human Behaviors by Combining Image Information and Speech Signal

By

ZINEDDINE SARHANI KAHHOUL

Thesis submitted to the department of Electrical Engineering in candidacy for the Degree of **Doctorate (3rd Cycle)** in **Automatic**.

Members of the jury:

President:	Zine Eddine Baarir	Profesor	University of Biskra
Supervisor:	Nadjiba Terki	Profesor	University of Biskra
Examiner:	Abdelkrim Ouafi	Profesor	University of Biskra
Examiner:	Abdelhamid Messaoudi	MCA	HNS RE2SD

2025/2026

Dedication

In loving memory of my mother, Naima,
who always believed this day was possible.

And for my father, Omar,
who taught me the perseverance to reach it.

To my dear brother and sisters,
for their constant love and support.

And to all my respected teachers,
for illuminating the path to knowledge.

And finally, to everyone I hold dear.

ACKNOWLEDGEMENTS

First and foremost, all praise and gratitude are due to Allah the Almighty. This work would not have been possible without His divine guidance and blessings.

I wish to express my deepest gratitude to my supervisor, Professor Nadjiba Terki. Her invaluable guidance, unwavering support, and expert mentorship have been instrumental to this research. My sincere appreciation also goes to the members of the defense committee for their time and insightful comments.

My sincere thanks are also extended to Professor Habiba Dahmani for her valuable collaboration and contribution to this study. I am grateful for the stimulating and supportive academic environment created by my colleagues from LI3CUB, VSC (University of Biskra), and the Electrical Engineering team. Within this community, I am particularly indebted to Ilyes Bennaissa, Tiar Mohamed , Moustari Abderaouf and Selma Boutaiba. It has been a privilege to work alongside you as fellow PhD students and to share in the success of our collaborative research and publications.

I gratefully acknowledge the support from the Algerian Ministry of Higher Education and the Faculty of Science and Technology at the University of Biskra.

Finally, my deepest gratitude is reserved for my family. To the loving memory of my dear mother, Naima, whose boundless love and belief in me remain my greatest inspiration. To my father, Omar, who taught me how to face life's challenges with courage. Thank you to my brother Sami, my sisters, and my brother-in-law Dhafer for your constant support. A special mention to my nephews, Iyad and Yazan, who bring immense joy and hope for the future to our lives.

Abstract

The automated recognition of human emotion is a cornerstone of modern affective computing, yet progress is often hindered by the limitations of unimodal analysis, the performance gap on real-world data, and a critical lack of resources for under-resourced languages. This thesis presents a comprehensive framework to address these challenges, with a deep focus on advancing the state-of-the-art in Automatic Speech Emotion Recognition (ASER).

The research makes three primary contributions. First, an efficient and lightweight architecture, the CBAM-DenseNet121, is proposed to resolve the trade-off between accuracy and computational complexity. By integrating an attention mechanism with a dense convolutional network, this model achieves highly competitive performance on the benchmark CREMA-D dataset while utilizing substantially fewer parameters than comparable state-of-the-art models.

Second, a novel high-accuracy framework is introduced, combining a custom DeepSpec-CNN with an architecturally diverse ensemble learning strategy. This approach reframes the classification problem using the control dimension of the Geneva Wheel of Emotions (GWE), establishing a new state-of-the-art performance on CREMA-D by significantly improving upon existing methods.

Finally, to address data scarcity, this thesis introduces the Open Your Heart (OYH) corpus, a new, large-scale dataset containing several hours of genuine emotional speech in the Algerian Arabic dialect. Comprehensive performance baselines were established on this challenging corpus using traditional machine learning models, providing a vital new benchmark for future research.

Collectively, this thesis advances the field through the dual contribution of novel, high-performance ASER models and the creation of an essential new corpus. The findings provide a robust foundation for building more nuanced, culturally aware, and socially intelligent systems.
Keywords: Automatic Speech Emotion Recognition (ASER), Deep Learning, Convolutional Neural Networks (CNN), Attention Mechanisms, Ensemble Learning, Spectrograms, Affective Computing, Algerian Arabic, Speech Corpus.

الملخص:

تهدف هذه الأطروحة إلى التعرّف على المشاعر والسلوكيات الإنسانية وتحليلها، مع تركيز عميق على تطوير آخر ما توصل إليه البحث في مجال **التعرّف الآلي على المشاعر من الكلام (ASER)** . في هذه الدراسة، نقترح ثلاث مساهمات أساسية لمواجهة التحديات الرئيسية في هذا المجال: أولاً، نموذج تعلّم عميق فعّال ومدعّم بآلية الانتباه للتصنيف القياسي للمشاعر؛ ثانيًا، إطار عمل جماعي عالي الدقة يعتمد على نهج بُعدي مبتكر؛ وثالثًا، مدوّنة لغوية جديدة وواسعة النطاق للهجة الجزائرية العربية قليلة الموارد. الفقرات التالية تقدم ملخصًا موجزًا لكل مساهمة.

يقترح الجزء الأول من هذا البحث معمارية خفيفة الوزن وعالية الكفاءة، وهي **CBAM-DenseNet121**، لمعالجة المفاضلة الحرجة بين الدقة والتعقيد الحسابي في أنظمة ASER . يعزز هذا النموذج قدرات إعادة استخدام الميزات القوية لشبكة DenseNet-121 الأساسية من خلال **وحدة الانتباه الكتلية الالتفافية (CBAM)**، مما يسمح للشبكة بالتركيز بشكل تكيّفي على الأنماط الزمنية-الترددية الأكثر بروزًا في سبكتروغرامات ميل اللوغاريتمية. أظهرت التجارب التي أجريت على مجموعة البيانات المعيارية **CREMA-D** أن النموذج المقترح يحقق دقة تنافسية عالية، متفوقًا على العديد من النماذج الحديثة مع الحفاظ على بنية خفيفة الوزن بعدد معلمات أقل بكثير.

تقترح هذه الرسالة أيضًا إطار عمل مبتكرًا وعالي الدقة يجمع بين معمارية مصممة خصيصًا، وهي **DeepSpecCNN**، واستراتيجية **تعلّم جماعي** متنوعة معماريًا. تقدم هذه الدراسة مخطط تصنيف جديدًا عبر إعادة صياغة مشكلة التصنيف السداسي إلى مهمة ثنائية أكثر قوة، استنادًا إلى **بُعد التحكّم** من **عجلة جنيف للمشاعر (GWE)** . يتم بعد ذلك تجميع تنبؤات الشبكات العصبونية الالتفافية المتعددة عبر آلية التصويت بالأغلبية لإنتاج تصنيف نهائي تآزري. عند تقييمه على مجموعة بيانات **CREMA-D**، أثبت النموذج الجماعي المكون من 5 نماذج أداءً هو الأحدث من نوعه، محققًا تحسنًا كبيرًا على الطرق الحالية لهذه المهمة.

أخيرًا، ولمعالجة النقص الحاد في الموارد للكلام التلقائي وغير الإنجليزي، تقدم هذه الأطروحة **مدوّنة افتح قلبك (OYH)** وهي مجموعة بيانات جديدة وواسعة النطاق تحتوي على عدة ساعات من الكلام العاطفي الحقيقي الذي تم جمعه من برنامج حواري تلفزيوني واقعي باللهجة الجزائرية العربية. تم تقطيع المدوّنة وتوصيفها بدقة باستخدام إطار GWE البُعدي. تم تأسيس خطوط أساس قوية للأداء باستخدام مجموعة من نماذج التعلّم الآلي التقليدية مع استراتيجية مُحسَّنة لاختيار الميزات، مما أدى إلى تحقيق دقة أساسية قوية وتوفير معيار قياسي حيوي للبحوث المستقبلية في اللغات قليلة الموارد.

الكلمات المفتاحية: التعرّف الآلي على المشاعر من الكلام(ASER) ؛ التعلّم العميق؛ الشبكات العصبونية الالتفافية (CNN)؛ آليات الانتباه؛ التعلّم الجماعي؛ السبكتروغرام؛ الحوسبة العاطفية؛ اللهجة الجزائرية؛ مدوّنة لغوية صوتية.

Publications in Journals

- **Kahhoul, Z.S.**, Terki, N., Dahmani, H., et al., "Automatic Speech Emotion Recognition for Arabic Dialects: A New Dataset and Machine Learning Framework," *Cluster Computing*, vol. 29, pp. 26, 2026.
- **Kahhoul, Z.S.**, Terki, N., Tiar, M.L., et al., "Ensemble Learning for Improved Speech Emotion Recognition: A Control Dimension Analysis of Log-Mel Spectrograms from the CREMA-D Dataset," *Signal, Image and Video Processing (SIViP)*, vol. 19, pp. 1135, 2025.
- **Kahhoul, Z.S.**, Terki, N., Benaissa, I., Aldwoah, K., Hassan, E.I., Osman, O., Boukhari, D.E., "Mathematical Analysis and Performance Evaluation of CBAM-DenseNet121 for Speech Emotion Recognition Using the CREMA-D Dataset," *Applied Sciences*, vol. 15, no. 17, pp. 9692, 2025.

Publications in International Conferences

- **Kahhoul, Z.S.**, Terki, N., Benaissa, I., Baarir, Z., "SpectoResNet: Advancing Speech Emotion Recognition...," in *Proc. of the 5th Int. Electronic Conf. on Applied Sciences (ECAS'24)*, Online, Dec 2024.
- **Kahhoul, Z.S.**, Boutiba, S., Dahmani, H., et al., "Algerian Database Under Development for Automatic Emotion Recognition," in *Proc. of the 1st Int. Conf. on Advances in Electronics, Control and Computer Technologies (ICAECCT'23)*, Biskra, Oct 2023.

Publications in National Conferences

- **Kahhoul, Z.S.**, Terki, N., Tiar, M.L., Boutiba, S., "Improving Speech Emotion Recognition: A Control-Based Approach...," in *Proc. of the 1st Nat. Conf. on Renewable Energies and Advanced Electrical Engineering (NCREAEE'25)*, Biskra, Jul 2025.

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

List of Acronyms

HCI	Human-Computer Interaction
ASER	Automatic Speech Emotion Recognition
IoT	Internet of Things
RQ	Research Question
OYH	Open Your Heart
BE-AWF	Backward Elimination of All Worst Features
CBAM	Convolutional Block Attention Module
CNN	Convolutional Neural Network
GWE	Geneva Wheel of Emotions
PAD	Pleasure-Arousal-Dominance
IEMOCAP	Interactive Emotional Dyadic Motion Capture Database
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
MSA	Modern Standard Arabic
LLDs	Low-Level Descriptors
HSFs	High-Level Statistical Features
openSMILE	open-source Speech and Music Interpretation by Large-space Extraction
ComParE	Computational Paralinguistics Challenge

F0	Fundamental Frequency
RMS	Root-Mean-Square
MFCCs	Mel-Frequency Cepstral Coefficients
FFT	Fast Fourier Transform
DCT	Discrete Cosine Transform
HNR	Harmonic-to-Noise Ratio
STFT	Short-Time Fourier Transform
DFT	Discrete Fourier Transform
dB	Decibel
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
SVM	Support Vector Machine
RBF	Radial Basis Function
KNN	K-Nearest Neighbors
GBM	Gradient Boosting Machines
DL	Deep Learning
ReLU	Rectified Linear Unit
MLP	Multilayer Perceptron
Adam	Adaptive Moment Estimation
TP	True Positives
FP	False Positives
FN	False Negatives

TN	True Negatives
UAR	Unweighted Average Recall
ANG	Anger
DIS	Disgust
FEA	Fear
HAP	Happy
SAD	Sad
NEU	Neutral
SOTA	State-of-the-Art
ViT	Vision Transformer
BFE	Backward Feature Elimination
ZCR	Zero-Crossing Rate

GENERAL INTRODUCTION

Background and Motivation

Emotions are the bedrock of human experience, profoundly influencing our daily interactions, cognitive processes, and social bonds. In an era where technology is ever more interwoven with the fabric of society, from ubiquitous voice assistants to sophisticated social robots, the imperative for computational systems to comprehend this emotional layer of communication has never been greater. This need is the driving force behind the vibrant field of Affective Computing, which seeks to bridge the gap between human emotion and computational technology [1]. The ultimate goal is to create systems that can recognize, interpret, and respond to human affects, facilitating more natural, empathetic, and effective Human-Computer Interaction (HCI) [2, 3].

Among the various modalities for conveying emotion, speech is arguably the most direct and information-rich channel. Unlike written text, which is limited to semantic content, the spoken word is layered with a wealth of paralinguistic information that reveals not just what is said, but how it is said. This information is encoded in acoustic cues such as prosody, voice quality, and spectral characteristics. The field of Automatic Speech Emotion Recognition (ASER) is dedicated to computationally decoding this emotional subtext from the audio signal alone [4, 5].

The potential applications of robust ASER systems are transformative. In healthcare, ASER can revolutionize diagnostics by detecting vocal biomarkers for mental health conditions like depression [6]. In education and commercial sectors, it can gauge user engagement and satisfaction in real-time. In pioneering fields like social robotics, ASER is indispensable for creating machines that can interact with humans in a more believable and socially aware manner [7, 8]. This technological pursuit, from its origins in traditional machine learning with handcrafted features to the modern paradigm of deep learning with spectrograms, represents a continuous effort to imbue technology with a fundamental aspect of human understanding.

Problem Statement

Despite significant progress, the practical, real-world application of ASER is hindered by several fundamental and interconnected challenges. The vast majority of research is conducted on datasets of "acted" emotions recorded in sterile environments, creating a critical domain gap; models trained on this data fail to generalize to the subtlety and variability of genuine, "in-the-wild" spontaneous speech [9]. This issue is compounded by a profound scarcity of resources for under-represented languages, particularly for linguistically complex regions like North Africa and its Arabic dialects, which impedes scientific progress and limits the fairness of affective technologies for billions of speakers [10]. Concurrently, the advent of deep learning has introduced an accuracy versus efficiency trade-off, where state-of-the-art performance is often achieved with massive, computationally expensive models that are impractical for deployment on resource-constrained devices. Finally, the field has predominantly relied on categorical emotion models (e.g., anger, joy, sadness), an approach that oversimplifies the true nature of human affect. The ambiguity and overlap between these discrete categories create inconsistent data labels and pose a significant challenge for classifiers, potentially limiting their robustness and performance ceiling. This thesis identifies and addresses these four critical problems as the core of its investigation.

Research Questions and Objectives

In response to the problems outlined above, this thesis seeks to answer the following primary research questions:

RQ1: How can a large-scale, spontaneous speech corpus for an under-resourced dialect (Algerian Arabic) be created and benchmarked to address the critical gaps in data authenticity and availability in the ASER field?

RQ2: Can a lightweight, attention-enhanced deep learning architecture be developed to achieve a competitive balance between accuracy and computational efficiency for standard multi-class speech emotion recognition on a benchmark dataset?

RQ3: Can a new state-of-the-art in ASER accuracy be achieved by combining an architecturally diverse ensemble of models with a dimensional classification framework derived from psychological theory to overcome the limitations of categorical emotion models?

To answer these questions, a set of clear objectives was established. The first objective was to collect, segment, and meticulously annotate a novel corpus of spontaneous Algerian

Arabic emotional speech, the OYH corpus, using the dimensional Geneva Wheel of Emotions framework. Following this, the goal was to establish comprehensive performance baselines on the new corpus using a suite of traditional machine learning classifiers and a sophisticated feature selection algorithm. In parallel, the research aimed to design, implement, and rigorously evaluate the CBAM-DenseNet121 architecture, a novel and efficient model for 6-class emotion recognition on the benchmark CREMA-D dataset. The final objective was to develop and validate a new, high-accuracy ensemble learning framework based on a custom DeepSpecCNN and other diverse architectures, applied to a dimensional 2-class emotion task on the same CREMA-D dataset.

Summary of Contributions

This dissertation makes four primary contributions to the field of Automatic Speech Emotion Recognition:

1. **A Novel Speech Corpus for a Low-Resource Dialect:** The primary contribution is the creation and public description of the **Open Your Heart (OYH) dataset**. This is a new, large-scale (6.3 hours) corpus of spontaneous, real-world emotional speech for the Algerian Arabic dialect. It directly addresses the critical scarcity of non-English, non-acted data and provides a challenging new benchmark for the research community.

2. **An Efficient, Attention-Enhanced ASER Model:** This thesis proposes the **CBAM-DenseNet121**, a novel deep learning architecture that successfully balances high accuracy with computational efficiency. The study demonstrates that by combining a parameter-efficient CNN backbone with a lightweight attention module, it is possible to outperform larger, more complex models, providing a practical solution for real-world deployment.

3. **A State-of-the-Art Ensemble Framework:** This work introduces a novel framework that achieves state-of-the-art accuracy on the CREMA-D benchmark. This was accomplished by combining two key innovations: the re-formulation of the classification task using a psychologically-grounded dimensional model (the GWE control dimension) and the application of a strategic, architecturally diverse **ensemble of CNNs**.

4. **Comprehensive Benchmarking on Diverse Tasks and Datasets:** This thesis provides a broad and detailed empirical evaluation of ASER techniques. It establishes strong baselines for a new, challenging spontaneous dataset using traditional machine learning

and, in parallel, pushes the state-of-the-art on a standard benchmark dataset using two distinct and novel deep learning approaches.

Thesis Outline

The remainder of this thesis is organized into four chapters and a general conclusion:

Chapter 1: Background and Literature Review. This chapter lays the foundational groundwork for the thesis. It provides a comprehensive review of emotional models, key datasets, and the evolution of feature representation paradigms in ASER, from handcrafted features to learned spectrograms.

Chapter 2: Methodological Framework. This chapter details the complete methodological arsenal employed throughout this thesis. It provides a deep, pedagogical dive into the foundational concepts of machine learning and deep learning, the specific architectures developed (including CNNs, attention, and ensembles), and the rigorous framework for model evaluation.

Chapter 3: Experiments on the CREMA-D Benchmark Dataset. This chapter presents the first set of experimental results, focusing on the development and validation of novel deep learning models on a standard benchmark. It is organized into two distinct studies: one focusing on the efficient CBAM-DenseNet121 model and the other on the high-accuracy ensemble framework.

Chapter 4: Experiments on the Novel OYH Algerian Arabic Corpus. This chapter directly addresses the challenges of data scarcity and real-world spontaneity. It provides a detailed account of the creation and annotation of the OYH dataset and presents the results of a comprehensive baseline study using traditional machine learning models.

General Conclusion. The final section concludes the thesis. It provides a holistic discussion synthesizing the findings from all experimental chapters, summarizes the key contributions, acknowledges the limitations of the research, and proposes several promising directions for future work.

CHAPTER 1

BACKGROUND: DATA AND ITS REPRESENTATION IN ASER

Contents

1.1 Introduction

This chapter establishes the foundational knowledge required to understand the research presented in this thesis. The field of Automatic Speech Emotion Recognition (ASER) is situated at the intersection of digital signal processing, machine learning, and affective science. A thorough understanding of its core principles, historical context, and the materials used for research is therefore essential before delving into novel methodologies. This review aims to provide a comprehensive overview of the theoretical models, datasets, and feature representation paradigms that constitute the state-of-the-art in ASER.

The chapter is structured to build a logical progression from the general problem domain to the specific technical details of data preparation. It begins by introducing the broader field of Affective Computing and contextualizing the specific importance and challenges of ASER. Following this, the chapter delves into the theoretical frameworks that define how emotions are modeled, contrasting the classical categorical approaches with the more nuanced dimensional models that are central to this thesis. We will then review the key datasets that have shaped the field, highlighting the critical gap in resources for non-English and spontaneous speech, particularly for the Arabic dialects of North Africa. The subsequent section provides a detailed technical review of the two major paradigms for feature extraction from audio: the traditional, engineered approach using handcrafted acoustic features, and the modern, learned approach using log-Mel spectrograms.

Finally, the chapter concludes by synthesizing this information to clearly identify the key research gaps in the literature related to data, its representation, and the need for new resources. This discussion will directly motivate the novel methodologies and experimental studies presented in the subsequent chapters of this thesis.

1.2 The Rise of Affective Computing and ASER

Emotions are a fundamental component of the human experience, profoundly influencing our daily interactions, cognitive processes, perceptions, and overall mental well-being [11]. The ability to perceive and interpret emotions in others is a cornerstone of social intelligence, enabling empathy, cooperation, and complex social bonding. As technology becomes increasingly interwoven with the fabric of modern society—from ubiquitous smartphones and voice assistants to sophisticated IoT ecosystems—there is a growing imperative for these computational systems to develop their own form of social intelligence. This need has given rise to the vibrant and rapidly expanding field of **affective computing**, which aims to bridge the profound gap

between human emotion and computational technology [1]. The ultimate goal of this field is not merely to build functional machines, but to create systems capable of recognizing, interpreting, processing, and even simulating human affects, thereby facilitating more natural, empathetic, and effective human-computer interaction (HCI) [2, 3].

Within this broad domain, **Automatic Speech Emotion Recognition (ASER)** has distinguished itself as a critical and highly active area of research [4, 5]. Speech is arguably the most direct, natural, and information-rich channel for human emotional expression. Unlike written text, which conveys only semantic content, speech is layered with a wealth of **paralinguistic information** that reveals *how* something is said. This information is encoded in a variety of acoustic cues. By analyzing properties such as **prosody** (the melody and rhythm of speech, including intonation patterns), **voice quality** (the timbre and stability of the voice), and **spectral characteristics** (the distribution of energy across different frequencies), ASER systems aim to decode the underlying emotional state of a speaker [12, 13].

The potential applications of robust ASER systems are transformative and span numerous sectors. In the domain of **healthcare**, ASER promises to revolutionize diagnostics and patient monitoring. For instance, by detecting subtle vocal biomarkers such as flat affect (monotonous speech), vocal tension, or psychomotor retardation (slowed speech rate), systems could provide early warnings for mental health conditions like depression and anxiety, or track the progression of neurological diseases like Parkinson's [6]. For individuals with autism spectrum disorder, ASER-powered tools could offer real-time feedback to help them better understand the emotional subtext in social conversations. In the **commercial sector**, particularly in customer service, ASER can analyze a caller's sentiment in real-time. By flagging calls with high levels of frustration or anger, companies can intelligently route issues to specialized human agents, improving customer satisfaction and retention. In **education**, an ASER system could gauge a student's engagement, confusion, or frustration from their vocal responses, allowing an intelligent tutoring system to adapt its teaching strategy in real-time. Furthermore, in pioneering fields like **social robotics and virtual agents**, ASER is indispensable. For a robot or virtual assistant to be a truly effective and accepted companion, it must be able to perceive and appropriately respond to its user's emotional state, moving interactions from purely transactional to genuinely relational [7, 8].

Despite this immense potential, ASER remains a formidable scientific challenge. The acoustic manifestation of emotion is a highly complex and subtle phenomenon. It is subject to significant **inter-speaker variability**, influenced by physiological factors like age and gender, as well as idiosyncratic traits like personality and baseline speaking style [9]. Moreover, **cultural background** plays a crucial role, as the norms for expressing certain emotions can

vary significantly across different societies. This inherent complexity is further compounded by **intra-speaker variability**; the same person may express the same emotion differently depending on the context. Finally, the transition from controlled laboratory environments to real-world applications introduces a host of environmental challenges. Factors such as **background noise**, **reverberation** in a room, and variations in microphone quality can all introduce artifacts that mask or distort the crucial emotional cues within the speech signal, making robust recognition a persistent obstacle [14]. This thesis confronts these core challenges by proposing and evaluating novel deep learning frameworks, exploring alternative emotional models, and developing new linguistic resources designed to push the boundaries of ASER in both benchmark and real-world conditions.

1.3 Theoretical Frameworks: Emotional Models and Datasets

Before a system can recognize an emotion, a fundamental question must be answered: what is an emotion, and how should it be represented? The choice of an emotional model is a foundational step that dictates the nature of the classification task and has significant implications for data annotation, model design, and performance evaluation.

1.3.1 From Categorical to Dimensional Emotion Models

Historically, much of the research in ASER has been guided by two competing yet complementary psychological frameworks for conceptualizing emotion: categorical and dimensional models. The choice between these frameworks is not merely a theoretical preference; it is a foundational decision that dictates the nature of the data annotation process, the structure of the classification task, and the very interpretation of a model's output.

1.3.1.1 The Categorical Approach: Basic Emotions Theory

The categorical approach, which has long been the dominant paradigm in ASER, is heavily influenced by the cross-cultural research of psychologist Paul Ekman and his colleagues [12]. Based on groundbreaking studies in the 1960s and 70s involving the analysis of facial expressions across different cultures (including the Fore people of Papua New Guinea, who had minimal exposure to the outside world), Ekman posited the existence of a small set of discrete, universal, and biologically innate "basic" emotions. This theory suggests that these core emotions are recognized and expressed similarly by all humans, regardless of language

or culture. The most commonly cited set in ASER includes **anger, happiness, sadness, fear, disgust**, and a non-emotional **'neutral'** state [15].

The adoption of this framework by the ASER community was driven by its practical advantages. Firstly, its labels are intuitive and align with everyday human language, which simplifies the process of data annotation where annotators assign one label from a fixed set. Secondly, it frames the complex problem of emotion recognition as a standard multi-class classification problem, which is a well-understood problem in machine learning and maps directly to standard algorithms like Support Vector Machines or the softmax output layer of a neural network.

However, the elegance and simplicity of the categorical approach belies the complexity of human affect. Its primary limitation is its potential for **oversimplification**. Real-world emotional experiences are rarely pure; they are often blended, subtle, or exist on a continuum. This leads to several significant challenges:

- **The "Affect Blend" Problem:** Humans frequently experience multiple emotions simultaneously, such as the bittersweet feeling of nostalgia (a blend of happiness and sadness). A strict categorical model cannot represent these mixed states, forcing an artificial choice that loses crucial information [9].
- **Lack of Intensity Representation:** The model does not inherently account for emotional intensity. It treats ‘irritation‘, ‘anger‘, and ‘rage‘ as the single category of "Anger," despite their vastly different levels of arousal and acoustic manifestations.
- **Annotation Ambiguity:** The "forced-choice" nature of the annotation process can lead to significant ambiguity and low inter-annotator agreement. When an expression is ambiguous between, for example, ‘fear‘ and ‘surprise‘, different annotators may select different labels, introducing noise and inconsistency into the ground-truth data.
- **Cultural Display Rules:** While the recognition of basic emotions may be universal, the social norms or "display rules" governing their expression are heavily culturally modulated. This can affect the way emotions are vocalized, posing a challenge for models trained on one cultural group and applied to another.

1.3.1.2 The Dimensional Approach: Mapping the Affective Space

To overcome the rigidity of the categorical model, many researchers have advocated for **dimensional models**. These models do not treat emotions as independent classes but rather as

points within a continuous multi-dimensional space, aiming to capture the underlying structure and relationships of emotional experience [16, 17].

The most influential of these is James A. Russell's **circumplex model of affect** [18]. This model maps emotional states onto a two-dimensional circular space defined by two orthogonal axes:

- **Valence:** The hedonic tone of the emotion, representing the pleasure-displeasure continuum. It ranges from highly positive (e.g., 'ecstatic') to highly negative (e.g., 'miserable').
- **Arousal:** The level of physiological and psychological activation, ranging from low (e.g., 'sleepy', 'calm') to high (e.g., 'excited', 'frenzied').

Within this space, any emotion can be represented as a point. For instance, 'anger' is a high-arousal, negative-valence emotion, while 'sadness' is a low-arousal, negative-valence emotion. This approach is powerful because it can represent emotional intensity (distance from the origin), similarity between emotions (proximity in the space), and transitions between states.

Building on this, other models introduced a third dimension. The **Pleasure-Arousal-Dominance (PAD) model**, for example, added the dimension of **Dominance** (or Control) to represent the sense of power or control the individual feels in a situation [19]. This brings us to the more recent and nuanced framework used in this thesis: the **Geneva Wheel of Emotions (GWE)**, shown in Figure 1.1. The GWE can be seen as a sophisticated hybrid model that organizes 20 distinct emotion families (providing intuitive categorical labels) within a two-dimensional space defined by **Valence** and **Control/Power** [20, 21].

The **Control** dimension is particularly relevant for ASER, as it has strong correlates in vocal production. A feeling of high control or agency is often associated with a more assertive and stable vocal pattern—a steady pitch, strong intensity, and clear articulation. Conversely, a feeling of low control or helplessness can manifest acoustically as a more hesitant, unstable, or weak voice—for instance, with a trembling pitch (high jitter), a breathy quality (low HNR), or a lower intensity. In this thesis, we leverage this strong theoretical link by using the control dimension as a novel classification strategy, grouping emotions into **High Control** (anger, happiness, disgust) and **Low Control** (fear, sadness, neutrality). The central hypothesis, supported by recent findings, is that such psychologically grounded dimensional groupings can lead to more stable and robust classification by clustering emotions that share more consistent underlying acoustic correlates than purely categorical labels might [22].

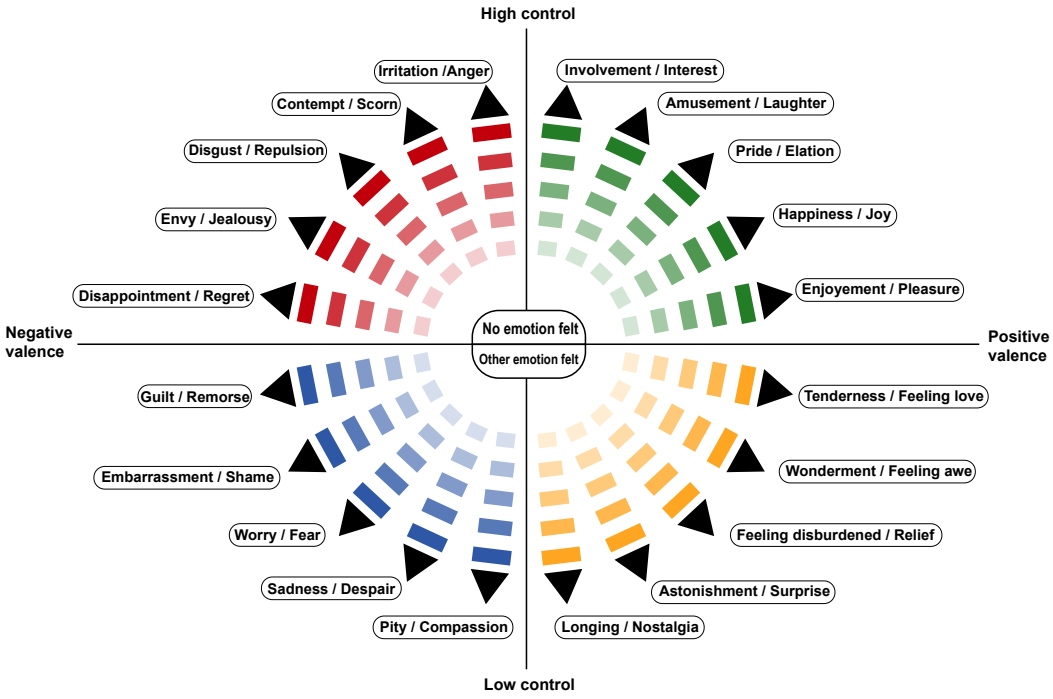


Figure 1.1: The Geneva Wheel of Emotions (GWE), illustrating the two primary dimensions of valence and control. The model organizes 40 emotion words into 20 distinct emotion families, providing a nuanced dimensional framework for affect.[21].

1.3.2 Corpora for Speech Emotion Recognition

The development and validation of ASER systems are fundamentally dependent on the availability of high-quality, annotated speech corpora. The nature of the dataset—particularly whether the speech is acted or spontaneous—profoundly impacts a model's real-world performance.

The field has matured thanks to several benchmark English-language datasets, which primarily feature acted emotions:

- **IEMOCAP (Interactive Emotional Dyadic Motion Capture Database):** One of the most widely used datasets in the community, IEMOCAP features approximately 12 hours of audiovisual data. It contains scripted readings but is most famous for its improvisational scenarios between pairs of actors, designed to elicit more natural emotional expressions. The data is annotated with both categorical labels and dimensional ratings (valence, arousal, dominance), making it invaluable for studying complex emotional phenomena [23].

- **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):** As a primary dataset for this thesis, CREMA-D offers 7,442 audio-visual clips from a large and ethnically diverse cohort of 91 actors. Actors uttered 12 specific sentences in six different emotional styles. The high-quality, noise-free recordings and the reliability of its labels, validated through a large-scale crowd-sourcing effort, make it an ideal benchmark for developing and comparing models under controlled conditions [24].
- **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):** This dataset is notable for its systematic design. It provides over 7,300 recordings from 24 actors expressing eight emotions at two distinct levels of intensity (normal, strong) and across two modalities (speech and song). This controlled variation is excellent for studying the nuances of how emotional intensity is encoded in the voice [25].

While these corpora have been instrumental, their predominance highlights a critical limitation in the field: a significant bias towards English and **acted, laboratory-controlled speech**. Acted data, while clean and emotionally balanced, may not accurately reflect the subtlety and messiness of genuine, spontaneous emotions. This creates a significant gap between model performance on benchmarks and their effectiveness in real-world applications. This gap is exacerbated by the lack of resources for the vast majority of the world's languages and dialects.

This issue is especially acute for Arabic. Despite being a major world language, research in Arabic ASER has been historically hindered by a lack of suitable corpora [26]. While important efforts have been made to create datasets for Modern Standard Arabic (MSA) and various national dialects (summarized in Table 1.1), these are often small in scale or, like their English counterparts, rely on acted speech. Furthermore, the significant linguistic diversity within the Arabic-speaking world means that a model trained on one dialect (e.g., from the Gulf) is unlikely to generalize well to another (e.g., from North Africa).

The Algerian linguistic landscape, as shown in Figure 1.2, is a prime example of this complexity. Daily life is dominated by spoken dialects (collectively known as "Darija"), which are themselves diverse and heavily influenced by Berber languages and French [27]. This creates a situation where models trained on MSA are poorly suited for practical applications. This critical lack of a large-scale, **spontaneous** speech corpus for the Algerian dialect served as a primary motivation for one of the core contributions of this thesis: the development of the "Open Your Heart" (OYH) dataset, a corpus collected from a real-world talk show to capture genuine emotional speech [10].

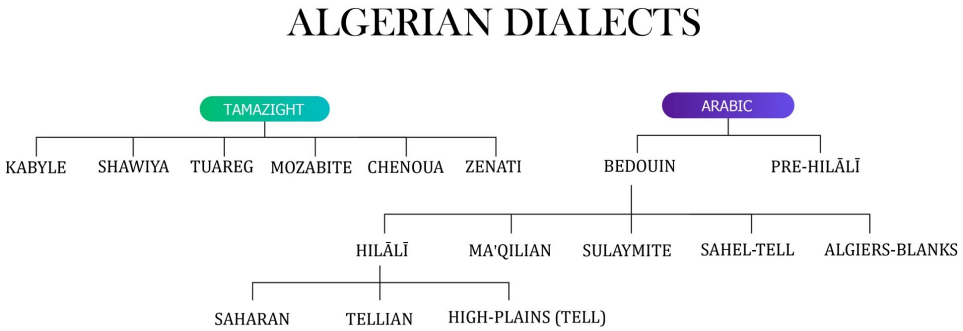


Figure 1.2: Map illustrating the rich diversity of Algerian dialects, highlighting the complex linguistic landscape that necessitates dialect-specific resources. Models trained on MSA or other Arabic dialects often fail to capture these local variations. (Source:[27]).

Table 1.1: An overview of major speech emotion databases (based on Table C1 from the OYH dataset paper).

Dataset	**Language**	**Type**	**Emotions Covered**	**Classification Methods Used**
IEMOCAP [23]	English	Actor-based	Happy, Sad, Anger, Neutral, etc.	Various
VAM Audio [28]	German	Spontaneous	Valence, Activation, Dominance	Various
CREMA-D [24]	English	Actor-based	Anger, Disgust, Fear, Happy, Sad, Neutral	Various (CNN, Transformers)
RAVDESS [25]	English	Actor-based	8 emotions (Calm, Happy, etc.)	Various
MELD [29]	English	Spontaneous (TV)	Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise	Multimodal Baselines
ShEMO [30]	Persian	Semi-spontaneous	Anger, Fear, Joy, Sadness, Surprise, Neutral	SVM, k-NN, Decision Tree

Dataset	Language	Type	Emotions Covered	Classification Methods Used
KEDAS [31]	MSA	Actor-based	Sadness, Fear, Anger, Happiness, Neutral	/
REGIM-TES [32]	Tunisian Arabic	Actor-based	Neutral, Sadness, Fear, Anger, Happiness	Naive Bayes, SVM
Emirati Speech DB [33]	Emirati Arabic	Spontaneous	Happiness, Sadness, Disgust, Anger, Fear	GMM-DNN, SVM, MLP
ADED [34]	Algerian Arabic	Actor-based	Anger, Fear, Sadness, Neutral	k-NN

1.4 From Signal to Representation: A Review of Feature Extraction in ASER

The process of converting a raw, continuous audio waveform into a structured format that a machine learning model can process is known as feature extraction. This is a critical stage that determines what information from the signal is preserved for classification, and the field has seen a major paradigm shift from handcrafted, engineered features to automatically learned representations.

1.4.1 Handcrafted Acoustic Features

The traditional approach to ASER, which dominated the field for decades, is predicated on a feature engineering paradigm. This methodology involves the meticulous design and extraction of a wide range of acoustic features believed to be correlated with emotional expression. These features, often called Low-Level Descriptors (LLDs), are quantitative measures that capture specific physical properties of the speech signal from short analysis frames (typically 20-40 ms in length). To create a single feature vector for an entire utterance, statistical functionals (e.g., mean, standard deviation, minimum, maximum, percentiles, regression coefficients) are

applied to the contours of these LLDs over time, resulting in a set of High-Level Statistical Features (HSFs).

This complex extraction process is typically automated using specialized toolkits. The most prominent among these is the **openSMILE (open-source Speech and Music Interpretation by Large-space Extraction)** toolkit [35], which provides standardized and reproducible feature sets. For instance, the ComParE (Computational Paralinguistics Challenge) feature set, used in one of the studies in this thesis, contains 6,373 HSFs, demonstrating the high dimensionality inherent in this approach [36]. These vast feature sets are typically organized into three principal categories, grounded in the principles of acoustic phonetics: prosodic, spectral, and voice quality features [37].

1.4.1.1 Prosodic Features

Prosodic features, often described as the "melody" of speech, relate to the suprasegmental aspects of an utterance, including its pitch, loudness, and rhythm. They are considered among the most powerful indicators of emotional state.

- **Fundamental Frequency (F0):** The F0 is derived from the rate of vibration of the vocal folds and is the primary acoustic correlate of perceived **pitch**. Its contour over an utterance carries immense emotional weight. Key metrics derived from the F0 contour include its mean (reflecting the speaker's baseline pitch), standard deviation (reflecting pitch range or vocal animation), and overall shape. For example, high-arousal emotions like **Anger** or **Joy** are often characterized by a high mean F0 and a wide F0 range, indicating an excited state. Conversely, low-arousal emotions like **Sadness** typically exhibit a low mean F0, a narrow, monotonous pitch range, and a slowly declining contour toward the end of phrases. The F0 in **Fear** can be exceptionally high, often with abrupt, uncontrolled jumps.

- **Energy and Intensity:** The energy of a speech signal, typically measured as the Root-Mean-Square (RMS) of the amplitude within a frame, corresponds to its perceived **loudness**. Like F0, the energy contour and its statistical properties are strong emotional cues. **Anger** is a classic high-energy emotion, characterized by sharp increases in intensity. **Joy** is also associated with high energy, though often with more dynamic variation than anger. **Sadness**, in contrast, is marked by low and decaying energy levels.

- **Duration and Speech Rate:** The temporal characteristics of speech are also highly modulated by emotion. This includes the overall **speech rate** (e.g., syllables per second),

the relative duration of voiced and unvoiced segments, and the length and frequency of pauses. A fast speech rate is a common indicator of **Joy** or **Anger**, while a significantly slowed speech rate is a hallmark of **Sadness**. The duration of pauses can also be revealing; for instance, longer and more frequent pauses might indicate the hesitation associated with **Sadness** or the tension before an outburst of **Anger**.

1.4.1.2 Spectral Features

Spectral features describe the frequency content of the speech signal, which is primarily shaped by the resonant properties of the vocal tract (i.e., the throat, mouth, and nasal cavities). While heavily tied to the phonetic content of what is being said, the spectral characteristics are also subtly modulated by emotion.

- **Mel-Frequency Cepstral Coefficients (MFCCs):** For decades, MFCCs have been the most dominant spectral features in both speech and speaker recognition, and they have been widely adopted in ASER. They provide a compact representation of the spectral envelope (the overall shape of the spectrum). Their calculation involves several steps inspired by human auditory perception: the Fast Fourier Transform (FFT) of a windowed frame is mapped onto the non-linear Mel frequency scale, its logarithm is taken, and finally, a Discrete Cosine Transform (DCT) is applied. The DCT decorrelates the spectral components, making them more suitable for statistical modeling [38]. While MFCCs primarily encode phonetic information, emotional states can alter articulation—for example, the tense articulation in **Anger** can shift spectral energy towards higher frequencies, which is captured by the MFCCs.

- **Formants:** These are the resonant peaks in the frequency spectrum, which are determined by the shape of the vocal tract and are fundamental to distinguishing vowels. The position and bandwidth of the first few formants (F1, F2, F3) can be influenced by the muscular tension in the vocal tract associated with different emotions.

- **Other Spectral Descriptors:** A variety of other features, such as the spectral centroid (the "center of mass" of the spectrum), spectral flux (the rate of change of the spectrum), and spectral roll-off, can also provide useful information about the timbre and brightness of the voice, which in turn relate to emotional expression.

1.4.1.3 Voice Quality Features

Voice quality features describe the characteristics of the glottal source—the sound produced by the vibrating vocal folds—and relate to perceptions of breathiness, roughness, or strain.

- **Jitter and Shimmer:** These are micro-prosodic perturbation measures. **Jitter** refers to the cycle-to-cycle variation in the fundamental frequency, while **Shimmer** refers to the cycle-to-cycle variation in amplitude. In a perfectly stable voice, these values would be near zero. Increased jitter and shimmer indicate instability in vocal fold vibration, which can be a sign of vocal tension or stress, often present in emotions like **Anger** or **Fear**. A creaky or harsh voice quality, sometimes associated with **Sadness**, can also result in high perturbation values.
- **Harmonic-to-Noise Ratio (HNR):** The HNR quantifies the ratio of periodic (harmonic) energy to aperiodic (noise) energy in the voice signal. A high HNR corresponds to a clear, resonant voice, while a low HNR corresponds to a noisy or breathy voice. Breathiness, a classic acoustic correlate of low-arousal states, is directly captured by a low HNR, making it a powerful feature for identifying **Sadness** or some expressions of **Fear**.

While this feature engineering paradigm is foundational and provides interpretable, phonetically grounded descriptors, it has significant limitations. The "curse of dimensionality" from extracting thousands of features necessitates complex and computationally expensive feature selection steps. More importantly, this approach struggles with the "semantic gap"—the fact that these low-level physical descriptors often have a complex, non-linear, and context-dependent relationship with high-level emotional concepts. It was precisely these limitations that motivated the field's shift towards the end-to-end learned representations discussed in the next section.

1.4.2 Learned Representations: Log-Mel Spectrograms

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), revolutionized feature extraction in ASER. It enabled a paradigm shift away from handcrafted features towards end-to-end models that learn the optimal representations directly from a semi-raw, visually intuitive format: the spectrogram. The key innovation was the adoption of the spectrogram as a standard input, which effectively transforms the audio analysis problem into an image recognition task. The generation pipeline, illustrated in Figure 1.3, is a multi-stage signal processing workflow designed to convert a 1D audio waveform into a 2D representation that is both information-rich and perceptually relevant.

1.4.2.1 Pre-Processing: Pre-Emphasis

Before the main transformation, a pre-emphasis filtering step is often applied to the raw audio signal, $x[n]$. The primary purpose of this step is to balance the frequency spectrum. Speech signals naturally have more energy at lower frequencies than at higher frequencies (a spectral tilt of approximately -6 dB/octave). Pre-emphasis is a high-pass filter that boosts the energy in the higher frequencies, which helps to improve the numerical stability of the subsequent Fourier Transform and can enhance the representation of important high-frequency details like fricatives. The filter is defined by the first-order difference equation:

$$y[n] = x[n] - \alpha x[n-1] \tag{1.1}$$

where $y[n]$ is the filtered signal and α is the pre-emphasis coefficient, typically set to a value between 0.95 and 0.98 [39].

1.4.2.2 Framing and Windowing

Speech signals are non-stationary, meaning their statistical properties change over time. However, over very short durations (e.g., 20-40 ms), they can be considered quasi-stationary. The process of **framing** involves segmenting the signal into these short, stationary frames. To avoid information loss at the boundaries of each frame, the frames are typically designed to overlap, with a common overlap being 50-75%.

A raw rectangular frame would result in sharp discontinuities at its edges, which introduces spurious high-frequency components in the frequency domain, an artifact known as spectral leakage. To mitigate this, each frame is multiplied by a **window function** that tapers the signal to zero or near-zero at its boundaries. A common choice, and the one used in this work, is the Hamming window, defined as:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right), \quad 0 \le n \le L-1 \tag{1.2}$$

where L is the length of the window (frame size).

1.4.2.3 Short-Time Fourier Transform (STFT) and the Power Spectrogram

With the signal segmented into windowed frames, the **Short-Time Fourier Transform (STFT)** is applied to each frame to determine its frequency content. The STFT computes the Discrete Fourier Transform (DFT) for each frame, yielding a series of complex-valued spectra over time. The discrete STFT is defined as:

$$\text{STFT}(m,k) = \sum_{n=0}^{L-1} y_m[n] \cdot e^{-j\frac{2\pi}{N}kn} \tag{1.3}$$

where $y_m[n]$ is the m-th windowed frame, N is the FFT size (number of frequency bins), and $\text{STFT}(m, k)$ is the resulting complex coefficient for the k-th frequency bin [40].

Since the phase information from the complex STFT output is often discarded in ASER as it is highly sensitive to noise, the analysis proceeds using the magnitude or power. The **power spectrogram** is computed by taking the squared magnitude of the STFT result for each time-frequency point:

$$P(m, k) = |\text{STFT}(m, k)|^2 \tag{1.4}$$

This results in a 2D matrix where the value at each point $P(m, k)$ represents the energy of the signal at a specific time m and frequency k.

1.4.2.4 Mel Scale Transformation

The linear frequency scale of the power spectrogram does not align well with human auditory perception. The human ear is far more sensitive to changes in low frequencies than in high frequencies. To model this, the linear frequency axis is warped onto the **Mel scale**, which is linear below 1 kHz and logarithmic above. The conversion is given by:

$$\text{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{1.5}$$

where f is the frequency in Hertz. This transformation is implemented using a **Mel filterbank**, which is a set of triangular band-pass filters (typically 20-128 filters) spaced according to the Mel scale. Each filter in the bank is multiplied by the power spectrum of a frame and the results are summed to obtain the energy in that specific Mel frequency band. This process effectively simulates the frequency resolution of the human cochlea and has been shown to be highly effective for both speech recognition and emotion recognition tasks [41].

1.4.2.5 Logarithmic Compression and Final Representation

The final step is to apply a logarithmic compression to the amplitudes of the Mel spectrogram. This serves two critical purposes:

1. It compresses the dynamic range of the values, making the features less sensitive to variations in signal energy (i.e., how loudly someone is speaking).

2. It aligns the amplitude representation with the human perception of loudness, which follows the decibel (dB) scale, a logarithmic unit.

The final **log-Mel spectrogram** is computed as follows:

$$S_{\text{log-mel}}(m,l) = \log\left(\sum_{k=0}^{N/2-1} P(m,k)\cdot H_l(k)\right) \tag{1.6}$$

where $H_l(k)$ represents the weight of the k-th frequency bin for the l-th Mel filter.

The significance of this entire transformation cannot be overstated. By converting a 1D time-series signal into a 2D "image," it became possible to leverage the immense power of **Convolutional Neural Networks (CNNs)**. A CNN can "look" at a spectrogram and automatically learn a hierarchical set of features—from simple, localized patterns like horizontal lines (stable tones), vertical lines (plosives), and basic formant shapes in the early layers, to complex, time-varying harmonic structures indicative of specific emotions in deeper layers. This end-to-end learning approach removes the need for manual feature engineering, bridges the "semantic gap," and has been shown to be a superior method for capturing the rich, nuanced information required for robust emotion recognition [42].

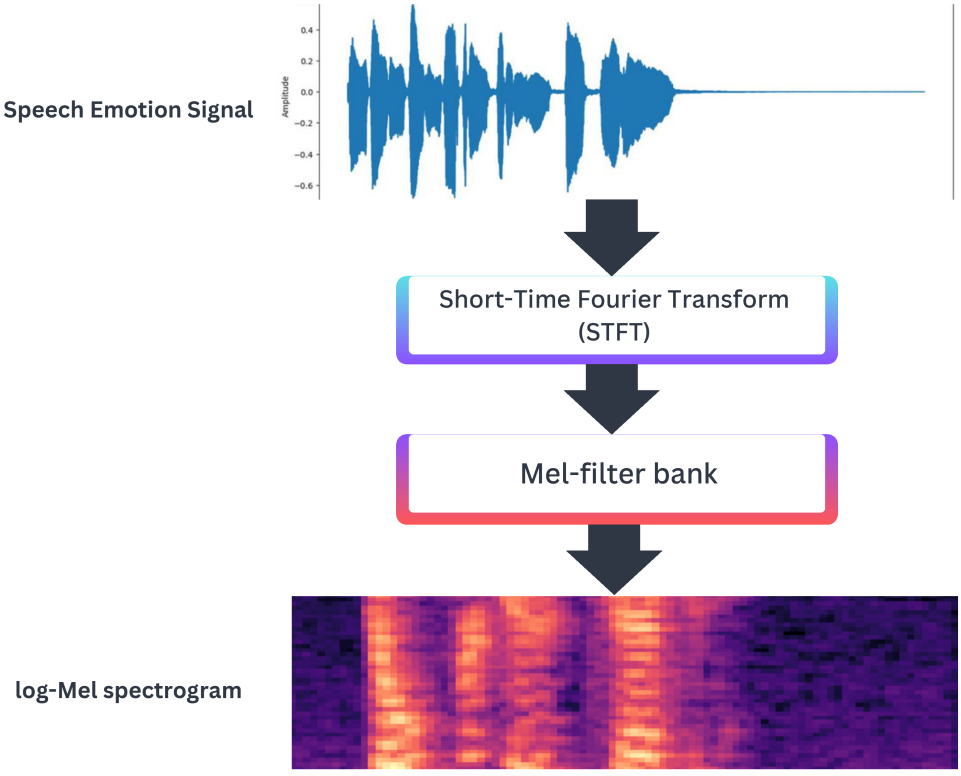


Figure 1.3: The log-Mel spectrogram feature extraction pipeline, converting a 1D audio waveform into a 2D representation suitable for deep learning models. (Source: [40]).

1.5 Research Gap and Thesis Positioning

This comprehensive review of the emotional models, datasets, and feature representations used in ASER reveals a field rich with opportunity but also defined by significant challenges. The groundwork laid by existing resources and techniques allows this thesis to be specifically positioned to address the following critical research gaps:

1. **The Need for Advanced Models to Interpret Rich Features:** While spectrograms provide a rich, high-dimensional representation of speech, they require sophisticated models to effectively learn the subtle, hierarchical patterns that encode emotion. The limitations of traditional models when applied to such complex data motivate the exploration of advanced deep learning architectures, which will be the central topic of the next chapter. There is a clear need for models that can not only achieve high accuracy but also operate efficiently, and for frameworks like ensemble learning that can maximize robustness.

2. **The Gap Between Lab-Based and Real-World Data:** The vast majority of high-quality corpora consist of acted speech recorded in sterile lab environments. This creates a significant "domain gap" between the data used for training models and the noisy, spontaneous speech encountered in real-world applications. This gap severely limits the generalizability of current state-of-the-art models.

3. **The Critical Scarcity of Low-Resource and Dialectal Corpora:** The focus on English and other high-resource languages has left many linguistic communities, including speakers of Arabic dialects like Algerian, underserved. Without appropriate in-domain data that captures local linguistic and cultural nuances, developing effective and fair ASER technology for these populations is impossible.

In summary, the field has established powerful methods for representing emotional speech but requires new, robust, and efficient models to analyze these representations. Furthermore, it desperately needs new, diverse, and realistic datasets to ensure that these models are not just accurate in the lab but are also effective and equitable in the real world. This thesis directly confronts these challenges. The following chapter will detail the specific classification models and experimental protocols designed to process the feature representations discussed here, applied to both a standard benchmark dataset and a novel, spontaneous dialectal corpus created to address the critical resource gap.

1.6 Conclusion

This chapter has laid the essential groundwork for the research presented in this thesis by providing a comprehensive review of the foundational elements of Automatic Speech Emotion Recognition. We began by contextualizing ASER within the broader field of affective computing, establishing its importance, applications, and the significant scientific challenges that motivate this work. We then delved into the theoretical frameworks that define how emotions are modeled, contrasting the traditional **categorical approach** with the more nuanced **dimensional models**, such as the Geneva Wheel of Emotions, which are central to the novel classification schemes explored in this research.

Furthermore, a critical review of the key datasets that have shaped the field highlighted a significant gap in resources, particularly the lack of large-scale, **spontaneous speech corpora** for under-resourced languages like the Arabic dialects of North Africa. Finally, the chapter examined the evolution of feature representation, moving from the interpretable but limited **handcrafted acoustic features** to the rich, high-dimensional data provided by **log-Mel spectrograms**.

The review has made it clear that while powerful representations like spectrograms exist, they require advanced computational models to effectively interpret the subtle, hierarchical patterns that encode emotion. Having established this foundation of data and its representation, the subsequent chapter will now introduce the specific computational methodologies—from traditional machine learning to advanced deep learning architectures—that will be employed to analyze this data and address the research gaps identified herein.

CHAPTER 2

METHODOLOGICAL FRAMEWORK FOR SPEECH EMOTION RECOGNITION

Contents

2.1 Introduction

This chapter details the core technical methodologies employed in this thesis to address the research gaps identified in the previous chapter. Having established the foundational concepts of data representation in ASER, this chapter now focuses on the models and frameworks used to classify those representations. The purpose of this chapter is to provide a comprehensive and self-contained guide to the algorithmic principles, architectural families, and specific models that underpin the experimental work of this thesis.

The chapter is structured to build from the general to the specific. It begins with an overview of the foundational principles of machine learning and deep learning, establishing the theoretical context for the models used. It then transitions into a deep and detailed examination of the specific architectures developed and employed in this thesis. This includes a thorough breakdown of the custom **DeepSpecCNN**, the **DenseNet-121** backbone, the **CBAM** attention module, and the final proposed **CBAM-DenseNet121** model. The principles of **ensemble learning** are also detailed. The chapter concludes by establishing the rigorous experimental framework for the subsequent chapters, formally defining the datasets, classification tasks, and the full suite of evaluation metrics used to assess model performance.

2.2 Foundational Concepts in Machine Learning

Machine Learning (ML) is a field of artificial intelligence in which algorithms and systems are designed to learn patterns and make predictions directly from data, without being explicitly programmed for each task [43]. The core principle of ML is to develop models that can learn a mapping function, f, from an input space, X, to an output space, Y, by optimizing their internal parameters based on the data it is exposed to. The field can be broadly categorized into several learning paradigms, defined by the nature of the available data and the learning objective [44].

2.2.1 Supervised vs. Unsupervised Learning

The vast landscape of machine learning is largely defined by the type of data and the problem to be solved. The two most fundamental paradigms are supervised and unsupervised learning.

2.2.1.1 Supervised Learning

This is the most common paradigm in machine learning and forms the basis for all classification experiments in this thesis. In supervised learning, the algorithm learns from a dataset

where each input sample $x \in X$ is paired with a correct, ground-truth output label $y \in Y$. The presence of these labels provides the "supervision" that guides the learning process. The model's objective is to learn a generalizable function $f(x) \approx y$ that can accurately predict the output for new, unseen inputs. This is achieved by iteratively adjusting the model's internal parameters to minimize the difference (or "error") between its predictions and the true labels. Supervised learning encompasses two primary task types:

- **Classification:** The goal is to predict a discrete class label from a finite set of categories. The output space Y is categorical. In the context of this thesis, this corresponds directly to the task of ASER, where the input is an audio feature vector and the output is an emotional category, such as {'Anger', 'Joy', 'Sadness', ...}. Classification can be binary (two classes), multi-class (more than two mutually exclusive classes), or multi-label (an instance can belong to multiple classes simultaneously, e.g., a speech sample could be both "happy" and "excited").

- **Regression:** The goal is to predict a continuous numerical value. The output space Y is composed of real-valued numbers. In ASER, this is relevant for dimensional emotion modeling, where the task could be to predict a valence score on a continuous scale from -1.0 to +1.0, or an arousal score from 1 to 9. The performance of regression models is typically measured using error metrics like Mean Squared Error (MSE).

2.2.1.2 Unsupervised Learning

In this paradigm, the model is provided with data that has no explicit labels. The objective is to discover inherent structures, patterns, or relationships within the data itself without any external guidance. Common unsupervised tasks include:

- **Clustering:** The task of grouping similar data points together into clusters, such that points within the same cluster are more similar to each other than to those in other clusters. For example, an unsupervised clustering algorithm like K-Means could be applied to thousands of unlabeled speech segments to see if natural acoustic-emotional groupings emerge from the data itself.

- **Dimensionality Reduction:** The process of compressing data into a lower-dimensional representation while preserving as much important information as possible. This is highly relevant for ASER when dealing with high-dimensional handcrafted feature sets (as described in Chapter 1), where techniques like Principal Component Analysis (PCA) can

be used to reduce a feature vector of several thousand dimensions to a more manageable size, combating the "curse of dimensionality" and improving computational efficiency.

2.2.2 Traditional Machine Learning Paradigms for ASER

Before the widespread adoption of deep learning, ASER research relied on a suite of traditional, or "classical," machine learning models. These models typically operate on handcrafted acoustic feature vectors and serve as powerful and interpretable baselines. The classifiers employed in the baseline study of this thesis are detailed below.

- **Support Vector Machines (SVM):** An SVM is a powerful discriminative classifier that operates by finding an optimal separating hyperplane between classes in a high-dimensional feature space [45]. For a two-class problem, the "optimal" hyperplane is the one that has the largest margin—the maximum distance to the nearest data points (the "support vectors") of any class. This maximization of the margin leads to better generalization. For data that is not linearly separable, SVMs employ the **kernel trick**. This technique uses a kernel function (such as the polynomial, sigmoid, or Radial Basis Function (RBF) kernel) to implicitly map the data into a much higher-dimensional space where a linear separation is possible, without ever having to explicitly compute the coordinates of the data in that space. The SVM is then controlled by hyperparameters like ‘C‘, which manages the trade-off between a smooth decision surface and classifying training points correctly, and ‘gamma‘, which defines the influence of a single training example in the RBF kernel.

- **K-Nearest Neighbors (KNN):** KNN is a simple, non-parametric instance-based learning algorithm. Unlike other models, it does not learn an explicit function during training; instead, it memorizes the entire training dataset. To classify a new data point, it finds the ’k’ closest training samples in the feature space using a distance metric (e.g., Euclidean distance). The new point is then assigned the class that is most common among its k neighbors. KNN is easy to interpret but can be computationally slow during inference and its performance can degrade significantly in high-dimensional feature spaces (the "curse of dimensionality"), where the concept of distance becomes less meaningful.

- **Decision Trees and Random Forest:** A **Decision Tree** is a non-parametric supervised learning method that predicts the class of a target variable by learning simple decision rules inferred from the data features. It builds a tree-like structure where internal nodes represent tests on features (e.g., "is mean F0 > 150 Hz?") and leaf nodes represent the

final class labels. While a single, deep decision tree is highly interpretable, it is prone to overfitting. A **Random Forest** is a powerful ensemble method that addresses this limitation by constructing a multitude of deep decision trees at training time [46]. It introduces randomness through two key mechanisms: **bagging** (each tree is trained on a random bootstrap sample of the data) and **feature randomness** (only a random subset of features is considered at each split). The final prediction is the mode of the classes predicted by the individual trees, which averages out their variances and results in a much more robust and accurate model.

- **Boosting Algorithms:** This family of ensemble algorithms builds a strong classifier by sequentially training a series of weak learners (typically shallow decision trees). Each new learner in the sequence focuses on correcting the errors made by the previous ones.
  - **AdaBoost (Adaptive Boosting):** The original boosting algorithm, AdaBoost iteratively trains weak learners and, at each step, increases the weights of the training samples that were misclassified by the previous learner. The final model is a weighted sum of all the weak learners, with more accurate learners receiving a higher weight [47].
  - **Gradient Boosting Machines (GBM):** A more generalized and powerful boosting framework. Instead of re-weighting samples, each new weak learner is trained to predict the pseudo-residuals (the gradient of the loss function) of the predecessor ensemble [48]. This makes it a direct gradient-based optimization algorithm.
  - **XGBoost, LightGBM, and CatBoost:** These are modern, highly optimized and scalable implementations of gradient boosting. **XGBoost** [49] is known for its performance and includes advanced regularization techniques. **LightGBM** [50] offers significant speed improvements using novel algorithms for building trees. **CatBoost** [51] is specialized for handling categorical features with high efficiency.

2.3 The Deep Learning Paradigm for ASER

Deep Learning (DL) is a subfield of machine learning based on artificial neural networks with multiple hidden layers (hence "deep"). The availability of vast datasets and powerful computational resources has allowed these deep architectures to achieve state-of-the-art performance on a wide range of tasks, including ASER [52]. The key advantage of deep learning is its ability to learn feature hierarchies automatically from the data, which is particularly beneficial for complex, high-dimensional inputs like audio spectrograms.

2.3.1 The Artificial Neuron: From Perceptron to Modern Activations

The simplest form of a neural network is the **perceptron**, a model of a single artificial neuron first proposed by Frank Rosenblatt in 1958 [53]. It was inspired by the biological neuron, which receives electrochemical signals through dendrites, integrates them in the cell body (soma), and, if a certain threshold is met, fires an output signal down the axon.

The artificial perceptron mimics this process mathematically. It takes a set of numerical inputs, $X = \{x_1, x_2, ..., x_n\}$, each multiplied by a corresponding **weight**, $W = \{w_1, w_2, ..., w_n\}$. These weights represent the "synaptic strength" of each input, determining its influence on the neuron's output. The weighted inputs are then summed together along with a **bias** term, b, which acts as an independent, learnable offset that can shift the activation function's decision boundary. This weighted sum, or pre-activation, S, is given by [54]:

$$S = \sum_{i=1}^{n} (x_i \cdot w_i) + b = W \cdot X + b \tag{2.1}$$

This sum is then passed through an **activation function**, f, to produce the final output, Y. The activation function introduces essential non-linearity into the model, allowing it to learn complex patterns. While early models used simple step functions, modern networks use differentiable functions like the Sigmoid, Hyperbolic Tangent (tanh), and, most commonly, the **Rectified Linear Unit (ReLU)**, defined as [55]:

$$f(S) = \max(0, S) \tag{2.2}$$

ReLU is computationally efficient and helps mitigate the vanishing gradient problem, making it the standard choice for most deep hidden layers today.

2.3.2 The Multilayer Perceptron (MLP) and Hierarchical Features

A single perceptron is limited to learning only linearly separable patterns. This was famously demonstrated by Minsky and Papert, who showed that it cannot solve the simple logical XOR problem [56]. To overcome this fundamental limitation, perceptrons are stacked together in layers to form a **Multilayer Perceptron (MLP)**.

An MLP consists of an input layer, one or more **hidden layers**, and an output layer. The key principle is that the presence of hidden layers, combined with non-linear activation functions, gives the MLP the power of a universal approximator, meaning it can theoretically approximate any continuous function [57]. This power comes from its ability to learn a **hierarchy of features**. For an ASER task, the first hidden layer might learn to combine raw acoustic features into basic phonetic detectors. A second hidden layer might combine these phonetic detectors to

recognize more complex syllabic or rhythmic patterns. Deeper layers can then combine these patterns to capture the high-level prosodic contours that are indicative of emotion.

2.3.3 Network Training: Backpropagation and Optimization

Neural networks learn by iteratively adjusting their weights and biases to minimize a **loss function**, which quantifies the error between the model's predictions and the true labels. For multi-class classification, the standard loss function is **Categorical Cross-Entropy**, which heavily penalizes confident but incorrect predictions. This minimization is achieved through an optimization algorithm, most commonly a variant of **gradient descent**.

The core mechanism is the **backpropagation algorithm** [58]. The process works as follows:

1. **Forward Pass:** A batch of inputs is fed through the network to generate predictions.
2. **Loss Calculation:** The loss function compares the predictions to the true labels to calculate a single error value.
3. **Backward Pass:** Using the chain rule from calculus, the algorithm computes the gradient (or derivative) of the loss function with respect to every single weight and bias in the network, starting from the output layer and moving backward. These gradients represent the direction and magnitude of the change needed for each parameter to reduce the error.
4. **Parameter Update:** The optimizer updates the parameters by taking a small step in the direction opposite to their gradient.

While standard gradient descent is effective, more advanced optimizers like **Adam (Adaptive Moment Estimation)** [59] are now standard. Adam adapts the learning rate for each parameter individually and incorporates momentum, leading to faster and more stable convergence.

2.3.4 Training Strategies

The process of teaching a deep learning model is known as training. The strategies employed can significantly impact performance.

- **Training from Scratch:** All network weights are initialized randomly and the model learns exclusively from the target dataset. This requires a very large dataset to avoid overfitting.
- **Transfer Learning & Fine-Tuning:** A more common approach is to use a model pre-trained on a massive dataset and adapt it to the target task. Recent work has demonstrated

the power of this approach by fine-tuning large, self-supervised models like WavLM on various speech and audio tasks, showing their ability to transfer learned representations effectively [60].

2.4 Architectures for Speech Emotion Recognition

This section provides a deep and detailed breakdown of the specific deep learning architectures and frameworks that were developed and utilized in the experimental work of this thesis.

2.4.1 Convolutional Neural Networks (CNNs) for Spectrogram Analysis

As established in Chapter 2, CNNs are exceptionally well-suited for analyzing spectrograms by treating them as images. The core components that enable this are the convolutional and pooling layers, which learn a hierarchy of time-frequency patterns. The specific models in this thesis build upon these foundational principles.

2.4.2 The DeepSpecCNN Architecture

For the study on dimensional emotion classification, a custom CNN architecture named **DeepSpecCNN** was designed specifically for the task of spectral analysis. As illustrated in Figure 2.1, the model is constructed for efficient hierarchical feature extraction.

The architecture is composed of four main convolutional blocks. Each block utilizes uniform 33 convolutions with "same" padding to preserve spatial dimensions during feature mapping, followed by a Rectified Linear Unit (ReLU) activation to introduce non-linearity. After each of the first three blocks, a max-pooling layer with a 22 kernel is applied for spatial down-sampling, progressively reducing the feature map dimensionality. After the final convolutional block, the feature maps are flattened into a 1D vector and passed to a dense layer with a softmax activation function to perform the final classification.

2.4.3 The DenseNet-121 Architecture

The second major architecture used in this thesis is **DenseNet-121**, a state-of-the-art CNN that is renowned for its parameter efficiency and strong performance [61]. Unlike traditional sequential networks, DenseNet is predicated on the principle of maximizing feature reuse through dense connectivity. As illustrated in Figure 2.2, the core of the architecture is the

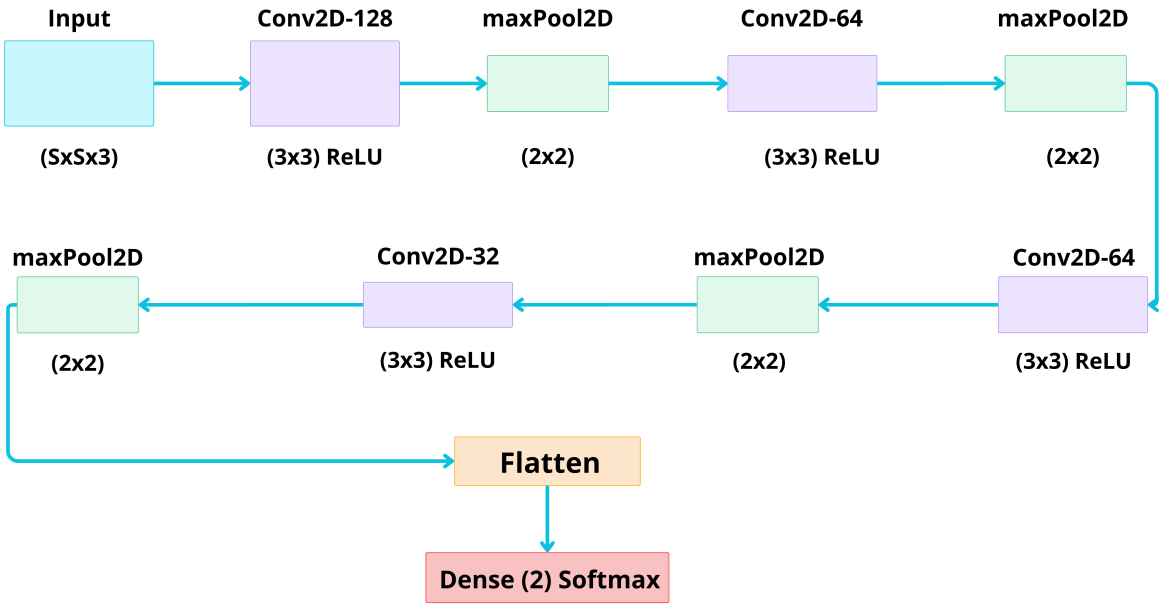


Figure 2.1: The proposed DeepSpecCNN architecture, designed for hierarchical feature extraction from spectrograms. It consists of four convolutional blocks followed by a dense classifier.

Dense Block. Within a dense block, each layer receives the feature maps from *all* preceding layers as its input, and its own output feature maps are passed on to *all* subsequent layers by concatenation. Between the dense blocks, **Transition Layers** are used to down-sample the feature maps.

2.4.4 Convolutional Block Attention Module (CBAM)

To enhance the representational power of the DenseNet121 backbone, this thesis integrates the **Convolutional Block Attention Module (CBAM)**. CBAM is a lightweight, "plug-and-play" module that refines feature maps sequentially along two separate dimensions: **Channel** and **Spatial**.

As illustrated in Figure 2.3, the CBAM sequentially infers attention maps to emphasize "what" features are important and "where" they are located.

The overall attention process for an input feature map $\mathbf{F} \in \mathbb{R}^{CHW}$ is summarized by:

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F} \tag{2.3}$$

$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}' \tag{2.4}$$

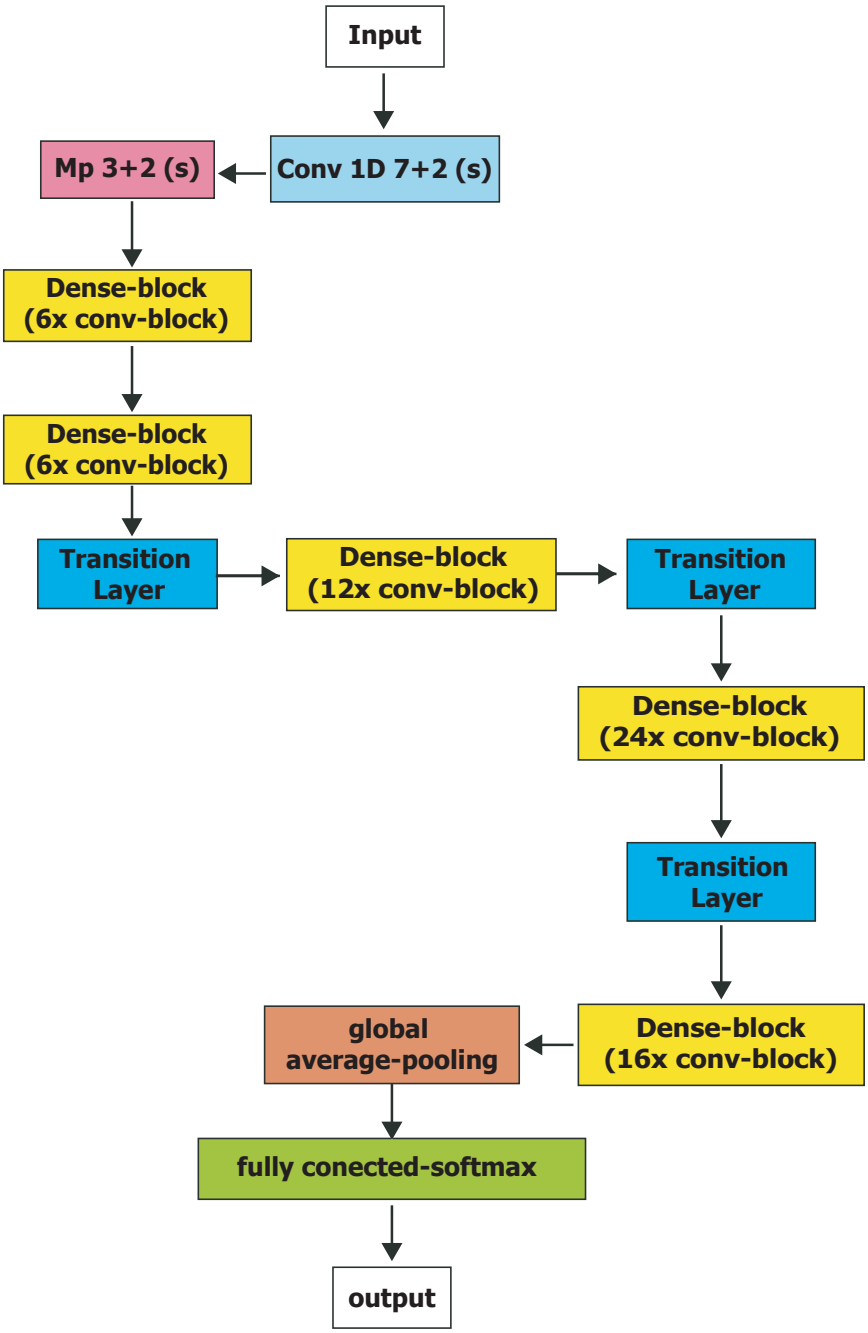


Figure 2.2: The architecture of DenseNet-121, illustrating the use of dense blocks and transition layers to encourage feature reuse and efficient information flow. (Source: [61, 62]).

where $\otimes$ denotes element-wise multiplication. The detailed structure of these sub-modules is illustrated in Figure 2.4.

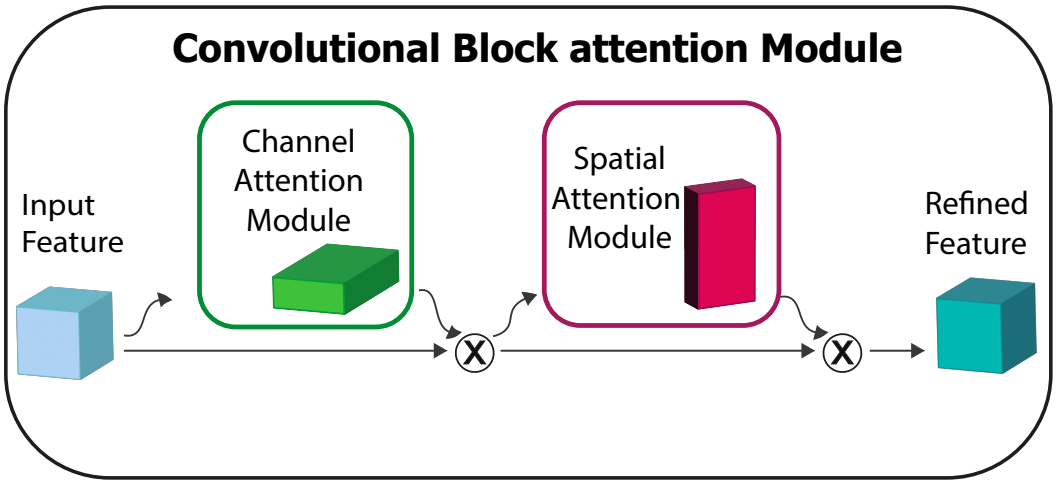


Figure 2.3: Overview of the Convolutional Block Attention Module (CBAM) illustrating the sequential application of channel and spatial attention. Source: [63].

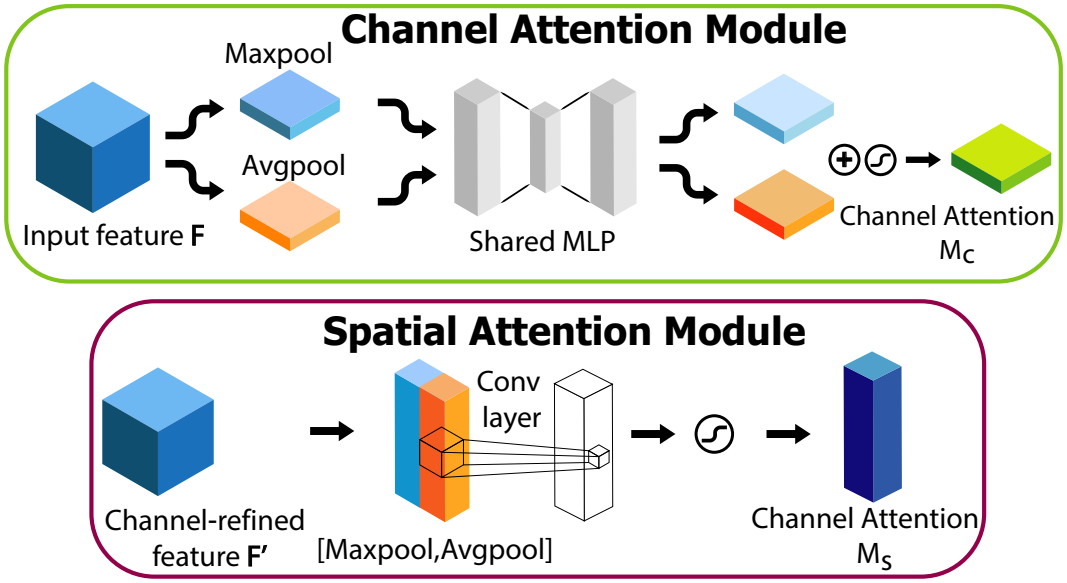


Figure 2.4: Detailed diagram illustrating the sub-modules of the CBAM: (a) Channel Attention Module and (b) Spatial Attention Module. Source: [63].

2.4.4.1 Channel Attention Module

This module exploits the inter-channel relationship of features. Since each channel is considered a feature detector, channel attention focuses on 'what' is meaningful. It uses both

average-pooling and max-pooling to aggregate spatial info, computed as:

$$\begin{aligned} M_c(F) &= \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \tag{2.5}$$

2.4.4.2 Spatial Attention Module

This module focuses on 'where' informative parts are located. It applies average and max pooling along the channel axis, concatenates them, and uses a 77 convolution to generate the spatial attention map:

$$\begin{aligned} M_s(F) &= \sigma(f^{77}([\text{AvgPool}(F); \text{MaxPool}(F)])) \\ &= \sigma(f^{77}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{2.6}$$

2.4.5 The Proposed CBAM-DenseNet121 Architecture

The first novel architecture proposed in this thesis, **CBAM-DenseNet121**, is formed by integrating the CBAM module into the DenseNet-121 backbone. The input spectrogram passes through the DenseNet-121 feature extractor, and the feature map produced after the final dense block is then fed into the CBAM module for refinement before being passed to the final classifier.

2.5 Ensemble Learning Framework

The second major framework proposed in this thesis is an **ensemble learning** approach. The specific strategy employed is a **Voting Ensemble**, which uses a majority vote from a committee of architecturally diverse CNNs (including DeepSpecCNN, LeNet, ResNet50, etc.). By aggregating their predictions, the uncorrelated errors of individual models tend to cancel out, leading to a more accurate and reliable final decision [64].

2.5.1 Ensemble Learning for Enhanced Robustness

While the development of a single, highly optimized model is a primary goal in machine learning, another powerful paradigm is **ensemble learning**. The core idea of this approach is that a "committee" of diverse models, when their predictions are combined, can often achieve better performance, robustness, and generalization than any single model in the ensemble could on its own [65]. The effectiveness of an ensemble relies on two key principles: **accuracy** and

diversity. The individual models (or "base learners") must perform better than random chance, and, crucially, they must be diverse, meaning they make different types of errors.

The power of ensembling can be formally understood through the lens of the **bias-variance tradeoff** [66]. A model's prediction error can be decomposed into three parts: bias (error from erroneous assumptions in the learning algorithm), variance (error from sensitivity to small fluctuations in the training set), and irreducible error. A single, complex model (like a deep decision tree) might have low bias but high variance, meaning it fits the training data perfectly but fails to generalize. Conversely, a simple model might have high bias but low variance. Ensemble methods are powerful because they provide strategies to reduce one or both of these error sources. There are several standard methods for creating effective ensembles:

- **Bagging (Bootstrap Aggregating):** Bagging is a technique designed primarily to **reduce variance**. It is particularly effective with high-variance, low-bias models like deep decision trees. The process is as follows:

  1. From the original training dataset of size N, create M new "bootstrap" datasets by sampling N instances with replacement. Each bootstrap dataset is a slightly different view of the original data.
  2. Train an independent base learner on each of the M bootstrap datasets. Because the training sets are different, the resulting models will be diverse.
  3. For a new prediction, aggregate the outputs of all M models. For classification, this is done via a majority vote (hard voting); for regression, the outputs are averaged.

  The **Random Forest** algorithm is a canonical and powerful extension of bagging, where in addition to training on bootstrap samples, each decision tree is further decorrelated by only considering a random subset of features at each split [46].

- **Boosting:** In contrast to the parallel nature of bagging, boosting trains base learners sequentially, with the primary goal of **reducing bias**. Each new model in the sequence is trained to correct the errors made by the existing ensemble.

  – **AdaBoost (Adaptive Boosting):** This method iteratively trains weak learners and, at each step, increases the weights of the training samples that were misclassified by the previous learner. This forces the next learner to focus more on these "hard" examples. The final model is a weighted sum of all the weak learners, where more accurate learners are given a higher weight in the final vote [47].

  - **Gradient Boosting Machines (GBM):** A more generalized and powerful boosting framework where each new weak learner is trained to predict the pseudo-residuals (the gradient of the loss function with respect to the predictions) of the predecessor ensemble. This turns the process into a direct, stage-wise optimization of the loss function [48].

- **Stacking (Stacked Generalization):** Stacking is a more advanced and often higher-performing ensemble technique that learns how to best combine the predictions from multiple different models [67]. The process involves two levels:

  1. **Level 0 (Base Models):** Several different base models (e.g., an SVM, a Random Forest, a CNN) are trained on the full training dataset.
  2. **Level 1 (Meta-Model):** A new model, the "meta-learner," is trained. The input features for this meta-learner are the outputs (predictions) of the base models from Level 0. The meta-learner's job is to learn the optimal combination of the base models' predictions to produce the final, improved prediction.

The approach used in this thesis falls into the category of a **Voting Ensemble**. Specifically, it uses a **majority vote** (or "hard voting") from a committee of architecturally diverse CNNs (including the custom DeepSpecCNN, LeNet, ResNet50, etc.). This strategy's success hinges on the principle of **architectural diversity**. Because each CNN has a different depth, connectivity pattern (e.g., residual vs. dense), and number of parameters, each one learns a slightly different representation of the data and is therefore likely to make uncorrelated errors. By taking a majority vote, the ensemble leverages this diversity to produce a final prediction that is more robust and accurate than any of its individual components. This strategy has proven highly effective in recent studies for both medical imaging and other complex classification tasks [68, 69].

2.6 Experimental Design and Evaluation Framework

The validity and impact of any computational study rest upon a meticulously designed and rigorously executed experimental framework. This section details the comprehensive methodology established for this thesis to ensure the scientific rigor, reproducibility, and objective evaluation of all models and findings. The framework is built upon three pillars: (1) the strategic selection and use of distinct datasets to address different research challenges; (2) the formulation of specific, hypothesis-driven classification tasks; and (3) a robust protocol

for model evaluation, including a clear data partitioning strategy and a suite of well-justified performance metrics. This systematic approach allows for a fair comparison between models and provides a solid empirical foundation for the conclusions drawn in this dissertation.

2.6.1 Corpora and Rationale for Selection

The experimental work in this thesis is strategically grounded in the use of two distinct datasets, each chosen to serve a specific research purpose and to address a different facet of the ASER problem.

- **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) [24]:** This corpus serves as the primary **benchmark dataset** for the development and validation of the novel deep learning architectures proposed in this thesis. As a high-quality, English-language collection of acted emotion, CREMA-D provides a controlled and acoustically clean environment ideal for testing new models. Its large number of speakers (91), clear categorical labels, and balanced class distribution make it a standard in the field, allowing for direct and meaningful performance comparisons with other state-of-the-art methods. The use of this benchmark is critical for demonstrating the architectural merits of the proposed models in a reproducible, standardized setting.
- **OYH (Open Your Heart) Dataset:** This novel corpus, a primary contribution of this thesis, was created to serve as a **real-world challenge dataset**. Comprising 6.3 hours of spontaneous emotional speech from an Algerian television talk show, it directly confronts the critical research gaps of domain mismatch and data scarcity for under-resourced dialects. Its defining characteristics—spontaneous "in-the-wild" speech, linguistic specificity to Algerian Arabic, and nuanced dimensional annotations—provide a challenging and ecologically valid testbed. The OYH corpus is used to establish baselines for a new, difficult problem domain and to ground the thesis in a practical, culturally relevant context.

2.6.2 Experimental Tasks and Hypotheses

The methodologies described in this chapter were applied to three distinct classification tasks, each designed to investigate a specific research question and test a core hypothesis of this thesis.

1. **Task 1: Dimensional Classification on the OYH Corpus.** The objective of this task is to establish a robust machine learning baseline for predicting Valence and Control

(each on a 3-point scale: Low, Medium, High) on the novel, spontaneous OYH dataset using traditional ML models. The central hypothesis is that for such complex, real-world data, performance is critically dependent on task-specific feature engineering, and that an advanced feature selection algorithm (BE-AWF) can identify distinct and more discriminative acoustic feature subsets for valence and control, respectively.

2. **Task 2: Categorical 6-Class Classification on CREMA-D.** This experiment addresses the standard ASER benchmark task of 6-class emotion recognition on the CREMA-D dataset. Its purpose is to investigate the trade-off between model accuracy and computational efficiency. The underlying hypothesis is that a lightweight, parameter-efficient CNN architecture enhanced with a targeted attention mechanism (the proposed CBAM-DenseNet121) can achieve a competitive performance against larger, more complex state-of-the-art models, making it a more practical solution for deployment.

3. **Task 3: Dimensional 2-Class Classification on CREMA-D.** The final experimental task aims to push the boundaries of classification accuracy on the CREMA-D benchmark. It involves a novel classification scheme based on the Geneva Wheel of Emotions (GWE), classifying emotions into a binary High-Control vs. Low-Control task. The hypothesis is twofold: first, that this psychologically-grounded dimensional grouping provides a more robust and less ambiguous classification target than discrete categories, and second, that an **ensemble** of architecturally diverse models will leverage this framework to achieve a new state-of-the-art accuracy by aggregating diverse "perspectives" on the data.

2.6.3 Evaluation Protocol and Performance Metrics

A standardized and rigorous evaluation protocol is essential for producing objective and reliable results. This protocol encompasses the data partitioning strategy and the selection of appropriate performance metrics.

2.6.3.1 Data Partitioning and Validation Strategy

To prevent data leakage and ensure that models are evaluated on entirely unseen data, a strict data partitioning strategy was employed. Each dataset was split into three mutually exclusive sets:

- A **Training Set**, which constitutes the largest portion of the data, used exclusively for learning the model's parameters.

- A **Validation (or Development) Set**, used during the training phase to tune model hyperparameters (e.g., learning rate, dropout rate) and for model selection (e.g., identifying the optimal training epoch to prevent overfitting).
- A **Test Set**, which is held out and used only once after all training and model selection is complete to provide a final, unbiased estimate of the model's generalization performance on new data.

The specific ratios for these splits (e.g., 70%/20%/10% or 70%/15%/15%) were chosen based on the dataset size and the specific requirements of each experimental study, as detailed in the corresponding chapters.

2.6.3.2 The Confusion Matrix

The cornerstone of performance analysis for any classification task is the **confusion matrix**. For a problem with N classes, this is an NN matrix where the rows represent the true, ground-truth labels and the columns represent the labels predicted by the model. Each cell (i, j) in the matrix contains the count of samples from true class i that were predicted as class j.

From this matrix, we can derive four key quantities for any given class C_i:

- **True Positives (TP_i):** The number of samples of class C_i that were correctly classified as C_i.
- **False Positives (FP_i):** The number of samples from other classes that were incorrectly classified as C_i.
- **False Negatives (FN_i):** The number of samples of class C_i that were incorrectly classified as a different class.
- **True Negatives (TN_i):** The number of samples from all other classes that were correctly not classified as C_i.

These fundamental counts form the basis for all other standard evaluation metrics.

2.6.3.3 Standard Evaluation Metrics

To provide a comprehensive and objective assessment of model performance, a standard suite of metrics was used. While **Accuracy**—the overall proportion of correct predictions—is a common metric, it can be highly misleading on datasets with imbalanced class distributions, a frequent issue in emotion recognition. Therefore, a more nuanced set of metrics, calculated on a per-class basis, was prioritized:

- **Precision:** This metric measures exactness for a single class C_i, answering the question: "Of all the samples the model predicted as class C_i, what fraction were actually class C_i?" It is crucial in applications where False Positives are costly. The per-class precision is calculated as:
$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$$
- **Recall (Sensitivity):** This metric measures completeness for a class C_i, answering the question: "Of all the true samples of class C_i in the dataset, what fraction did the model correctly identify?" It is vital in scenarios where missing a positive instance (a False Negative) is a critical error.
$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$
- **F1-Score:** This is the harmonic mean of Precision and Recall for a given class C_i. It provides a single, balanced measure of a model's performance for that class. The use of the harmonic mean ensures that the score is high only if both metrics are high. These per-class scores are often averaged to produce a single F1-score for the model.
$$\text{F1-Score}_i = 2\frac{\text{Precision}_i \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$
- **Unweighted Average Recall (UAR):** The UAR is the macro-average of the recall scores calculated for each individual class (Recall_i). It is arguably the most important single metric for evaluating ASER models on imbalanced datasets, as it treats each emotional category as equally important, regardless of its prevalence in the test set. This provides a fair and robust measure of how well the model performs across the entire emotional spectrum.

2.6.4 Implementation and Computational Environment

To ensure the reproducibility of the research presented, all experiments were conducted using publicly available and well-documented software libraries. The primary programming language was Python (version 3.8+). For the traditional machine learning experiments, the **Scikit-learn** library was used. For the deep learning experiments, the **PyTorch** framework was employed. Audio processing and feature extraction were performed using the **Librosa** library for spectrogram generation and the standardized **openSMILE** toolkit for handcrafted features. Model training was accelerated using NVIDIA Tesla T4 and V100 Graphics Processing Units (GPUs) in a cloud computing environment.

2.7 Conclusion

This chapter has detailed the complete methodological arsenal that underpins the experimental work of this thesis. The objective was to build a comprehensive and foundational understanding of the algorithms and frameworks used, moving from general principles to the specific architectural families and models relevant to ASER. We began with an overview of machine learning, covering the principles of supervised learning and the mechanics of **traditional classifiers** like SVMs and Random Forests that serve as a performance baseline.

Subsequently, the chapter provided a deep and pedagogical dive into the **deep learning paradigm**, constructing the conceptual framework from the fundamental artificial neuron up to the modern mechanics of network training. Building on this, we examined the core architectures for spectrogram analysis, providing a thorough, illustrated explanation of the proposed **DeepSpecCNN** and **CBAM-DenseNet121** models, as well as the robust strategy of **ensemble learning**. Finally, the complete experimental protocol and a rigorous framework for performance assessment were established.

With this comprehensive methodological framework now in place, we are fully equipped to apply these powerful tools to concrete research problems. The following chapters will transition from theory to practice, presenting the experimental application of these models and evaluation principles to the benchmark CREMA-D dataset and the novel OYH corpus to develop and validate new state-of-the-art solutions for speech emotion recognition.

CHAPTER 3

EXPERIMENTS ON THE CREMA-D BENCHMARK DATASET

Contents

3.1 Introduction

Having established the foundational data representations and methodological principles in the preceding chapters, this chapter transitions from theory to practice. It presents the core of the deep learning experimentation conducted for this thesis, focusing on the development and validation of novel models on a standard, high-quality benchmark dataset. The use of a benchmark like the **Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)** is a critical step in any rigorous research endeavor. It provides a controlled, clean, and widely accepted environment for testing new architectures, allowing for direct and fair comparison with state-of-the-art methods from the broader research community.

This chapter details two distinct but complementary studies performed on the CREMA-D dataset, each designed to address a key challenge in modern ASER. These studies correspond to two central research questions of this thesis:

1. Can a single, lightweight Convolutional Neural Network architecture, enhanced with modern attention mechanisms, achieve a competitive balance between classification accuracy and computational efficiency for the standard 6-class emotion recognition task?

2. Can reframing the classification problem using a psychologically-grounded dimensional model and leveraging an ensemble of diverse architectures push the boundaries of state-of-the-art accuracy?

To answer these questions, the chapter is structured into two main sections. **Study 1** details the development, training, and evaluation of the **CBAM-DenseNet121** model. This study focuses on the crucial trade-off between performance and efficiency, presenting a detailed analysis of the model's performance on the 6-class task and validating its architectural components through a rigorous ablation study. **Study 2** then shifts focus from efficiency to achieving maximal accuracy. It introduces a novel classification framework based on the Geneva Wheel of Emotions and presents the **DeepSpecCNN** model and its integration into a powerful, high-performance **ensemble learning** framework.

Together, these two studies provide a comprehensive exploration of deep learning for ASER on benchmark data, yielding both a practical, efficient model and a state-of-the-art, high-accuracy framework. The findings from this chapter will lay the groundwork for the subsequent chapter, which will investigate the challenges of applying these methods to a more complex, real-world dataset.

3.2 Study 1: An Efficient Attention-Enhanced Architecture for 6-Class Emotion Recognition

The first study addresses the pressing need for ASER models that are not only accurate but also computationally efficient and practical for real-world deployment. While large, complex models often achieve the highest accuracy, their significant computational overhead limits their feasibility. This study investigates the potential of a carefully designed, lightweight CNN architecture enhanced with an attention mechanism to achieve a competitive balance between these two goals.

3.2.1 Experimental Setup

3.2.1.1 Dataset and Task

The foundation of this study is the **Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)** [24], a prominent and widely-used English-language benchmark in the affective computing community. The selection of this corpus was deliberate and strategic for several reasons. First, as a benchmark, it provides a standardized platform for evaluating the proposed **CBAM-DenseNet121** architecture, enabling direct and fair comparisons against a large body of existing state-of-the-art methods. Second, the dataset consists of high-quality, acoustically clean recordings captured in a controlled laboratory environment. This minimizes the influence of confounding variables like background noise or reverberation, thereby isolating the evaluation to the architectural merits of the model itself—a critical requirement when the primary goal is to assess the impact of specific design choices, such as the integration of an attention mechanism. Finally, the corpus features a large and ethnically diverse set of 91 actors, which provides a degree of speaker variability that is essential for training a generalizable model. Although CREMA-D is a multimodal dataset containing both audio and video, this study operates within a strictly unimodal ASER framework, utilizing only the audio channel.

The experimental objective is framed as the quintessential ASER problem: a **6-class categorical emotion recognition** task. While Chapter 1 acknowledged the theoretical limitations of discrete emotion models, this categorical approach was intentionally chosen for this study to align with the predominant paradigm in the literature, thus ensuring the comparability of our results. The model is tasked with classifying each speech utterance into one of the six canonical "basic" emotions: **Anger, Disgust, Fear, Happy, Sad**, or a non-emotional **Neutral** state.

As detailed in Table 3.1, the CREMA-D dataset is exceptionally well-suited for this task due to its balanced class distribution. The five core emotional categories contain an identical

number of samples (1,271 each), with the Neutral class being only slightly less represented (1,087 samples). This high degree of balance mitigates the risk of the model developing a bias towards a majority class during training and simplifies the interpretation of performance metrics. It provides an ideal scenario for rigorously assessing the model's ability to discriminate between distinct emotional expressions.

Table 3.1: Label Distribution of the CREMA-D Dataset for the 6-Class Task.

Emotion	**Abbreviation**	**Number of Samples**
Anger	ANG	1271
Disgust	DIS	1271
Fear	FEA	1271
Happy	HAP	1271
Sad	SAD	1271
Neutral	NEU	1087
Total		**7442**

3.2.1.2 Data Partitioning and Preprocessing

A robust and unbiased evaluation protocol requires a meticulous data partitioning and preprocessing strategy. This section details the steps taken to prepare the CREMA-D dataset for model training and to ensure that the final performance evaluation is a true measure of the model's generalization capability.

The dataset was partitioned into three distinct, non-overlapping sets: a training set (70%), a validation set (20%), and a test set (10%). This 70/20/10 split is a standard practice that allocates a substantial majority of the data for model learning, provides a sufficiently large validation set for reliable hyperparameter tuning and prevention of overfitting, and reserves a final holdout set for an unbiased performance assessment. Crucially, this split was performed with speaker independence. This means that all utterances from any single actor were confined to only one of the three sets. This practice is essential in ASER to prevent the model from simply memorizing the vocal characteristics of specific speakers, forcing it instead to learn generalizable acoustic patterns of emotion that are speaker-agnostic. The partitioning was also stratified, ensuring that the proportional representation of each of the six emotion classes was maintained across all three sets, thereby guaranteeing that the validation and test sets are representative of the overall data distribution. The resulting sample distribution is detailed in Table 3.2.

Table 3.2: Data Partitioning for Study 1.

Total Samples	Train Set (70%)	Validation Set (20%)	Test Set (10%)
7442	5209	1488	745

The preprocessing pipeline was designed to convert the raw audio signals into a format suitable for consumption by a Convolutional Neural Network. As established in Chapter 1, all audio files were transformed into log-Mel spectrograms. This two-dimensional representation is perceptually motivated, mimicking the frequency response of the human ear, and has become the de facto standard for ASER, effectively reframing the audio classification problem as an image recognition task.

To ensure uniformity for batch processing, the resulting spectrograms were resized to a fixed dimension of 128×256 pixels. Here, 128 represents the number of Mel frequency bins, a resolution that captures sufficient spectral detail, while 256 represents the number of time frames, achieved by padding or truncating the audio clips. This size strikes a balance between preserving spectro-temporal information and maintaining computational feasibility. Finally, to leverage the power of transfer learning, a critical step was performed. Although spectrograms are inherently single-channel (grayscale), the channel dimension was replicated three times to match the 224×224×3 input shape expected by the DenseNet121 backbone, which was pre-trained on the massive, three-channel RGB ImageNet dataset. This technique allows the model to utilize the rich, hierarchical visual features learned from millions of natural images as a powerful starting point for learning the patterns within the speech spectrograms.

3.2.1.3 Data Augmentation

To enhance the model's ability to generalize and to mitigate the risk of overfitting, data augmentation techniques were applied exclusively to the training set. During training, spectrogram images were subjected to random color jittering with a 50% probability. This included minor, random adjustments to the image's **brightness, contrast, saturation, and hue**. This process creates new, slightly modified training examples at each epoch, forcing the model to learn more robust and invariant features. Validation and test spectrograms were not augmented.

3.2.1.4 Model Architecture and Training Protocol

This study employs the novel **CBAM-DenseNet121** architecture, the design principles of which were established in Chapter 2. The choice of this specific architecture is grounded in a strategic effort to balance high classification accuracy with computational efficiency. The

model is built upon a **DenseNet-121** backbone, which was selected for its characteristic dense connectivity pattern. This design encourages feature reuse throughout the network, improves the flow of information and gradients, and has been shown to achieve competitive performance with significantly fewer parameters than comparable architectures like ResNet.

To further enhance the representational power of this efficient backbone, a **Convolutional Block Attention Module (CBAM)** was integrated after the final dense block. The purpose of this lightweight module is to enable the model to perform feature refinement, adaptively learning to emphasize the most informative features both along the channel axis ("what" is important) and the spatial axis ("where" in the spectrogram is important). The complete layer-wise architecture and parameter distribution are summarized in Table 3.3. With a total of approximately **7.1 million** trainable parameters, the resulting model is positioned as a relatively lightweight yet powerful solution, designed for practical applicability where computational resources may be constrained.

Table 3.3: Layer-wise Parameter Summary of the CBAM-DenseNet121 Model.

Block	**Output Shape**	**Parameters**
Input	[batch, 3, 128, 256]	0
Conv0	[batch, 64, 64, 128]	9,408
DenseBlock 1	[batch, 256, 32, 64]	335,040
Transition 1	[batch, 128, 16, 32]	33,280
DenseBlock 2	[batch, 512, 16, 32]	919,680
Transition 2	[batch, 256, 8, 16]	132,096
DenseBlock 3	[batch, 1024, 8, 16]	2,837,760
Transition 3	[batch, 512, 4, 8]	526,336
DenseBlock 4	[batch, 1024, 4, 8]	2,158,080
CBAM	[batch, 1024, 4, 8]	132,260
Classifier	[batch, 6]	8,198
Total Trainable Parameters	-	**7,092,138**

The model was trained using a carefully selected set of modern deep learning practices and hyperparameters. The optimization was driven by the **Categorical Cross-Entropy** loss function, which is the standard and most appropriate choice for multi-class classification tasks. For parameter updates, the **AdamW optimizer** was employed. This optimizer is a robust variant of the popular Adam algorithm that decouples weight decay from the gradient updates, a modification that has been shown to improve generalization performance.

The training was configured to run for a maximum of **100 epochs** with a batch size of **32**. The initial learning rate was set to 110^{-3} with a weight decay of 110^{-4} for regularization. To

dynamically manage the learning rate, a **cosine annealing scheduler** was used. This scheduler gradually decreases the learning rate following a cosine curve over the course of an epoch, which can help the model settle into broader and more robust minima in the loss landscape. Furthermore, to prevent overfitting and reduce unnecessary computation, an early stopping mechanism was implemented. The training process was monitored, and if the validation loss did not show improvement for a patience of 10 consecutive epochs, the training was halted, and the model weights from the epoch with the best validation performance were saved for the final evaluation.

3.2.2 Results and Analysis

3.2.2.1 Training Dynamics

The learning process of the model over 100 epochs is visualized in Figure 3.1. A detailed analysis of the curves reveals several key dynamics. The training accuracy (blue line) exhibits a steep and rapid increase, reaching nearly 100% (**99.98%**) by the end of training. This indicates that the model has sufficient capacity to fully memorize or "fit" the training data. The validation accuracy (orange line) also shows a healthy learning trajectory, rising steadily before beginning to plateau around epoch 40 at a final value of **67.49%**.

The loss curves provide a complementary view. The training loss (blue) steadily decreases towards zero, confirming that the optimization process was successful. The validation loss (orange) decreases in tandem for the first 40 epochs, after which it flattens and shows slight fluctuations. The divergence point between the training and validation loss curves (around epoch 40) is the classic indicator of the onset of overfitting. However, the fact that the validation loss remains stable and does not significantly increase suggests that the regularization techniques (like weight decay and dropout) were effective in preventing severe overfitting.

3.2.2.2 Test Set Performance and Confusion Matrix Analysis

The final, definitive performance of the trained model was evaluated on the unseen test set. The model achieved a strong overall test accuracy of **71.26%**. To ensure a fair assessment given the slight class imbalance, the F1-Score and Unweighted Average Recall (UAR) are the most critical metrics. The model obtained an excellent F1-Score of **71.25%** and a UAR of **71.01%**, demonstrating a robust and balanced performance across all six emotion classes.

A detailed, class-wise analysis is provided by the confusion matrix in Figure 3.2. The diagonal elements, representing correct predictions, show the model's strengths. It performs exceptionally well in identifying **Angry (ANG)**, **Happy (HAP)**, and **Neutral (NEU)** emotions,

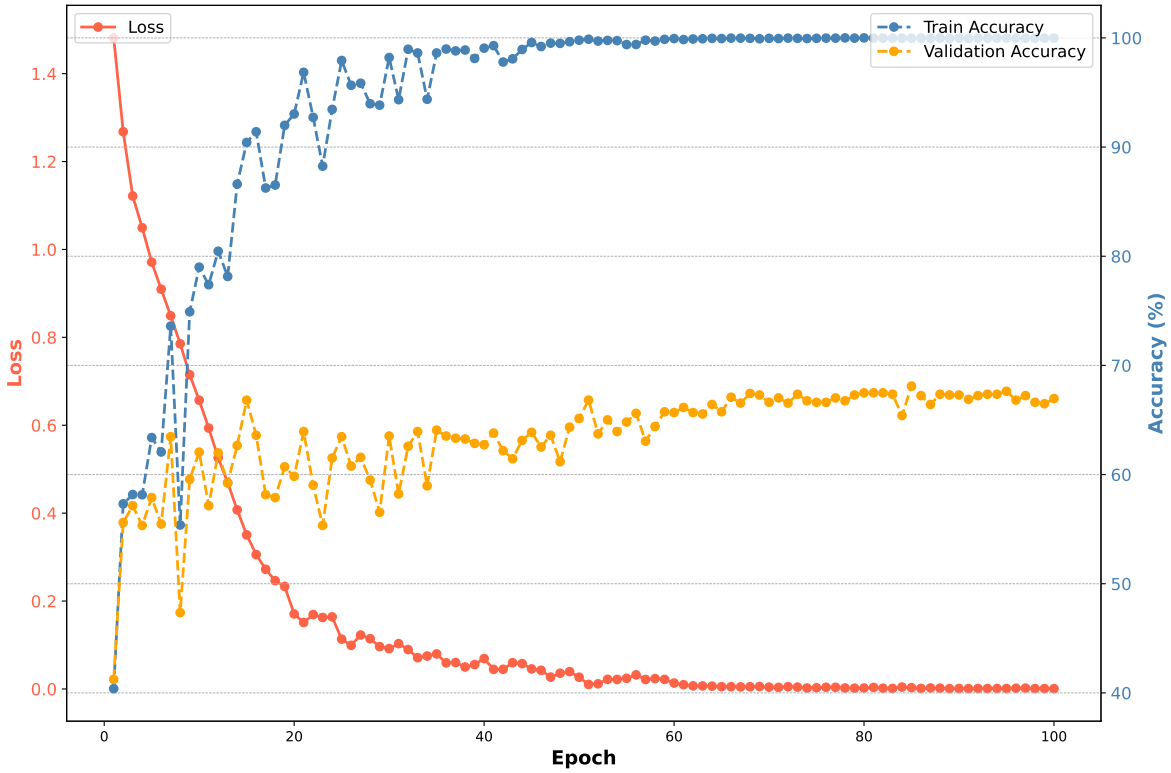


Figure 3.1: Analysis of Training and Validation Performance for the CBAM-DenseNet121 model, showing accuracy and loss curves over 100 epochs.

correctly classifying 104, 95, and 91 samples respectively for these classes. However, the off-diagonal elements reveal specific areas of confusion. The most significant challenge for the model is distinguishing between low-arousal, negative-valence emotions. For instance, out of 127 true **Fear (FEA)** samples, 21 were misclassified as **Sad (SAD)**. Similarly, **Disgust (DIS)** was most often confused with Sad (16 samples). These results indicate that the acoustic features of these emotions have significant overlap in the learned feature space, a common and persistent challenge in the field of ASER.

3.2.2.3 Comparison with State-of-the-Art (SOTA)

A central goal of this study was to achieve competitive accuracy while maintaining computational efficiency. Table 3.4 compares our proposed model against other state-of-the-art architectures on the CREMA-D dataset. Our CBAM-DenseNet121 model achieves a test accuracy of **71.26%** with only **~7.1 million** trainable parameters. This performance modestly surpasses more complex transformer-based models like LeRaC + SepTr (70.95%), while using less than half the number of parameters. It significantly outperforms the much larger Vision

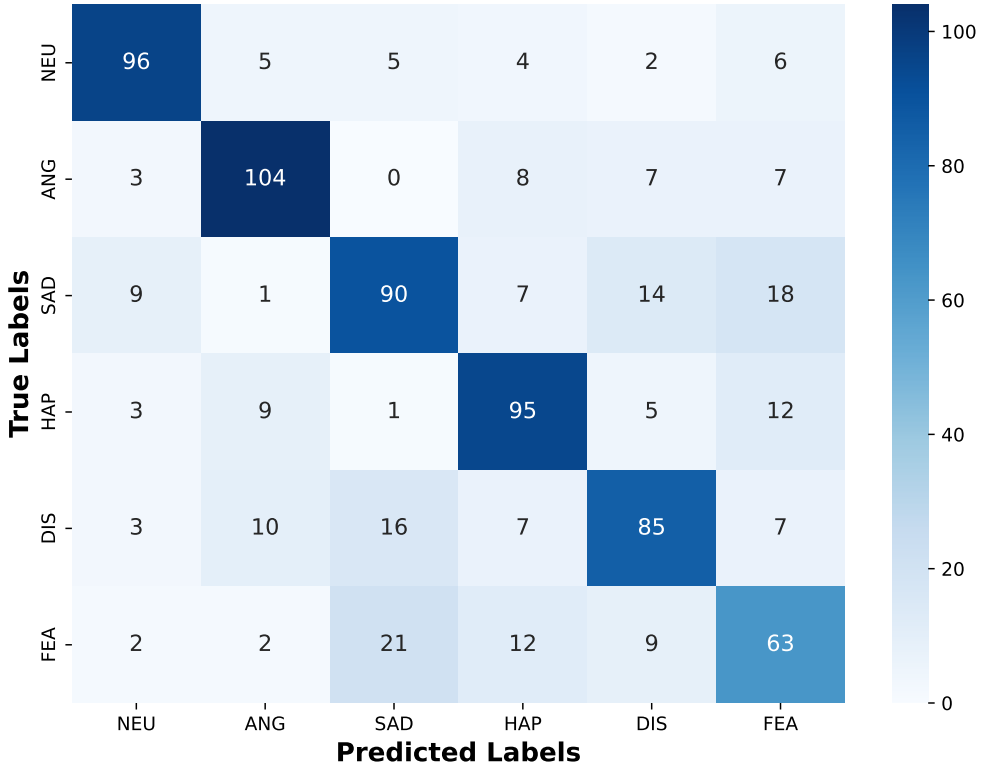


Figure 3.2: Confusion Matrix of the CBAM-DenseNet121 model on the CREMA-D test set for the 6-class task.

Transformer (ViT), which has ~86M parameters but achieves only 67.81% accuracy. This highlights the strength of a well-designed CNN architecture with strong inductive biases for processing spectrograms. The efficiency of our model is a direct result of the parameter-efficient DenseNet backbone and the lightweight nature of the CBAM module, demonstrating a clear advantage for practical applications where computational resources are limited.

Table 3.4: Performance and Complexity Comparison of Proposed Model Against Baseline and SOTA Architectures on CREMA-D. Our model is in **bold**.

Model	**Accuracy (%)**	**Parameters (Approx.)**
Lin et al. (2025) [70]	82.00	Large (not specified)
CBAM-DenseNet121 (Ours)	**71.26**	**~7.1M**
LeRaC + SepTr [71]	70.95	~15M
SepTr [72]	70.47	~15M
ResNet-18 + SPEL [73]	68.12	~11.7M
ViT (Audio Spectrogram Transformer) [74]	67.81	~86M
SpectoResNet [75]	65.20	~11.3M

3.2.3 Ablation Study

To scientifically validate the architectural choices made in this study, a rigorous ablation study was conducted to isolate and evaluate the impact of the two key components: the CBAM attention mechanism and the dropout rate.

3.2.3.1 Evaluating the Impact of CBAM

To scientifically validate the architectural design of the proposed model, a rigorous **ablation study** was conducted. This process is critical in deep learning research as it isolates the contribution of individual components to the overall performance, moving beyond conjecture to provide empirical evidence of their efficacy. In this study, the primary goal was to quantify the impact of the **Convolutional Block Attention Module (CBAM)**. To achieve this, the module was integrated into several distinct, pre-trained CNN backbones, and the performance was compared against the baseline versions of those same architectures.

The results, presented in Table 3.5, reveal a nuanced and architecture-dependent impact. The most compelling finding is the clear synergistic effect between CBAM and the **DenseNet121** backbone. Integrating the attention module provided a substantial performance boost across all key metrics, most notably increasing the overall accuracy by over 2 percentage points from 69.25% to **71.26%** and the UAR from 69.34% to **71.01%**. This strong positive interaction is likely due to DenseNet's core mechanism of feature concatenation; as each layer receives a rich pool of feature maps from all preceding layers, the CBAM module can effectively act as a selection filter, learning to prioritize the most discriminative features for the final classification. As visualized in Figure 3.3, this enhancement was particularly impactful for acoustically similar and challenging classes like 'Disgust' (DIS) and 'Fear' (FEA), strongly suggesting that the attention mechanism successfully guided the model's focus toward the subtle spectro-temporal cues that differentiate these easily confused emotions.

The effect on other backbones was less pronounced. Adding CBAM to **ResNet50** yielded a modest but consistent improvement, suggesting that standard residual blocks can also benefit from explicit feature recalibration. Conversely, and quite interestingly, adding CBAM to the highly optimized **MobileNetV2** architecture resulted in a slight degradation in performance. A plausible hypothesis for this counterintuitive result is architectural redundancy. MobileNetV2's inverted residual blocks already contain an inherent form of feature gating through their linear bottlenecks, which may make an external attention module functionally redundant or even disruptive to its finely-tuned information flow.

In summary, this ablation study empirically confirms that the CBAM module is a valuable

component of the proposed architecture. However, it also critically demonstrates that the benefits of attention are not universal but are highly dependent on the base architecture with which they are paired. The results validate our central hypothesis that a synergy exists between the feature aggregation of DenseNet and the feature refinement of CBAM, making their combination particularly potent for this ASER task.

Table 3.5: Ablation Study: Evaluating the performance impact of adding the CBAM module to different CNN architectures.

Model	**Accuracy (%)**	**F1 (%)**	**Precision (%)**	**UAR (%)**	**Recall (%)**
MobileNetV2	69.92	69.79	69.97	69.72	69.73
MobileNetV2 + CBAM	67.91	67.93	68.17	67.92	68.51
ResNet50	55.21	54.56	56.64	55.40	57.74
ResNet50 + CBAM	56.95	56.41	56.87	56.95	57.78
DenseNet121 (Finetuned)	69.25	69.05	69.44	69.34	69.67
DenseNet121 (FT + CBAM)	**71.26**	**71.25**	**71.30**	**71.01**	**70.34**

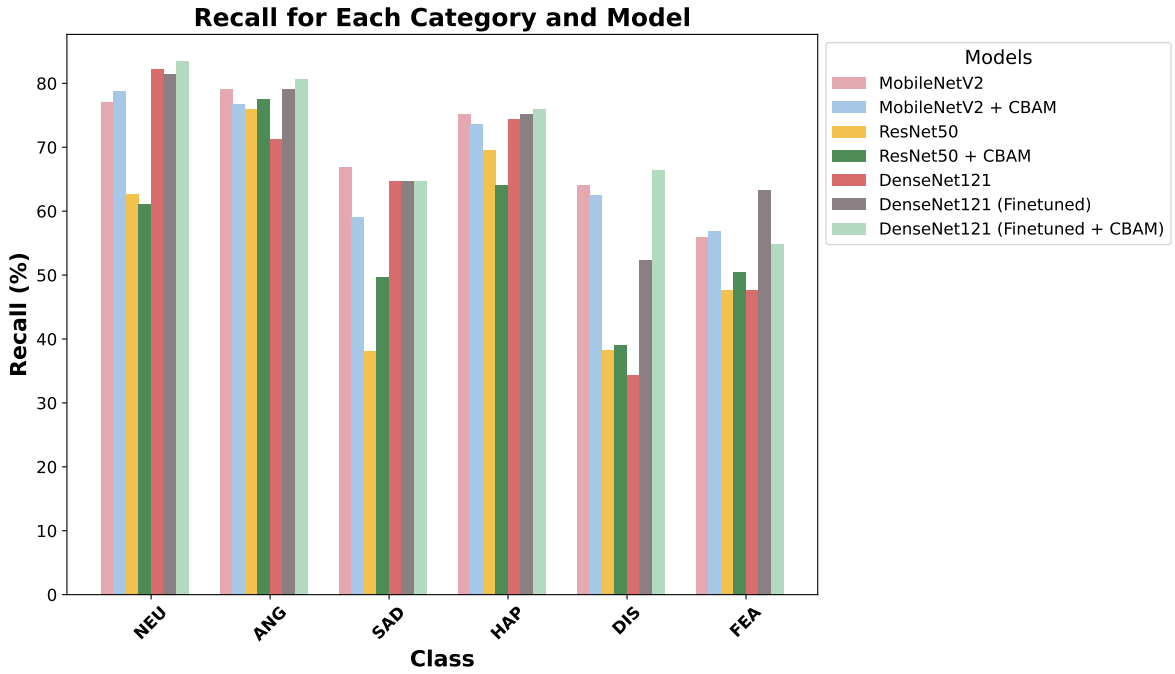


Figure 3.3: Class-wise recall for each model in the ablation study, highlighting the significant improvement for 'Disgust' (DIS) and 'Fear' (FEA) after adding CBAM to DenseNet121.

3.2.3.2 Impact of Dropout Rate

Dropout is a critical stochastic regularization technique designed to combat overfitting in deep neural networks. By randomly setting the activations of a fraction of neurons to zero during each forward pass of training, dropout prevents neurons from forming complex co-adaptations on the training data. This forces the network to learn more robust and redundant feature representations, thereby improving its ability to generalize to unseen data. The dropout rate, which defines the probability of a neuron being deactivated, is therefore a critical hyperparameter that must be carefully tuned to find the optimal balance between underfitting and overfitting.

To empirically determine this optimal rate for the proposed **CBAM-DenseNet121** architecture, an ablation study was conducted, with the results summarized in Table 3.6. The analysis of the performance across different rates reveals a clear trend. At lower rates (e.g., 0.1 to 0.3), the model's performance improves steadily, but it is not yet optimal. This suggests that the regularization effect is insufficient to fully prevent the model from overfitting to the intricacies of the training set.

The peak performance is clearly achieved at a dropout rate of **0.5**, which emerged as the empirical "sweet spot" for this architecture and task. At this level, the regularization is aggressive enough to effectively break up spurious co-adaptations and promote generalization, leading to the best scores across all key metrics, including an accuracy of **71.26%** and a UAR of **71.01%**. As the dropout rate increases beyond this point (e.g., hypothetically to 0.6 or 0.7), one would expect performance to decline due to underfitting, where the network's capacity is so heavily penalized that it struggles to learn the underlying patterns in the data. This meticulous tuning was a crucial step in finalizing the model's configuration, confirming that a 50% dropout rate provides the ideal regularization strength to maximize the performance of the CBAM-DenseNet121 model.

Table 3.6: Ablation Study: Effect of Dropout Rate on the final CBAM-DenseNet121 model's performance.

Dropout	Accuracy (%)	UAR (%)	Precision (%)	F1 Score (%)
0.1	67.65	67.70	67.79	67.61
0.2	68.18	68.05	68.60	67.79
0.3	69.92	69.66	70.04	69.57
0.4	68.05	67.96	67.88	67.87
0.5	**71.26**	**71.01**	**71.30**	**71.25**

3.2.4 Discussion of Study 1

The results of this first study successfully demonstrate that a carefully designed, attention-enhanced CNN can achieve a highly competitive balance between classification accuracy and computational efficiency for the standard 6-class ASER task. The proposed **CBAM-DenseNet121** architecture provides a strong affirmative answer to our first research question, presenting a validated, lightweight, and effective model for practical speech emotion recognition. The analysis, however, merits a deeper interpretation of the findings and their implications.

A key insight from this study is the validation of a well-crafted CNN architecture in an era increasingly dominated by larger Transformer-based models. The fact that the ~7M parameter CBAM-DenseNet121 outperformed the significantly larger ~86M parameter Vision Transformer is particularly telling. This result underscores the continued importance of architectural **inductive bias**. CNNs, with their inherent properties of locality and translation invariance, are naturally suited to processing the grid-like structure of spectrograms. This innate structural understanding makes them highly data-efficient for this task. Transformers, lacking this built-in bias, often require vast amounts of data to learn spatial hierarchies, which may explain their comparatively lower performance here.

Furthermore, the ablation study provided strong empirical evidence for the contribution of the CBAM module. Its success highlights that for ASER, not all parts of a spectrogram are equally important. Emotional cues are often encoded in subtle, localized spectro-temporal events—such as formant shifts, pitch inflections, or bursts of energy. The attention mechanism's ability to adaptively focus on these salient regions, while suppressing less relevant information, proved crucial for improving the model's ability to distinguish between challenging, easily confused emotions like 'Fear' and 'Disgust'.

It is crucial, however, to contextualize these findings within the limitations of the study's experimental setup. The use of the **CREMA-D dataset**, while ideal for controlled architectural benchmarking, means the model was exclusively tested on acted, acoustically clean speech. The impressive performance achieved here does not guarantee similar success in real-world scenarios, which are characterized by spontaneous emotional expressions, background noise, and dialectal variations. This well-known **"domain gap"** between laboratory data and "in-the-wild" speech remains a formidable challenge.

This study acts as a fundamental first step by establishing an efficient modeling method under controlled conditions. The successful performance of the CBAM-DenseNet121 architecture on benchmark data validates the transition toward more difficult research areas, specifically evaluating the model's effectiveness against spontaneous speech and identifying appropriate methodologies for emotion recognition within under-resourced dialectal contexts.

3.3 Study 2: State-of-the-Art Accuracy via Ensemble Learning on a Dimensional Emotion Framework

While Study 1 successfully developed an efficient, lightweight model for practical applications, this second study shifts the research objective from a balance of performance and efficiency to the pursuit of the highest possible classification accuracy. The central research question here is how to overcome the inherent performance ceilings of single-model, categorical ASER systems to establish a new state-of-the-art on a benchmark dataset. To accomplish this, a two-pronged strategy was devised, combining a theoretically-grounded reformulation of the classification task with a powerful, multi-model learning paradigm.

The first strategy directly confronts the limitations of the categorical emotion model, a core problem identified in Chapter 1. The ambiguity and acoustic overlap between discrete classes like 'Fear' and 'Sadness' often create a noisy learning signal, limiting a classifier's performance. To mitigate this, this study reframes the complex 6-class problem into a simpler, more robust binary task based on the **control dimension** of the Geneva Wheel of Emotions (GWE). The underlying hypothesis is that by grouping emotions along a psychologically meaningful axis that has strong acoustic correlates (e.g., vocal stability, intensity, and articulation), we can create a classification task with a clearer and more learnable decision boundary than one based on discrete categorical labels alone.

The second strategy leverages the principle of **ensemble learning** to push beyond the capabilities of any single architecture. Even a highly optimized model has its own intrinsic biases and will make certain types of errors. The core hypothesis of this approach is that a committee of architecturally diverse CNNs, each learning a slightly different representation of the data, will produce uncorrelated errors. By aggregating their individual predictions through a majority voting mechanism, these errors can be effectively cancelled out, resulting in a collective decision that is more robust and accurate than that of any individual constituent model. This study therefore investigates the combined power of these two strategies, aiming to demonstrate that a psychologically-informed task representation paired with a strategic ensemble framework provides a superior pathway to achieving state-of-the-art accuracy in speech emotion recognition.

3.3.1 The Dimensional Classification Framework

As established in Chapter 1, a significant performance bottleneck in ASER arises from the inherent ambiguity and acoustic overlap between discrete emotion categories. The confusion

between classes like 'Fear' and 'Sadness' introduces label noise and creates a complex, often ill-defined decision boundary for classifiers. To circumvent this limitation, this study proposes a strategic reformulation of the classification task itself, moving from a purely categorical approach to one grounded in a dimensional model of affect: the **Geneva Wheel of Emotions (GWE)** [20], previously illustrated in Figure 1.1.

Instead of treating emotions as six independent classes, we map them onto the GWE's **control dimension**. This dimension, distinct from the more common arousal (activation) axis, pertains to an individual's sense of power, agency, and control over a situation. This choice is theoretically motivated by its strong and intuitive link to vocal production. A state of **High Control**—a feeling of power or assertiveness—is often acoustically manifested through a more stable and controlled voice: a steady pitch contour (low jitter), consistent intensity, clear articulation, and often a faster speech rate. Conversely, a state of **Low Control**—a feeling of helplessness or submissiveness—is typically associated with less regulated vocal patterns: a trembling or unstable pitch (high jitter and shimmer), a weaker or breathier voice quality (low Harmonic-to-Noise Ratio), and more hesitant speech.

Based on this strong theoretical and acoustic link, the six emotions from the CREMA-D dataset were mapped into two psychologically-grounded macro-classes. The **High Control** class comprises emotions where the speaker feels a sense of agency, namely **Anger, Happiness, and Disgust**. The **Low Control** class consists of emotions associated with a lack of agency or power: **Fear, Sadness, and Neutrality**.

This transformation simplifies the complex 6-class problem into a more tractable binary classification task. As detailed in Table 3.7, this re-mapping also has the practical benefit of creating a highly balanced dataset, which is ideal for training and evaluation. The central hypothesis of this framework is that this binary task will prove more robust and yield higher classification accuracy, as the model is learning to distinguish between two macro-classes defined by a clearer and more consistent set of underlying acoustic correlates than those of the six discrete emotions.

Table 3.7: Class Distribution for the Binary Dimensional Task.

Macro-Class	**Number of Samples**	**Percentage of Dataset**
High Control	3630	48.76%
Low Control	3812	51.24%
Total	**7442**	**100%**

3.3.2 Experimental Setup

3.3.2.1 Data Partitioning and Base Model Selection

For this study, the CREMA-D dataset was partitioned into a **training set (70%)**, a **validation set (15%)**, and a **test set (15%)**. This 70/15/15 split was chosen to provide a substantial corpus for training while dedicating a larger and more statistically robust test set for the final performance evaluation, which is critical when the goal is to confidently establish a new state-of-the-art. As in the previous study, the partitioning was conducted with strict **speaker independence** to ensure that the model's generalization capability to new speakers was being fairly assessed.

The cornerstone of the ensemble learning strategy is the principle of **diversity**. A successful ensemble relies on combining base learners that are not only individually accurate but also make different types of errors. To achieve this, a diverse set of 11 CNN architectures, representing different families and eras of computer vision research, were selected as base learners. The hypothesis is that their varied architectural designs, depths, and connectivity patterns will cause them to learn complementary feature representations, leading to the decorrelated errors necessary for a high-performing ensemble. The selected models include:

- **Custom Architecture:** The bespoke **DeepSpecCNN**, designed in Chapter 2 specifically for spectrogram analysis.
- **Pioneering Architectures:** Foundational models like **LeNet-5** and **AlexNet** that established the core principles of CNNs.
- **Very Deep Sequential Architectures:** Models like **VGG16** and **VGG19**, known for their deep, homogenous stacks of 3x3 convolutions.
- **Architectures with Advanced Connectivity: ResNet50**, which introduced residual connections to train extremely deep networks, and **DenseNet121**, which uses dense connections to maximize feature reuse.
- **Lightweight, Efficient Architectures:** Modern models designed for computational efficiency, including **SqueezeNet**, **MobileNet**, and **EfficientNetB0**.

By curating this architecturally diverse portfolio, we create a rich pool of base learners, forming a robust foundation for testing our ensemble learning hypothesis.

3.3.2.2 Ensemble Strategy and Training Protocol

The core of this study's methodology is its ensemble strategy, designed to be both powerful and nuanced. The specific approach employed was **soft voting**. In this framework, each of the k base models in an ensemble outputs a vector of predicted probabilities for each class. These probability vectors are then averaged across all models, and the final ensemble prediction is the class that corresponds to the highest average probability. This method was chosen over hard (majority) voting as it leverages more information from each model—not just its final decision, but also its confidence in that decision. This often leads to superior performance, particularly when the constituent models are well-calibrated.

To determine the optimal composition of the ensemble, several configurations of increasing size were systematically constructed. This was done by progressively adding the best-performing individual models in a greedy, forward-selection manner. This approach allows for a clear analysis of how performance scales with ensemble size and helps identify the point at which adding more models ceases to provide a benefit, ensuring a final ensemble that is both accurate and as efficient as possible.

To ensure a fair comparison and a consistent foundation for the ensemble, all 11 base models were trained independently but under an identical training protocol. The optimization was guided by the **Binary Cross-Entropy loss function**, which is the standard and mathematically appropriate choice for a binary classification task. The **Adam optimizer** was used for its adaptive learning rate capabilities and fast convergence. Each model was trained for a maximum of **64 epochs**, with an **early stopping** mechanism based on the validation set's performance to prevent overfitting and save the best-performing version of each model. To accommodate the diverse set of selected architectures, all input Log-Mel spectrograms were resized to a uniform dimension of 224224 **pixels**. This specific size is the standard input dimension for many of the ImageNet-pre-trained models used in this study (e.g., VGG, ResNet), ensuring compatibility across the entire model portfolio.

3.3.3 Results and Analysis

3.3.3.1 Individual Model Performance

The independent training and evaluation of the 11 diverse CNN architectures served a dual purpose: first, to assess their individual merits on the dimensional classification task, and second, to generate a pool of base learners for the subsequent ensemble experiments. The performance of each model on the held-out test set is summarized in Table 3.8. The analysis reveals a stark

stratification in performance, providing critical insights into architectural suitability for this ASER task.

A clear group of **top-performing models** emerged, led by the custom-designed **DeepSpecCNN**, which achieved the highest accuracy of any single model at **74.75%**. Notably, the classic and relatively simple **LeNet** architecture also performed exceptionally well (74.57% accuracy), suggesting that for this well-defined binary task, a direct and less complex architecture is highly effective at capturing the most salient spectro-temporal features.

Conversely, a distinct group of modern, lightweight models designed for computational efficiency were **significant underperformers**. Architectures such as **MobileNet**, **SqueezeNet**, and **EfficientNetB0** failed to learn the task effectively, achieving accuracies near 52%, which is only slightly above the chance baseline given the class distribution. The extremely low precision scores (~26%) for these models indicate a probable model collapse, where they were unable to learn discriminative features and defaulted to a simplistic prediction strategy. A strong hypothesis for this failure is that the aggressive parameter-reduction techniques central to these models (e.g., depthwise separable convolutions) may discard the subtle, fine-grained harmonic and textural information within the spectrograms that is crucial for distinguishing between the nuanced vocal states of high and low control.

Table 3.8: Performance evaluation of the individual CNN models on the binary High/Low Control task.

Model	**Metrics (%)**			
	Accuracy	**Precision**	**F1-Score**	**Recall**
DeepSpecCNN (Ours)	**74.75**	74.70	74.70	74.69
LeNet	74.57	74.55	74.55	74.60
AlexNet	72.87	72.85	72.75	72.72
DenseNet121	70.99	71.80	70.91	71.36
ResNet50	70.54	71.10	70.50	70.84
VGG16	70.18	70.24	70.18	70.27
VGG19	70.09	70.09	69.92	69.89
ZFNet	68.84	68.81	68.81	68.84
MobileNet	52.19	26.09	34.29	50.00
EfficientNetB0	52.19	26.09	34.29	50.00
SqueezeNet	52.19	26.09	34.29	50.00

Deeper insights are revealed by the training dynamics, illustrated by the validation accuracy and loss curves in Figure 3.4 and Figure 3.5, respectively. The top-performing models like DeepSpecCNN and LeNet can be characterized as high-variance “specialists”; they learn very quickly and achieve high peak accuracy, but their validation loss begins to increase after 12-15

epochs, signaling a rapid onset of overfitting. In contrast, more complex models with advanced regularization properties, like **DenseNet121** and **ResNet50**, behave as more stable "generalists." While their peak accuracy is slightly lower, they maintain a very low and stable validation loss throughout training, suggesting superior generalization. This diversity in learning behavior is not a weakness but is, in fact, the ideal condition for ensembling. It provides a mix of high-accuracy "specialist" models and robust "generalist" models, setting a perfect foundation to test the hypothesis that their combination will yield a more accurate and reliable final classifier.

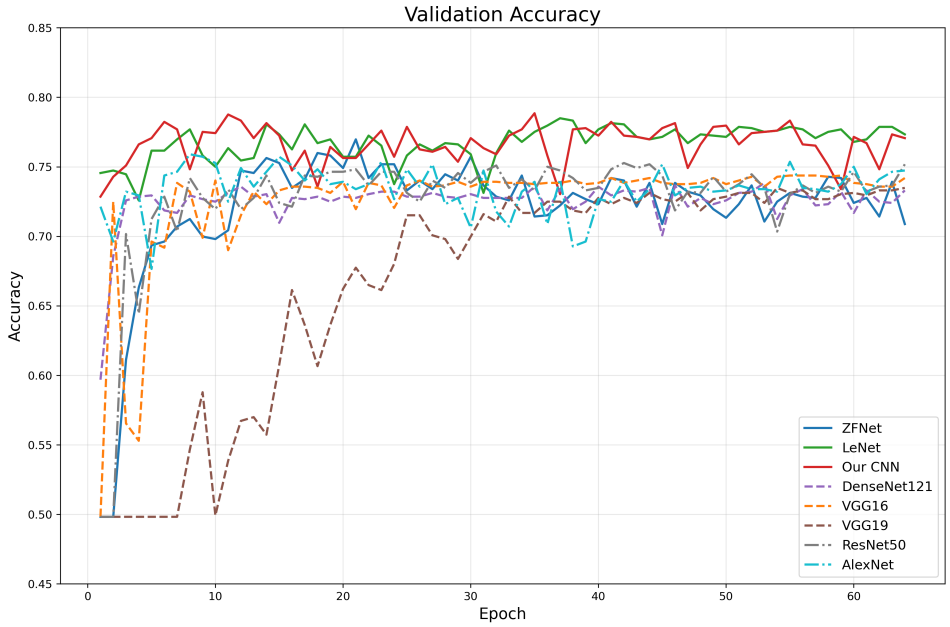


Figure 3.4: Validation accuracy for each individual model over 64 epochs.

3.3.3.2 Ensemble Model Performance

The core finding of this study, summarized in Table 3.9, is the validation of our primary hypothesis: that a strategically constructed ensemble of diverse models can significantly outperform any single architecture. By combining the predictions of the individual models via **soft voting**, a substantial boost in performance was immediately realized. A 3-model ensemble comprising the top performers (DeepSpecCNN, LeNet, AlexNet) already achieved an accuracy of 77.26%, a significant leap of nearly 2.5 percentage points over the best individual model.

The peak performance was achieved with the **5-model ensemble**, which reached a final test accuracy of **77.70%**. The composition of this optimal ensemble provides a compelling insight into the power of diversity. The peak performance was not achieved by simply aggregating the five best models, but by combining the high-accuracy, high-variance "specialist" models

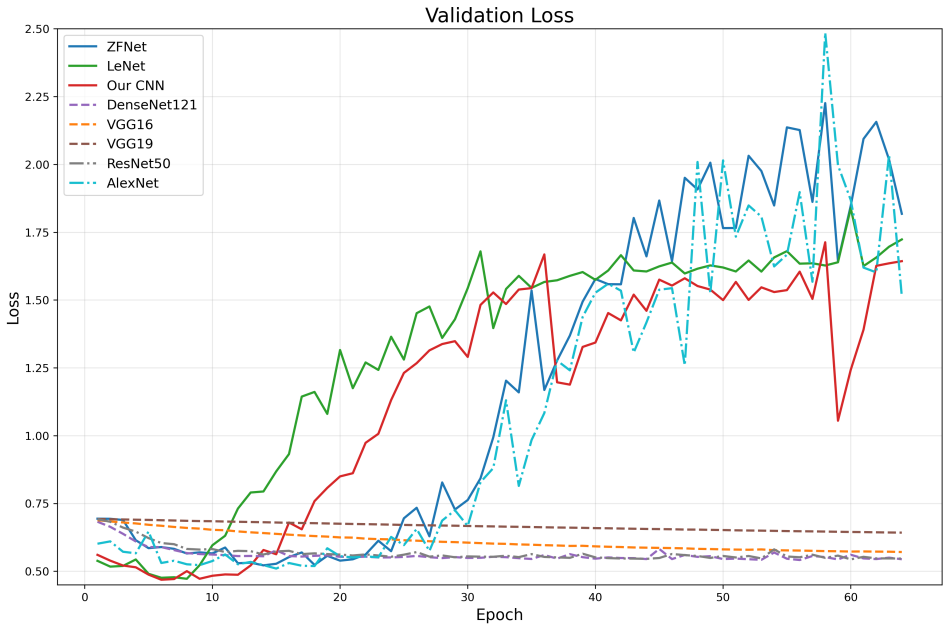


Figure 3.5: Validation loss for each individual model over 64 epochs.

(DeepSpecCNN, LeNet, AlexNet) with the two most stable and robust "generalist" models (DenseNet121 and ResNet50). This suggests an optimal synergy: the well-regularized predictions of the generalists likely tempered the overfitting tendencies of the specialists, creating a more balanced and robust collective decision. The confusion matrix for this 5-model ensemble, shown in Figure 3.6, confirms its highly balanced predictive performance for the two macro-classes.

Crucially, the study also demonstrated that the benefits of ensembling are not limitless. Adding more, weaker models beyond this 5-member committee led to a slight but consistent degradation in performance. This is because including less accurate or poorly-calibrated models in a soft voting scheme can introduce more noise than signal, diluting the high-confidence predictions of the stronger members. This result underscores a key principle of ensemble construction: strategic, selective ensembling that prioritizes both model quality and diversity is more effective than indiscriminately aggregating all available models. The 5-model configuration therefore represents the empirically-determined optimal balance for this task.

3.3.3.3 Comparison with State-of-the-Art

The culminating result of this study is the establishment of a new state-of-the-art accuracy on the CREMA-D dataset for this classification task. As detailed in the comparative evaluation in Table 3.10, the proposed **5-model ensemble** achieved a final accuracy of **77.70%**. This

Table 3.9: Performance metrics for the ensemble models of increasing size.

Metric	3-Model (%)	5-Model (%)	7-Model (%)	All-Model (%)
Precision	77.22	75.70	77.51	76.10
Recall	77.24	78.70	77.56	76.15
F1 Score	77.23	77.20	77.51	76.08
Accuracy	77.26	**77.70**	77.52	76.09

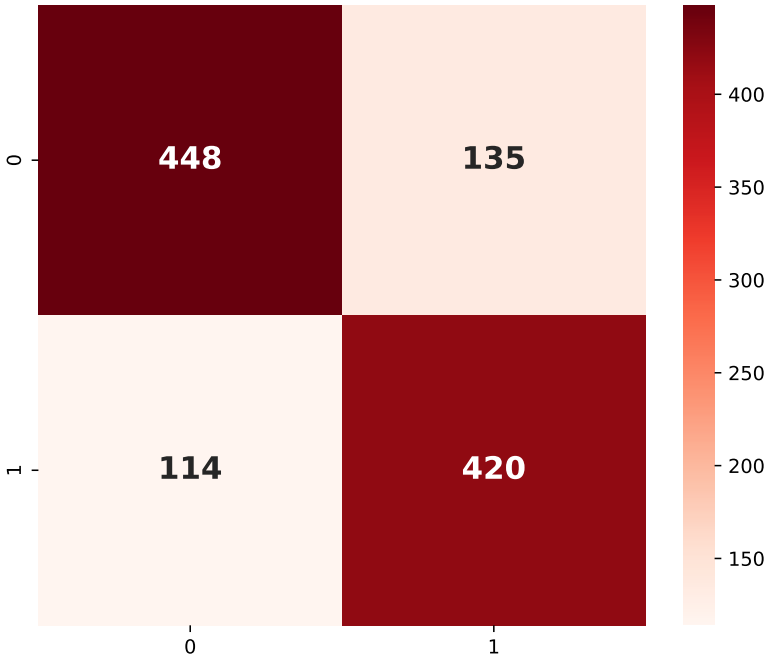


Figure 3.6: The confusion matrix for the 5-model ensemble, which achieved the highest accuracy of 77.70% on the binary High/Low Control task.

represents a substantial advance over previously reported methods, marking a **5 percentage point absolute increase** over the prior leading ensemble approach (ResNet-18 with SPEL at 72.70%), a relative improvement of nearly 7%.

This significant performance gain is not attributable to a single architectural tweak but to the synergistic effect of the two core strategies employed in this study. First, a major contribution to this result is the **psychologically-grounded task reformulation**. By shifting from a noisy 6-class categorical problem to a more robust binary task based on the 'control' dimension, we provided the models with a cleaner, more discriminative learning signal. This fundamental change in the problem definition likely raised the potential performance ceiling for any given architecture.

Second, the strategic use of **architectural diversity** within the ensemble proved superior to

other approaches. Unlike methods that ensemble variations of a single model, our framework combines fundamentally different CNN families—from high-variance "specialists" like LeNet to stable "generalists" like DenseNet121. This curated diversity ensured that the constituent models learned complementary representations of the emotional speech, leading to a more powerful and reliable collective decision.

In conclusion, this result validates the central hypothesis of Study 2. The combination of a dimensional emotion model that simplifies the classification task and a strategically diverse ensemble that maximizes predictive power provides a superior framework for achieving high-accuracy speech emotion recognition.

Table 3.10: A Comparative Performance Evaluation of Our Proposed Ensemble Models Against State-of-the-Art Methods on CREMA-D.

Method	**Approach**	**Accuracy (%)**
Audio Spectrogram Transformer [74]	Transformer-based global attention	67.81
SpectoResNet [75]	CNN with residual connections	65.20
SepTr with LeRaC [71]	Separable transformer with learning curriculum	70.95
ResNet-18 with SPEL [73]	Ensemble with self-paced learning	72.70
Our DeepSpecCNN (Individual)	Custom CNN for spectral analysis	74.75
Our 5-Model Ensemble	**Strategic Ensemble with Majority Voting**	**77.70**

3.3.4 Discussion of Study 2

The findings of this second study provide a strong and unequivocal answer to our second research question. The results demonstrate that the proposed two-pronged strategy—reframing the ASER task using a **psychologically-grounded dimensional model** and applying a **strategic ensemble of diverse architectures**—is a highly effective method for maximizing classification accuracy. The fact that the individual DeepSpecCNN model already surpassed existing state-of-the-art methods, and that the 5-model ensemble pushed this boundary even further to achieve a new benchmark accuracy of **77.70%**, validates the efficacy of this novel framework.

The success of this approach can be deconstructed into two synergistic contributions. First, the task reformulation based on the GWE's control dimension was a foundational element. By simplifying the problem from a noisy 6-class categorical task to a more robust and theoretically-grounded binary one, we effectively "cleaned" the learning signal, providing the models with a less ambiguous and more acoustically consistent target. This highlights a critical, often-

overlooked principle: integrating insights from psychological theory directly into the problem definition can be as impactful as purely architectural innovations. Second, the success of the selective ensemble design confirmed that it is the *diversity* of the constituent models, not just their quantity, that drives performance gains. The optimal 5-model ensemble combined high-variance "specialists" with stable "generalists," creating a robust committee that leveraged the complementary strengths of each member.

This study, when contrasted with Study 1, illuminates a crucial trade-off in applied ASER research. The CBAM-DenseNet121 model from Study 1 represents the **"practical"** solution: a single, efficient architecture optimized for deployment in resource-constrained environments. The ensemble framework from this study represents the **"maximal performance"** solution: a more computationally complex approach designed to achieve the highest possible accuracy, suitable for offline analysis or high-stakes applications where performance is the paramount concern. Together, these studies provide two distinct and valuable contributions to the field.

However, it is essential to acknowledge the context of this achievement. This new state-of-the-art was established on the CREMA-D dataset—a corpus of acted, acoustically clean speech. While this is a critical step for validating the framework under controlled conditions, its robustness and performance on genuine, "in-the-wild" spontaneous speech remain unproven. This work on a benchmark dataset, therefore, serves as a crucial but preliminary phase. It has yielded a powerful set of validated techniques, paving the way for the next and most critical challenge: applying these insights to the noisy, complex, and culturally specific domain of spontaneous dialectal speech, which will be the central focus of the next chapter.

3.4 Conclusion

This chapter presented a comprehensive and systematic investigation of novel deep learning frameworks on the benchmark CREMA-D dataset, successfully achieving two distinct but equally important goals in ASER research. The first core contribution, detailed in **Study 1**, was the development and validation of the **CBAM-DenseNet121**, an efficient, lightweight architecture. This study demonstrated that a carefully designed, attention-enhanced CNN can resolve the critical trade-off between accuracy and computational efficiency, yielding a practical model suitable for real-world deployment. The second major contribution, presented in **Study 2**, was a novel framework that achieved a new state-of-the-art accuracy on the dataset. This was accomplished not merely through architectural improvements, but by synergistically combining a psychologically-grounded task reformulation with the collective predictive power of a strategic, architecturally diverse ensemble of CNNs.

Taken together, the findings from these two studies provide a holistic and insightful view of modern deep learning approaches to ASER. They deliver strong empirical evidence for the value of targeted attention mechanisms in refining feature representations for complex acoustic signals. Furthermore, they powerfully illustrate the capacity of ensemble learning to break through the performance ceilings of single-model systems, confirming that curated diversity is a key driver of state-of-the-art performance.

Having developed and rigorously validated this powerful toolkit of modeling techniques on the clean, controlled proving ground of a benchmark dataset, the research has reached a critical juncture. The logical and necessary next step is to test the true robustness and applicability of these advanced methods. The subsequent chapter will therefore pivot from the idealized laboratory setting to confront the far more challenging and ecologically valid domain of spontaneous, noisy, and dialectal speech, applying these learnings to the novel OYH Algerian Arabic corpus.

CHAPTER 4

EXPERIMENTS ON THE NOVEL OYH ALGERIAN ARABIC CORPUS

Contents

4.1 Introduction

Having developed and validated advanced deep learning frameworks on a standardized benchmark dataset in the previous chapter, this chapter now confronts one of the most significant challenges in modern ASER: the domain gap between clean, acted laboratory data and noisy, spontaneous, real-world speech. As established in Chapter 1, a critical barrier to progress, particularly for non-English languages, is the scarcity of suitable corpora that reflect authentic emotional expression.

This chapter details a study designed to address this critical resource gap for the Arabic language, with a specific focus on the under-resourced Algerian dialect. The primary objective is twofold. First, this chapter introduces and meticulously documents the creation, annotation, and composition of the **Open Your Heart (OYH) corpus**, a novel and substantial dataset of spontaneous emotional speech collected from a real-world television talk show. This contribution provides a valuable new resource to the research community for studying genuine emotional expression in a complex linguistic environment.

Second, the chapter aims to establish a comprehensive and robust performance baseline on this new corpus. To achieve this, a suite of eleven traditional machine learning models are systematically evaluated. This study moves beyond simple classification to investigate the crucial role of data preprocessing and feature engineering, employing a sophisticated, iterative feature selection algorithm, the **Backward Elimination of All Worst Features (BE-AWF)**, to identify the most discriminative acoustic features for the dimensional emotion tasks of valence and control. The findings presented herein not only quantify the performance of classical ASER techniques on this challenging new dataset but also provide critical insights into the acoustic correlates of emotion in spontaneous Algerian Arabic speech.

4.2 The OYH Corpus: A Novel Dataset for Algerian Arabic

The creation of a new dataset is a significant undertaking, motivated by a clear and pressing need within the research community. The OYH corpus was developed to address the limitations of existing resources and to provide a foundation for future research in Arabic ASER.

4.2.1 Motivation and Data Collection

The impetus for creating the OYH corpus stems directly from two critical, intersecting research gaps identified in Chapter 1: the **domain gap** between acted and spontaneous speech, and the **data scarcity gap** for under-resourced languages. The vast majority of ASER research

has relied on acted datasets recorded in sterile, acoustically controlled environments. While valuable for benchmarking, this has produced models with poor generalizability to the complexities of authentic, "in-the-wild" human interaction. This problem is severely compounded for the Arabic language, where immense dialectal variation means models trained on Modern Standard Arabic (MSA) or one regional dialect fail to capture the unique phonetic and prosodic nuances of another.

The Algerian dialect, "Darija," is a prime example of this challenge. It is not a monolith but a complex continuum of local dialects heavily influenced by Berber and French, making it a particularly under-resourced linguistic environment (as shown in Figure 1.2). To address this gap, data was sourced from a place of genuine emotional expression: the popular and culturally significant Algerian television talk show, **"Open Your Heart" (*Eftah Qalbak*)**. This program was uniquely suited for this research due to its format, which centers on mediating deeply personal issues like family reconciliations and lost friendships. This context naturally elicits a wide spectrum of powerful, unscripted emotions from its participants.

The unscripted dialogues provide a rich source of **ecologically valid** data. This means the corpus captures emotional speech as it occurs in a naturalistic social setting, complete with the hesitations, emotional blends, and subtle paralinguistic cues that are characteristic of genuine communication and starkly absent in the often-exaggerated portrayals found in acted corpora.

Data was collected by recording 14 full episodes of the show that aired between October 2016 and April 2019. This "in-the-wild" collection method presents a classic methodological trade-off between **control and authenticity**. While laboratory recordings offer high control over emotional balance and acoustic quality, they do so at the cost of authenticity. Conversely, our approach sacrifices control over the emotional content and recording environment to gain unparalleled authenticity. For the primary objective of this research—to build a corpus that reflects the true nature of spontaneous emotional speech—this trade-off was a deliberate and necessary choice.

4.2.2 Dataset Statistics and Segmentation

The resulting OYH dataset is a substantial collection, comprising approximately **6.3 hours** of audio recordings, segmented into **6,167 individual clips**. The data features 43 unique speakers (27 male, 16 female) with a diverse age range from 21 to 81 years old. A detailed breakdown of the dataset's composition is provided in Figure 4.1.

The initial recordings were processed to create a standardized, high-quality dataset. All audio extracts were converted to a mono channel format and downsampled to a **16 kHz sampling rate** to balance audio fidelity with computational efficiency. An advanced normalization process,

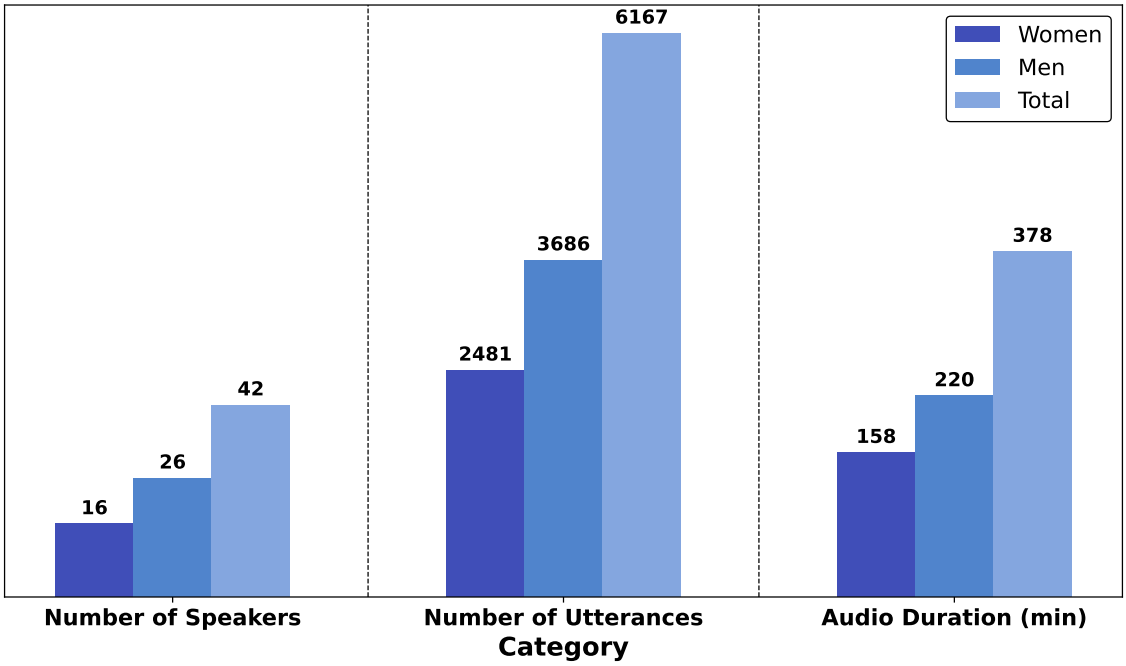


Figure 4.1: Statistical overview of the OYH dataset, showing the distribution of audio files and total duration by speaker gender.

using the Min-Max scaling approach, was applied to ensure consistent acoustic characteristics and equalize volume across all clips.

The segmentation process involved several stages. Full one-hour recordings were first divided into smaller segments containing a single thematic discussion. These video clips were then manually segmented into distinct turns of speech, and subsequently into individual sentences for fine-grained analysis. To ensure data quality, any utterances that were distorted by background applause, music, or other significant disruptions were excluded from the final dataset. The speakers were categorized into four groups: guests (G), moderators (M), main speakers (P), and supporting speakers (S). This classification allows for a more nuanced analysis of the interaction dynamics. Figure 4.2 shows the distribution of utterances and speakers across these categories, highlighting that the majority of the data comes from the main guest speakers.

4.2.3 Annotation Methodology

A core contribution of the OYH corpus is its nuanced annotation, designed to move beyond the limitations of simplistic categorical labels. To achieve this, the dataset was annotated using the dimensional framework of the **Geneva Wheel of Emotions (GWE)**, as detailed in Chapter 1 and shown again in Figure 1.1. This model was specifically chosen because its two primary

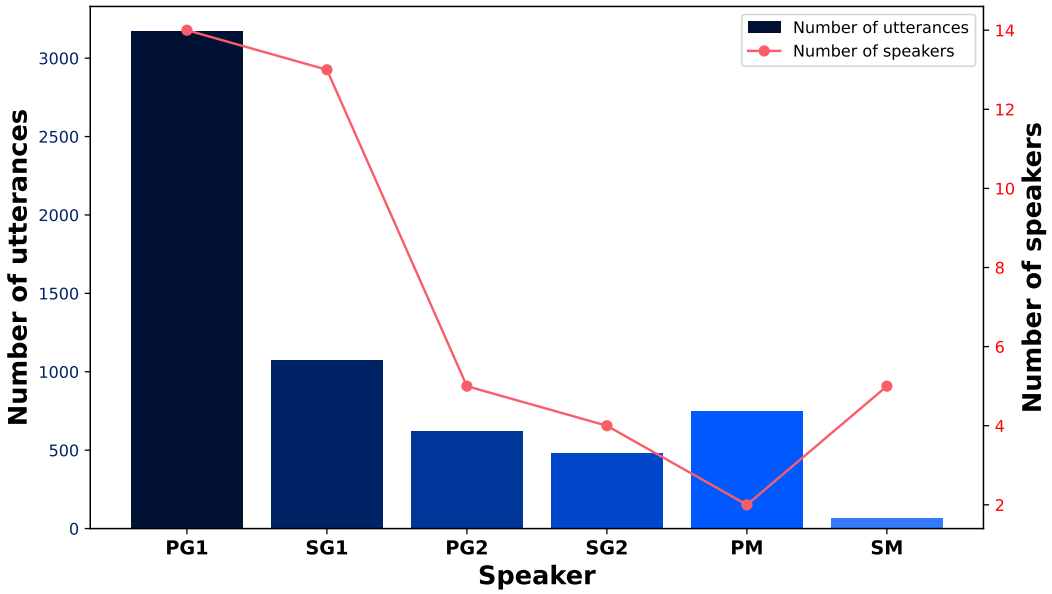


Figure 4.2: Number of utterances and unique speakers per speaker category within the OYH dataset.

axes, **valence** (the pleasantness of an emotion) and **control** (the sense of agency or power), are particularly well-suited for speech analysis, as the control dimension has strong, direct acoustic correlates.

The annotation was a meticulous, multi-stage process performed by a team of three trained annotators, all of whom are **native speakers of Algerian Arabic**. This was essential to ensure that the subtle linguistic and cultural-specific cues within the speech were accurately interpreted. For each of the 6,167 audio clips, each annotator independently assigned a continuous score on a scale from -5 (highly negative/low control) to +5 (highly positive/high control) for both the valence and control dimensions. To ensure reliability, the final score for each clip was derived by averaging the three ratings. In cases of significant disagreement between annotators (defined as a standard deviation greater than 1.5), the raters convened to discuss the specific clip and reach a consensus, thereby strengthening the consistency and quality of the final ground-truth labels.

To frame the problem for standard machine learning classifiers, the resulting continuous scores were then discretized into three conceptually meaningful classes for each dimension: **Low, Medium, and High**. This three-level scheme provides a good balance between capturing emotional nuance and ensuring that the classes are sufficiently distinct for a classification task. The thresholds were set as follows: scores in the range [-5.0, -1.5] were mapped to 'Low',

scores between (-1.5, 1.5) were mapped to 'Medium', and scores in the range [1.5, 5.0] were mapped to 'High'.

The final distribution of utterances, shown in Figure 4.3, reveals a natural class imbalance. This is not an artifact of the binning process but rather an authentic reflection of the corpus's content; the prevalence of 'Medium' valence (49%) and 'High' control (41.2%) is characteristic of the often intense and emotionally charged but varied nature of the talk-show dialogues. This inherent imbalance makes the OYH corpus a more realistic and challenging benchmark for ASER research. As will be detailed in the methodology, this imbalance was addressed during the experimental phase through up-sampling techniques.

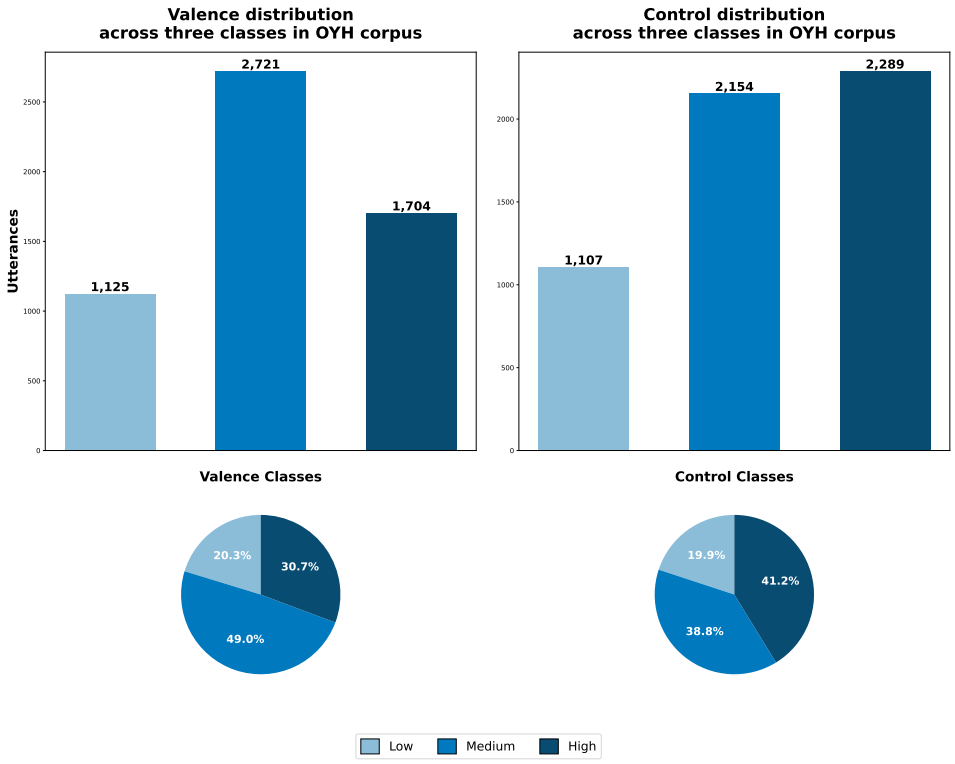


Figure 4.3: Distribution of utterances across the three discretized classes (Low, Medium, High) for the Valence and Control dimensions in the OYH dataset.

Table 4.1: Mapping of emotion families from the Geneva Wheel of Emotions to their corresponding valence and control scores. The columns labeled 1 through 5 represent five increasing levels of emotional intensity for each family.

((a))

Emotions with High Control and Negative Valence.

Emotions	Levels	1	2	3	4	5
Irritation	Valence	-0.25	-0.5	-0.75	-0.75	-1
Anger	Control	2	2.5	3	4	5
Contempt	Valence	-1	-1.25	-1.5	-1.75	-2
Scorn	Control	1.75	2.25	2.75	3.5	4.25
Disgust	Valence	-1.5	-1.75	-2.25	-2.75	-3.25
Repulsion	Control	1.5	1.75	2.25	2.75	3.25
Envy	Valence	-1.75	-2.25	-2.75	-3.5	-4.25
Jealousy	Control	1	1.25	1.5	1.75	2
Disappointment	Valence	-2	-2.5	-3	-4	-5
Regret	Control	0.25	0.5	0.75	0.75	1

((b))

Emotions with High Control and Positive Valence.

Emotions	Levels	1	2	3	4	5
Involvement	Valence	0.25	0.5	0.75	0.75	1
Interest	Control	2	2.5	3	4	5
Amusement	Valence	1	1.25	1.5	1.75	2
Laughter	Control	1.75	2.25	2.75	3.5	4.25
Pride	Valence	1.5	1.75	2.25	2.75	3.25
Elation	Control	1.5	1.75	2.25	2.75	3.25
Happiness	Valence	1.75	2.25	2.75	3.5	4.25
Joy	Control	1	1.25	1.5	1.75	2
Enjoyment	Valence	2	2.5	3	4	5
Pleasure	Control	0.25	0.5	0.75	0.75	1

((c))

Emotions with Low Control and Negative Valence.

Emotions	Levels	1	2	3	4	5
Guilt	Valence	-2	-2.5	-3	-4	-5
Remorse	Control	-0.25	-0.5	-0.75	-0.75	-1
Embarrassment	Valence	-1.75	-2.25	-2.75	-3.5	-4.25
Shame	Control	-1	-1.25	-1.5	-1.75	-2
Worry	Valence	-1.5	-1.75	-2.25	-2.75	-3.25
Fear	Control	-1.5	-1.75	-2.25	-2.75	-3.25
Sadness	Valence	-1	-1.25	-1.5	-1.75	-2
Despair	Control	-1.75	-2.25	-2.75	-3.5	-4.25
Pity	Valence	-0.25	-0.5	-0.75	-0.75	-1
Compassion	Control	-2	-2.5	-3	-4	-5

((d))

Emotions with Low Control and Positive Valence.

Emotions	Levels	1	2	3	4	5
Tenderness	Valence	2	2.5	3	4	5
Felling love	Control	-0.25	-0.5	-0.75	-0.75	-1
Wonderment	Valence	1.75	2.25	2.75	3.5	4.25
Felling awe	Control	-1	-1.25	-1.5	-1.75	-2
Disburdened	Valence	1.5	1.75	2.25	2.75	3.25
Relief	Control	-1.5	-1.75	-2.25	-2.75	-3.25
Astonishment	Valence	1	1.25	1.5	1.75	2
Surprise	Control	-1.75	-2.25	-2.75	-3.5	-4.25
Longing	Valence	0.25	0.5	0.75	0.75	1
Nostalgia	Control	-2	-2.5	-3	-4	-5

4.3 Experimental Methodology

This section details the specific methods used to establish a machine learning baseline on the OYH corpus.

4.3.1 Acoustic Feature Extraction and Selection

To establish a robust and interpretable baseline on the novel OYH corpus, this study employed the traditional handcrafted feature paradigm. This methodological choice was deliberate. While end-to-end deep learning models, as explored in the previous chapter, are powerful, their features are often uninterpretable. The handcrafted approach, in contrast, allows for an explanatory analysis to identify the specific, physically-meaningful acoustic properties that are most salient for expressing emotion in spontaneous Algerian Arabic—a key scientific goal when introducing a new corpus for an under-resourced dialect.

Following this paradigm, a comprehensive set of **6,373 acoustic features** was extracted from each audio clip using the renowned **openSMILE** toolkit [35] with its standard **ComParE 2013** configuration file. This feature set is designed to be exhaustive, capturing a wide range of acoustic phenomena including prosodics (fundamental frequency, energy contours), voice quality (jitter, shimmer, harmonic-to-noise ratio), and spectral characteristics (Mel-Frequency Cepstral Coefficients (MFCCs), spectral flux, formants).

However, working with such a high-dimensional feature set introduces the well-known **"curse of dimensionality,"** which increases the risk of model overfitting, raises computational costs, and demands larger amounts of training data. To mitigate this, a robust feature selection process was essential to distill this large set into a more compact, informative, and generalizable subset. For this purpose, we employed **wrapper-based feature selection methods**. Unlike simpler filter methods that evaluate features independently of a model, wrapper methods use the performance of a specific classifier to score and select feature subsets. While more computationally intensive, this approach is often superior as it finds features that are optimally suited for the chosen learning algorithm.

Two specific backward elimination algorithms were used, as illustrated in Figure 4.4 and Figure 4.5. The first is the standard **Backward Feature Elimination (BFE)**, an iterative algorithm that starts with the full feature set and greedily removes one feature at a time based on which removal least degrades (or most improves) classifier performance. The second, an enhanced version developed for this work, is the **Backward Elimination of All Worst Features (BE-AWF)**. This approach improves upon BFE by being less greedy; at each iteration, it re-evaluates the entire remaining feature set to remove the feature that has the least overall impact

on model performance. This more exhaustive search aims to find a more robust and globally optimal feature subset. For both algorithms, a Support Vector Machine (SVM) was used as the core classifier to guide the selection process.

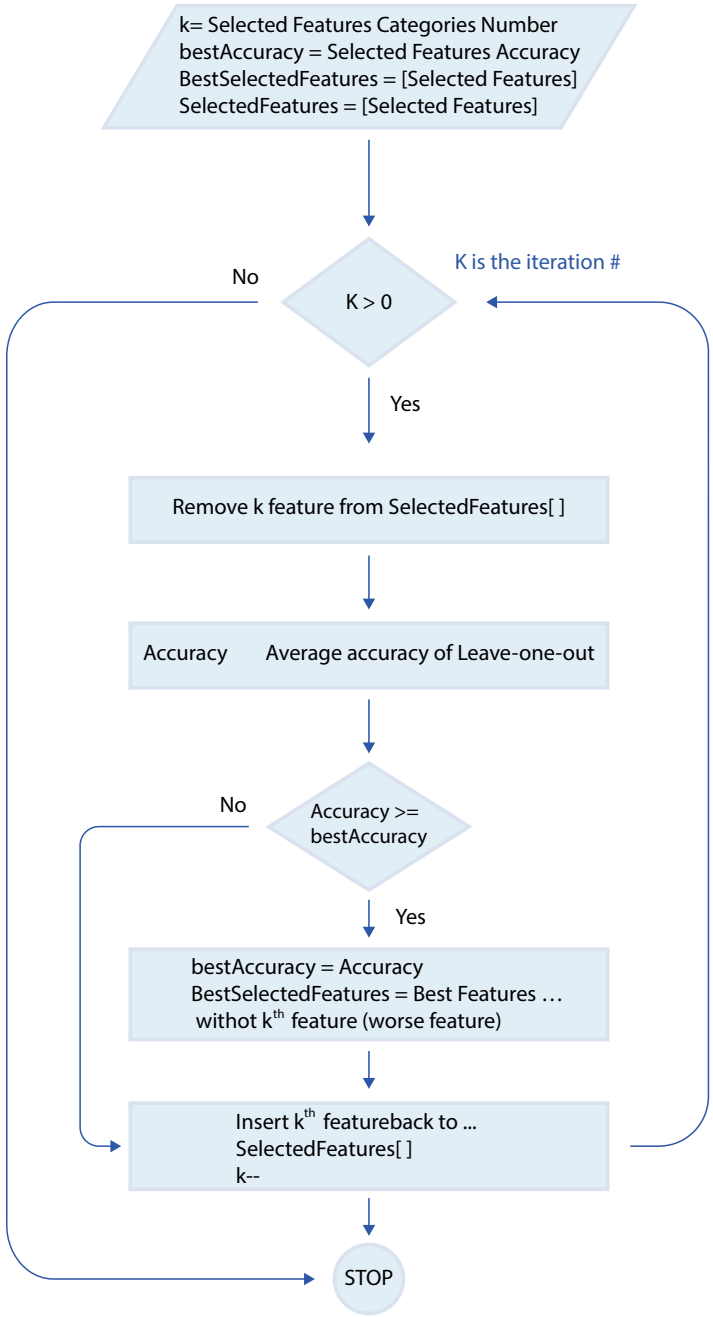


Figure 4.4: Flowchart illustrating the standard Backward Feature Elimination (BFE) algorithm used for feature selection.

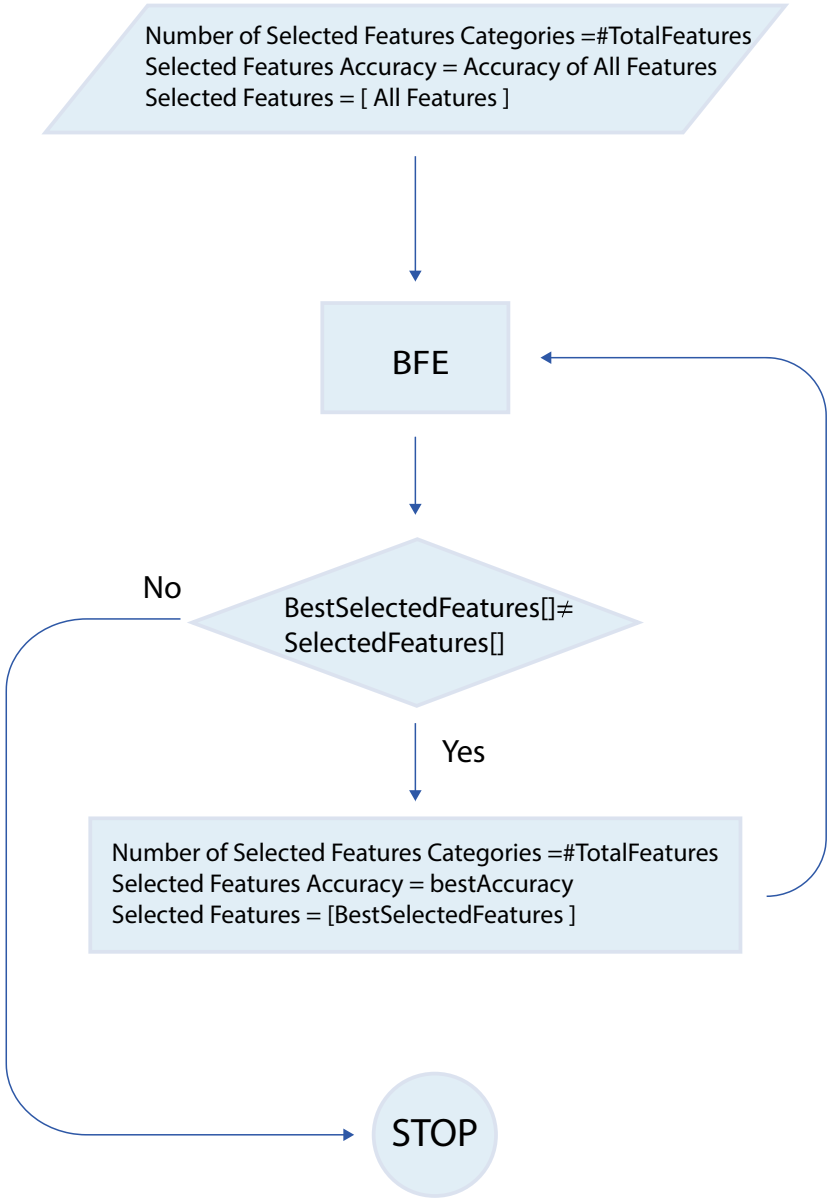


Figure 4.5: Flowchart illustrating the enhanced Backward Elimination of All Worst Features (BE-AWF) algorithm developed for feature selection.

4.3.2 Classification Models

To establish a comprehensive performance baseline on the OYH corpus, a wide-ranging comparative analysis was conducted using a suite of eleven traditional machine learning

classifiers. The selection was deliberately diverse, spanning several major families of algorithms to determine which class of model is most effective for this novel task. The chosen models, whose theoretical principles were detailed in Chapter 2, include:

- **Kernel-based Methods:** A **Support Vector Machine (SVM)**, renowned for its ability to find optimal separating hyperplanes in high-dimensional feature spaces.
- **Instance-based Methods: K-Nearest Neighbors (KNN)**, a simple yet effective non-parametric model that serves as a strong baseline.
- **Ensemble Methods (Bagging):** The **Random Forest** classifier, a powerful model known for its robustness against overfitting by averaging the predictions of many decorrelated decision trees.
- **Ensemble Methods (Boosting):** A comprehensive set of state-of-the-art gradient boosting models, including **AdaBoost**, **Gradient Boosting Machines (GBM)**, **XGBoost**, **LightGBM**, and **CatBoost**. This family of models is often dominant in tasks involving tabular or structured data and was included to test the upper limits of performance with classical methods.

This systematic evaluation across different algorithmic families ensures a thorough and unbiased assessment, aimed at identifying the most suitable and highest-performing models for dimensional emotion recognition on spontaneous Algerian Arabic speech.

4.3.3 Data Partitioning and Preprocessing

A rigorous protocol for data handling was established to ensure the validity and reproducibility of the experimental results. The OYH dataset, containing 6,167 clips from 43 speakers, was methodically partitioned into three subsets with strict **speaker independence**. This means all clips from any given speaker were confined to a single set, which is a critical step to ensure the model is evaluated on its ability to generalize to new, unseen speakers, not on its ability to recognize speakers it was trained on. The final split was **70% for the training set, 20% for the development (validation) set, and 10% for the test set**.

Two critical preprocessing steps were then applied to the feature sets to prepare them for the classifiers.

First, to address the natural class imbalance documented in the previous section (and shown in Figure **??**), a strategy of **up-sampling** was employed. Standard classifiers are often biased towards the majority class, as their optimization objective is typically to maximize overall

accuracy. To counteract this, instances from the minority classes were randomly duplicated in the training set to create a balanced class distribution for the model to learn from. Crucially, this up-sampling was **applied only to the training data**, ensuring that the validation and test sets remained in their original, natural distribution for an unbiased evaluation of the model's real-world performance.

Second, all acoustic features were normalized using **Min-Max scaling**. This is an essential step for the many machine learning algorithms (like SVM and KNN) whose performance is dependent on the scale of the input features. Normalization transforms every feature to a common range of [0, 1], ensuring that features with large value ranges (like energy) do not disproportionately influence the model's learning process over features with smaller ranges (like jitter). The profound and necessary impact of this step is demonstrated empirically in Table 4.2 and Table 4.3. As shown, normalization dramatically improved the SVM's F1-score on the valence task from a near-random 16.96% to a meaningful 43.34%, confirming its status as an indispensable part of the preprocessing pipeline.

Table 4.2: Impact of Normalization on SVM performance for the Valence task.

complexity	Without Normalization		With Normalization	
	Accuracy (%)	F1 score (%)	Accuracy (%)	F1 score (%)
1e-4	32.11	16.96	42.57	41.38
1e-3	32.11	16.96	41.11	42.5
1e-2	32.11	16.96	40.79	41.72
1e-1	32.11	16.96	41.11	42.08
1	32.11	16.96	42.66	43.34

Table 4.3: Impact of Normalization on SVM performance for the Control task.

complexity	Without Normalization		With Normalization	
	Accuracy (%)	F1 score (%)	Accuracy (%)	F1 score (%)
1e-4	46.55	33.46	48.74	46.74
1e-3	46.55	33.46	49.96	48.4
1e-2	46.55	33.46	45.5	43.79
1e-1	46.55	33.46	43.14	42.09
1	46.55	33.46	43.8	43.59

4.4 Results and Analysis

4.4.1 Impact of Feature Selection and Acoustic Correlates

The application of the BE-AWF feature selection algorithm was a critical step, not only for model optimization but also as an analytical tool to uncover the most salient acoustic correlates of emotion in spontaneous Algerian Arabic. The process successfully distilled the vast 6,373-feature set into compact, more powerful subsets and, most significantly, provided strong empirical evidence that the dimensions of valence and control have distinct acoustic footprints.

Analysis for the Valence Dimension

As illustrated in Figure 4.6, the optimal feature subset for predicting **valence** was found to be a combination of **F0, HNR, MFCCs, and general spectral features**. This is a highly informative result. The inclusion of MFCCs and other spectral descriptors, which capture the resonant properties of the vocal tract, suggests that the valence of an emotion is strongly encoded in the overall **timbre and articulatory setting** of the voice. For instance, positive emotions are often associated with a more relaxed vocal tract and clearer articulation, while negative emotions can lead to pharyngeal tension that alters the spectral shape. The retention of **F0** (pitch contour) and **HNR** (voice quality) confirms that prosodic modulation and the degree of breathiness or harshness are also key carriers of valence information in this dialect.

Analysis for the Control Dimension

The analysis for the **control** dimension yielded a different, equally insightful feature set: **F0, HNR, RMS Energy, and ZCR** (Figure 4.7). While F0 and HNR are shared, indicating their fundamental role in affect, the unique inclusion of **RMS Energy** (Root-Mean-Square energy, a direct measure of signal intensity or loudness) and **ZCR** (Zero-Crossing Rate, a correlate of spectral brightness and noisiness) is the critical finding. This strongly implies that the psychological dimension of control and agency is conveyed through more direct physical cues of **vocal effort and physiological activation**. High-control states (e.g., assertiveness, anger) are intuitively linked to higher vocal intensity (high RMS Energy), while low-control states (e.g., sadness, submission) are linked to lower intensity. This clear acoustic separability between the feature sets for valence and control is a major finding of this study. It empirically validates the use of this dimensional framework for ASER, demonstrating that these psychological constructs have distinct physical manifestations in the speech signal.

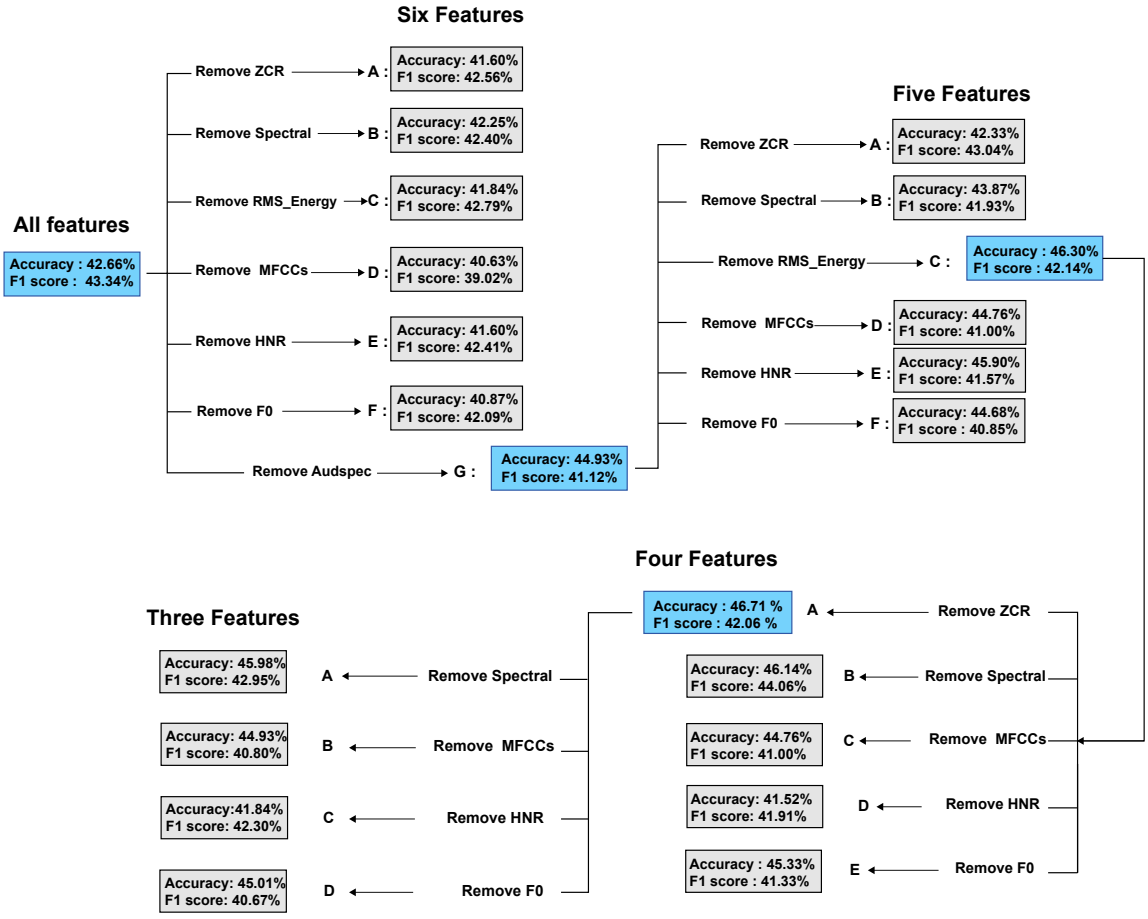


Figure 4.6: Performance of SVM classification on the valence dimension as feature categories are eliminated using the BE-AWF Algorithm.

4.4.2 Classifier Performance for Valence Prediction

Using the optimal feature subset identified for valence, the full suite of eleven machine learning classifiers was evaluated. The comparative results, shown in Figures 4.8 and 4.9, clearly indicate that tree-based ensemble models significantly outperformed other algorithmic families. The **Random Forest** classifier emerged as the top-performing model, achieving a peak accuracy of **58.47%** and an F1-score of **52.41%**. This superiority suggests that the decision boundaries separating the valence classes in this high-dimensional acoustic space are highly complex and non-linear, a challenge for which tree-based ensembles are particularly well-suited.

It is crucial to contextualize this performance. While an accuracy of 58.47% may seem modest compared to results on acted data, it is a strong and significant outcome for a three-class classification task on spontaneous, noisy, "in-the-wild" speech, where the chance baseline is

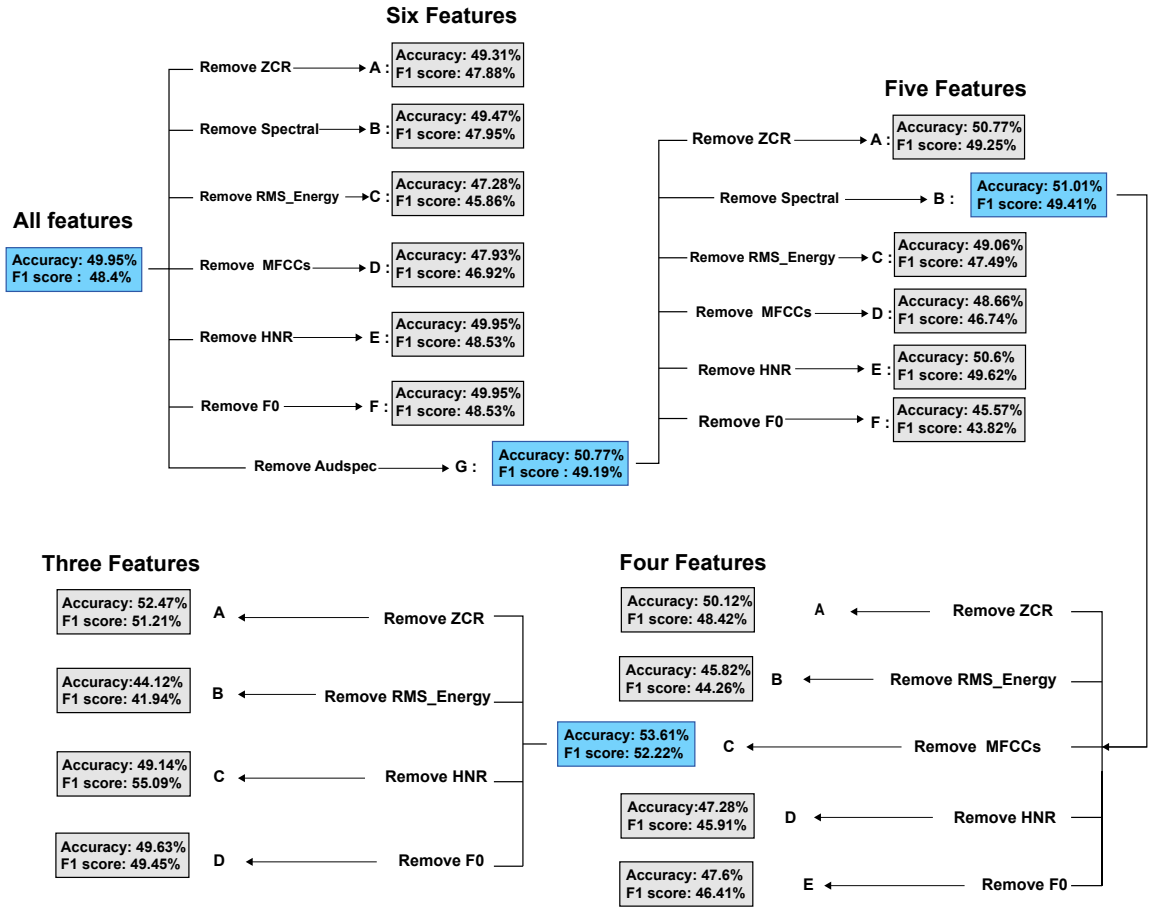


Figure 4.7: Performance of SVM classification on the control dimension as feature categories are eliminated using the BE-AWF Algorithm.

33.3%. This result establishes a robust baseline for this challenging new corpus.

However, a deeper analysis of the best model's confusion matrix (Figure 4.10) reveals a critical challenge that pervades real-world ASER. While the model demonstrates high precision and recall for the majority 'Medium' valence class, its ability to correctly identify the minority 'Low' valence class is severely limited. This finding highlights the profound difficulty of recognizing under-represented emotional states in spontaneous data. Even with up-sampling applied during training, the model struggles, which suggests that the acoustic cues for 'Low' valence are either too subtle and acoustically similar to neutral speech, or that the limited number of unique examples in the original data was insufficient for the model to learn a generalizable pattern.

The detailed performance metrics for the Random Forest classifier across various estimator counts (summarized in Table 4.4) reveal a consistent upward trend in predictive reliability. As the number of estimators increases from 100 to 600, the model achieves its peak F1-score of

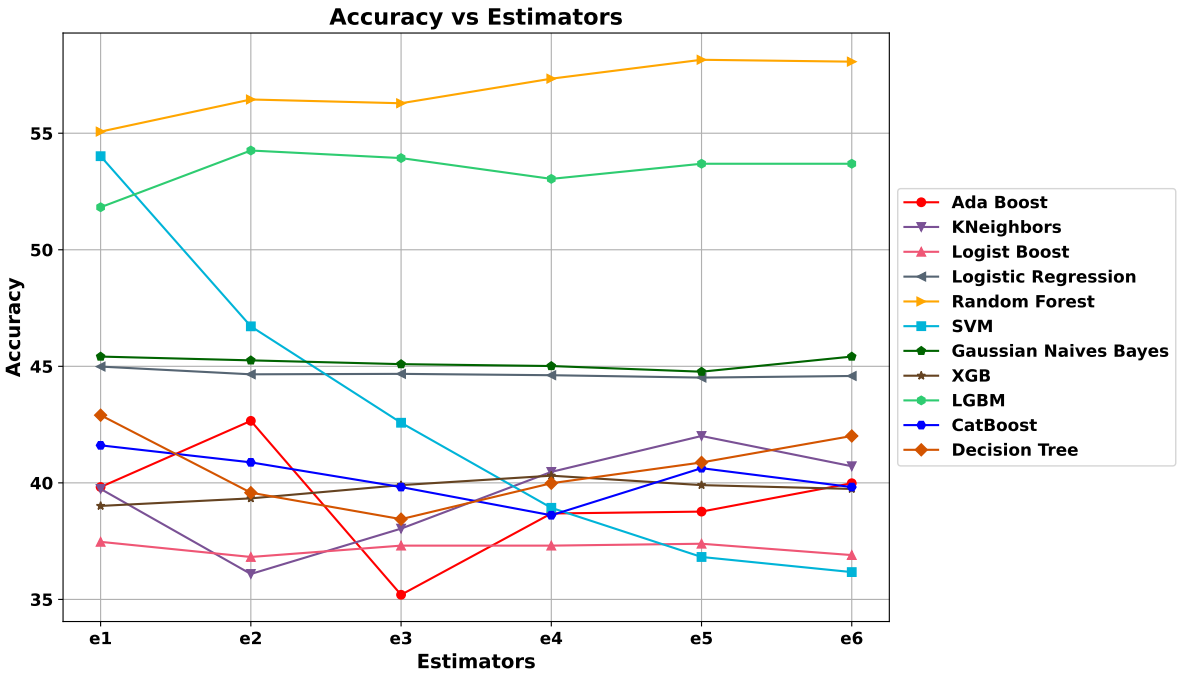


Figure 4.8: Model performance comparison: accuracy across different estimators for the valence dimension task.

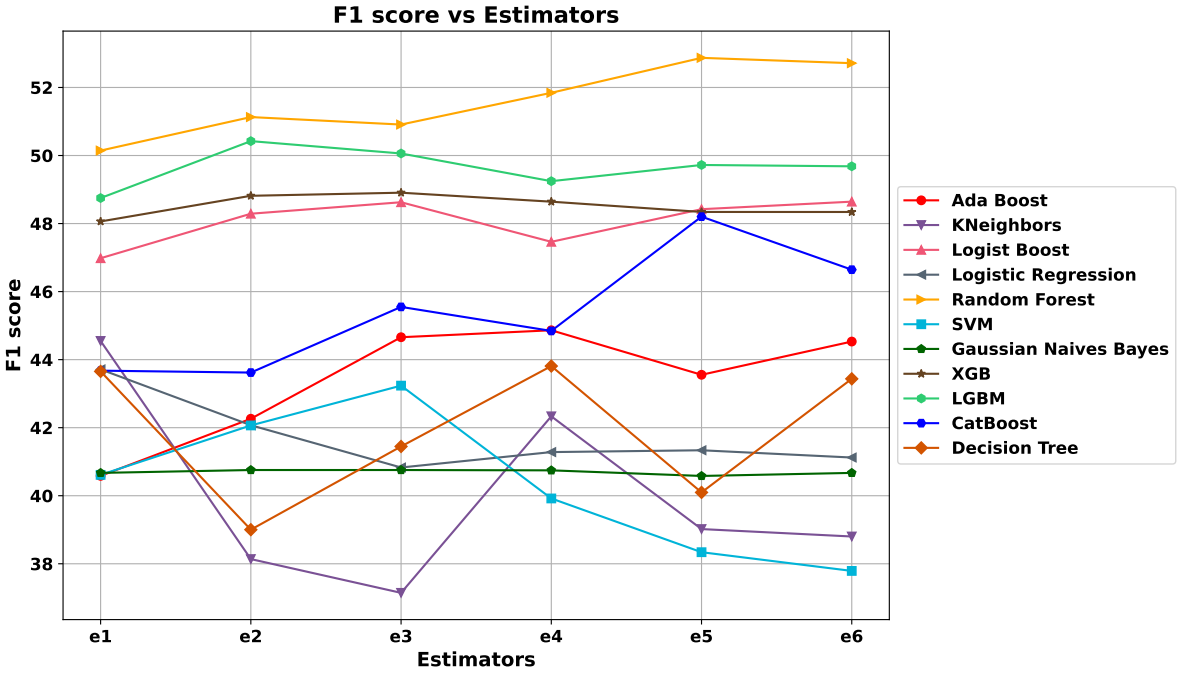


Figure 4.9: Model performance comparison: F1-score across different estimators for the valence dimension task.

52.41%, demonstrating that a larger ensemble of decision trees is better equipped to handle the vocal variability inherent in the OYH dataset.

Table 4.4: Detailed performance metrics for the Random Forest classifier on the valence dimension across various estimator counts.

Estimators	Accuracy	F1 Score	Precision	UAR	Class L Recall	Class M Recall	Class H Recall
100	51.09	46.96	45.53	34.9	2.1	74.67	27.82
200	54.66	50.16	49.54	37.6	2.8	79.02	31.08
300	55.96	51.00	50.30	38.1	1.4	81.48	31.33
400	57.42	52.04	51.93	39.0	1.4	83.94	31.58
500	57.42	51.71	51.99	38.8	1.4	84.80	30.08
600	58.47	52.41	53.85	39.4	1.4	86.83	29.82

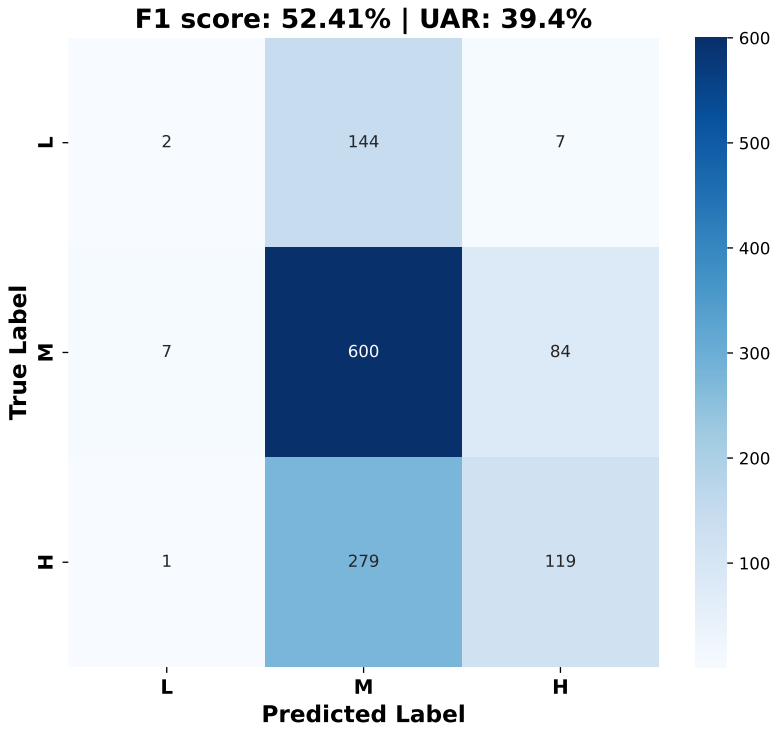


Figure 4.10: Confusion Matrix for the best-performing Random Forest classifier on the valence dimension task.

4.4.3 Classifier Performance for Control Prediction

A parallel comparative analysis was performed for the control dimension using its own distinct, optimal feature set. The results, presented in Figures 4.11 and 4.12, reveal another compelling finding: the best-performing classifier for control was different from that for valence.

Interestingly, the **Support Vector Machine (SVM)** emerged as the top-performing model for this task, achieving a peak accuracy of nearly **59.53%** and an F1-score of over 57%. The success of a margin-based classifier like SVM, in contrast to the tree-based ensembles that excelled on the valence task, further reinforces the distinction between the two dimensions. This suggests that the relationship between the more direct acoustic cues for control (RMS Energy, ZCR) and their corresponding classes may form a more cohesively separable decision boundary within a high-dimensional feature space, a problem for which SVMs are exceptionally powerful. Random Forests, conversely, were better suited to the more complex, fragmented decision space defined by the spectral and timbral features of valence.

Similar to the valence task, an accuracy of nearly 60% is a strong and significant result for this challenging three-class problem, establishing another robust baseline. The analysis of the SVM's confusion matrices at different complexity levels (Figure 4.13) also highlights a classic bias-variance trade-off. At lower complexity, the model is more regularized, leading to more balanced performance across all classes but with a lower overall accuracy. At higher complexity, the model achieves higher overall accuracy by fitting the data more closely, but this comes at the cost of reduced performance on the minority classes. This trade-off between peak accuracy and balanced, equitable performance across emotional classes is a central challenge in real-world affective computing.

The detailed performance indicators for the SVM classifier across different complexity levels (summarized in Table 4.5) reveal how the regularization parameter *C* dictates the model's predictive reliability.

Table 4.5: Detailed performance indicators for the SVM classifier on the control dimension at different complexity levels.

Complexity	Accuracy	F1 Score	Precision	UAR	Class L recall	Class M recall	Class H recall
1e-5	48.42	31.98	38.08	33.2	0	0.37	99.33
1e-4	59.53	57.04	54.86	42.9	0	60.52	68.28
1e-3	53.61	52.22	51.09	39.3	2.06	57.36	58.6
1e-2	48.01	46.81	45.91	35.4	3.09	45.62	57.43
1e-1	45.42	44.12	43.46	33.6	4.12	37.62	59.1
1.0	44.36	42.89	42.37	32.8	4.12	33.89	60.27

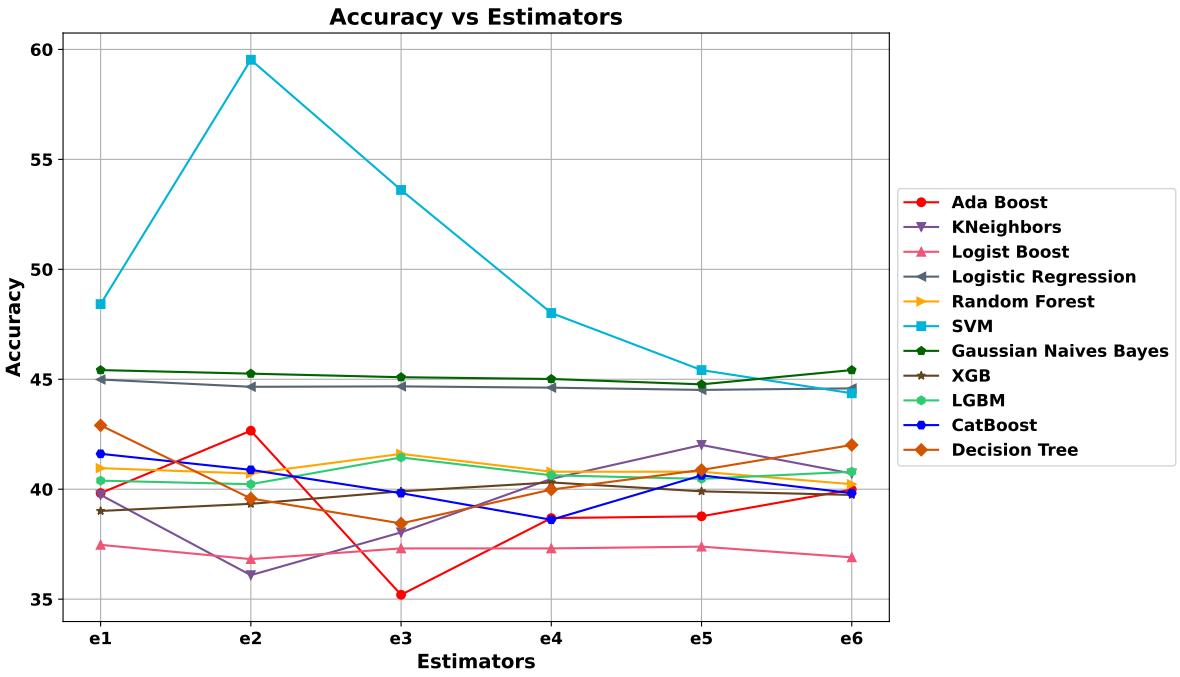


Figure 4.11: Model performance comparison: accuracy across different estimators/complexities for the control dimension task.

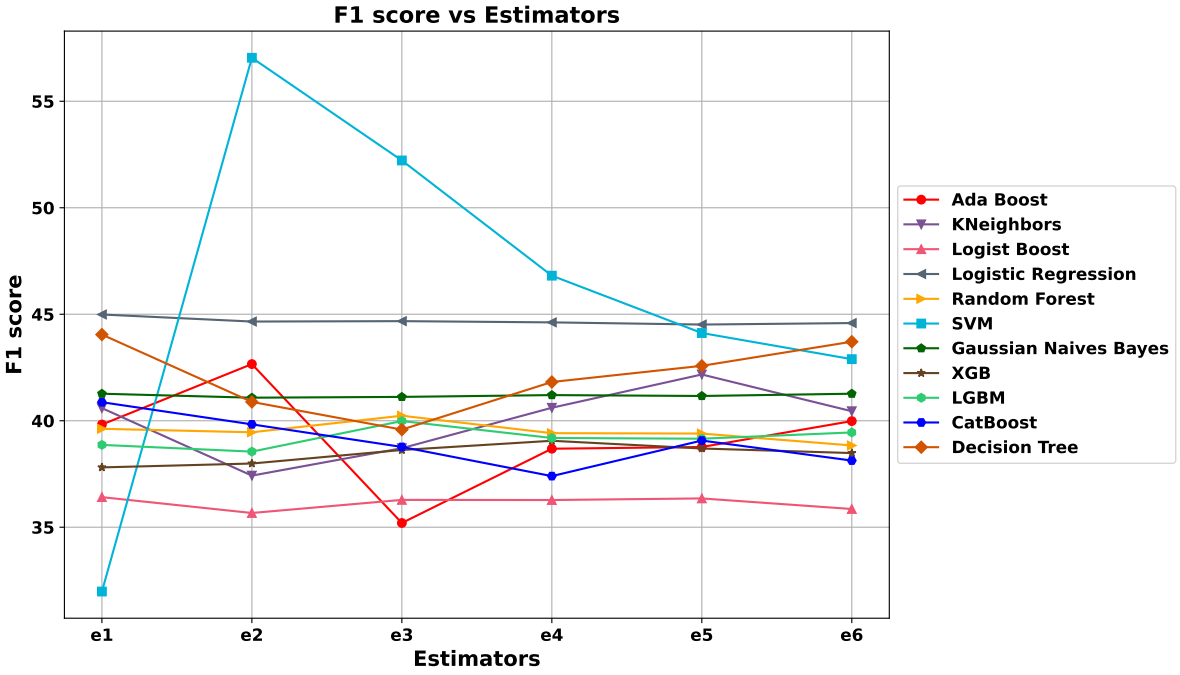
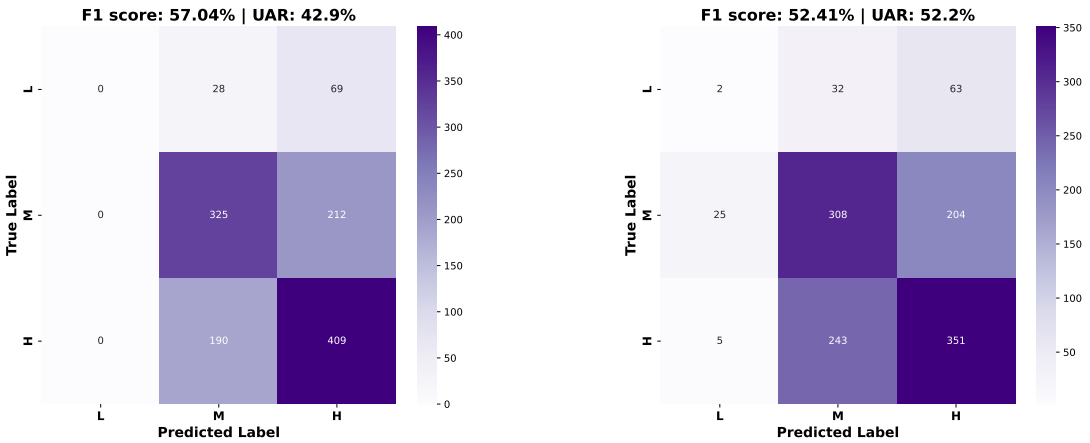


Figure 4.12: Model performance comparison: F1-score across different estimators/complexities for the control dimension task.



((a)) Complexity = 0.00001 ((b)) Complexity = 0.001000

Figure 4.13: Confusion Matrices for the SVM classifier on the control dimension task at two different complexity levels, showing the trade-off between overall accuracy and balanced class performance.

4.5 Conclusion

This chapter successfully detailed two major contributions to the field of Speech Emotion Recognition. The first and most significant is the creation, annotation, and public presentation of the **OYH corpus**, a novel, large-scale dataset for the under-resourced Algerian Arabic dialect. By capturing over six hours of spontaneous, genuine emotional speech, this corpus provides a vital resource for moving ASER research beyond the limitations of acted, English-language data and into more realistic and diverse domains.

The second contribution is the establishment of comprehensive and robust machine learning baselines on this new corpus. Through a meticulous methodology involving advanced feature selection and the comparative evaluation of eleven classifiers, this study established the state-of-the-art performance for this task, achieving a peak accuracy of **58.47%** with a Random Forest model for valence prediction and **59.53%** with an SVM for control prediction. The research also yielded critical insights, most notably that the acoustic features most salient for predicting valence are different from those for predicting control, underscoring the importance of task-specific feature optimization.

The performance levels achieved, while strong for such a challenging task, also highlight the immense difficulty of spontaneous, dialectal ASER when compared to the higher accuracies obtained on the clean, benchmark CREMA-D data in Chapter 3. This contrast sets the stage perfectly for the final discussion of this thesis. The following chapter will synthesize the findings

from both experimental chapters, comparing and contrasting the challenges and outcomes of working on benchmark versus in-the-wild data and discussing the broader implications for the field.

GENERAL CONCLUSION

This dissertation embarked on a systematic investigation into the complex and challenging domain of Automatic Speech Emotion Recognition (ASER). Motivated by the imperative for more accurate, efficient, and culturally inclusive affective computing systems, this research confronted several of the field's most persistent problems: the performance gap between acted and spontaneous speech, the scarcity of resources for under-resourced languages, the trade-off between model accuracy and efficiency, and the limitations of categorical emotion models. Through a series of methodical studies, this work has contributed novel deep learning architectures, validated new classification frameworks, and presented a vital new dataset to the research community. In doing so, it has successfully answered its primary research questions, affirming that it is possible to create and benchmark a new spontaneous corpus for a low-resource dialect; that a lightweight, attention-enhanced model can master the accuracy-efficiency trade-off; and that a new state-of-the-art in performance can be achieved by synergistically combining psychological theory with an architecturally diverse ensemble of models.

The research presented in this dissertation has made four principal and lasting contributions to the field. First and foremost is the creation of the **OYH dataset**, a large-scale corpus of spontaneous Algerian Arabic speech that not only addresses the critical need for more ecologically valid data but also promotes greater equity and inclusivity in affective computing research. Second, this thesis proposed and validated the **CBAM-DenseNet121**, a novel architecture that provides a pragmatic blueprint for developing ASER systems that are both highly accurate and computationally efficient enough for real-world deployment. Third, it introduced a **state-of-the-art ensemble framework**, validating a powerful paradigm wherein insights from psychology are fused with diverse computational models to overcome the performance plateaus of single-model systems. Finally, as a whole, this thesis represents a comprehensive cross-corpus empirical analysis, bridging the gap between theoretical benchmarking on clean data (CREMA-D) and the challenges of real-world applicability on a noisy, spontaneous dialectal corpus (OYH).

Limitations and Future Directions

In the interest of academic rigor, the deliberate scoping of this research defines clear and exciting avenues for future inquiry. The experimental work remained unimodal, focusing exclusively on the rich information within the audio signal. A natural and significant extension would be to incorporate the visual modality, exploring the fusion of the acoustic features developed herein with visual cues from facial expressions, such as action units and gaze patterns, to build a more holistic and robust multimodal emotion recognition system. Furthermore, while the findings are grounded in two primary datasets, the generalizability of these models to other types of spontaneous speech or other dialects presents an open and important research question.

These limitations directly inspire several promising directions for future work. The most immediate step is to apply the advanced deep learning models from Chapter 3 to the OYH corpus to establish a deep learning benchmark on spontaneous dialectal speech. This would also enable a formal investigation of the domain gap via cross-corpus generalization experiments, exploring domain adaptation techniques to improve model robustness. Finally, a state-of-the-art approach would be to leverage large, self-supervised foundation models for speech, such as WavLM [60]. Pre-training such a model on a large, unlabeled corpus of Algerian Arabic before fine-tuning on the OYH dataset could unlock unprecedented performance and potentially enable effective few-shot learning on other under-resourced Arabic dialects.

Closing Remarks

In closing, the research presented in this thesis has navigated the complex landscape of ASER, from the controlled environment of benchmark datasets to the challenging, authentic nature of real-world dialectal speech. By systematically addressing key problems of model efficiency, accuracy, and data scarcity, this work has delivered a tripartite contribution: a practical attention-based model, a novel state-of-the-art ensemble framework, and a vital new corpus for the under-resourced Algerian Arabic dialect. The findings herein reinforce the immense power of deep learning to capture the intricate patterns of human emotion, while simultaneously highlighting the critical importance of psychologically-grounded frameworks and the foundational need for diverse, representative data. It is hoped that the models and, in particular, the OYH corpus provided by this research will serve as a valuable resource for the scientific community, paving the way for future innovations and contributing to the ultimate goal of creating computational systems that are not just emotionally intelligent, but more humane, responsible, and culturally aware.

BIBLIOGRAPHY

[1] D. M. Schuller and B. W. Schuller, "A review on five recent and near-future developments in computational processing of emotion in the human voice," *Emotion Review*, vol. 13, no. 1, pp. 44–50, 2021.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[3] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[4] H. Waheed, S.-U. Hassan, R. Nawaz, N. Aljohani, G. Chen, and D. Gasevic, "Early prediction of learners at risk in self-paced education: A neural network approach," *Expert Systems with Applications*, vol. 213, 2022.

[5] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digital Signal Processing*, vol. 110, p. 102951, 2021.

[6] J. M. Garcia-Garcia, V. M. Penichet, M. D. Lozano, and A. Fernando, "Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions," *Universal Access in the Information Society*, vol. 21, no. 4, pp. 809–825, 2022.

[7] G. K. Verma, "Emotion recognition from facial expression in a noisy environment," in *Multimodal Affective Computing*, pp. 75–96, Bentham Science Publishers, 2023.

[8] R. Stock-Homburg, "Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research," *International Journal of Social Robotics*, vol. 14, no. 2, pp. 389–411, 2022.

[9] L. F. Barrett, "Solving the emotion paradox: Categorization and the experience of emotion," *Personality and Social Psychology Review*, vol. 10, no. 1, pp. 20–46, 2006.

[10] H. Dahmani, H. Hussein, B. Meyer-Sickendiek, and O. Jokisch, "Natural arabic language resources for emotion recognition in algerian dialect," in *Arabic Language Processing: From Theory To Practice: 7th International Conference, ICALP 2019*, pp. 18–33, 2019.

[11] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[12] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[13] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[14] Y.-W. Chen, J. Hirschberg, and Y. Tsao, "Noise robust speech emotion recognition with signal-to-noise ratio adapting speech enhancement," *arXiv preprint arXiv:2309.01164*, 2023.

[15] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations," *Psychological Review*, vol. 99, no. 3, pp. 561–565, 1992.

[16] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[17] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[18] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[19] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. The MIT Press, 1974.

[20] V. Sacharin, K. Schlegel, and K. R. Scherer, "Geneva Emotion Wheel rating study," tech. rep., University of Geneva, Swiss Center for Affective Sciences, 2012.

[21] V. Sacharin, K. Schlegel, and K. R. Scherer, "Geneva emotion wheel rating study," *NCCR Affective Sciences, University of Geneva*, 2012.

[22] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[24] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[25] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[26] A. A. A.-S. Khalil, "Real-time anger detection in arabic speech dialogs," Master's thesis, King Fahd University of Petroleum and Minerals (Saudi Arabia), 2011.

[27] K. Abainia, "DZDC12: a new multipurpose parallel algerian Arabizi–French code-switched corpus," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 419–455, 2020.

[28] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*, pp. 865–868, 2008.

[29] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[30] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, pp. 1–16, 2019.

[31] M. Belhadj, I. Bendellali, and E. Lakhdari, "KEDAS: A validated arabic speech emotion dataset," in *2022 International Symposium on iNnovative Informatics of Biskra (ISNIB)*, pp. 1–6, 2022.

[32] M. Meddeb, M. BenAmmar, and A. Alimi, "Towards a recommendation system for tv programs based on human behavior," in *The International Conference on Control, Engineering Information Technology, CEIT*, pp. 180–182, 2013.

[33] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion recognition using hybrid Gaussian mixture model and deep neural network," *IEEE access*, vol. 7, pp. 26777–26787, 2019.

[34] H. Horkous and M. Guerti, "Recognition of emotions in the Algerian Dialect Speech," *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 245–254, 2021.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[36] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. Interspeech 2012*, pp. 254–257, 2012.

[37] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[38] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2000.

[39] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Prentice Hall, 2011.

[40] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR, 2001.

[41] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[42] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5089–5093, 2018.

[43] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[44] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd ed., 2009.

[45] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[46] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[47] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," vol. 55, pp. 119–139, 1997.

[48] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[49] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[50] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, 2017.

[51] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31*, 2018.

[52] M. Bertalmío, ed., *Denoising of Photographic Images and Video Fundamentals, Open Challenges and New Trends*. UK: Springer, 2018.

[53] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[54] A. Chaudhary, K. S. Chouhan, J. Gajrani, and B. Sharma, *Deep Learning With PyTorch in Machine Learning and Deep Learning in Real-Time Applications*. 2020.

[55] H. Kinsley and D. Kukieła, *Neural Networks from Scratch in Python*. 2020.

[56] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[57] A. Romero, *Assisting the training of deep neural networks with applications to computer vision*. PhD thesis, University of Barcelona, 2015.

[58] G. Saint-Cirgue, *Apprendre le Machine Learning en une semaine*. 2019.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[60] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," vol. 16, pp. 1505–1518, 2022.

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[62] I. Tareq, B. Elbagoury, S. El-Regaily, and E.-S. El-Horbaty, "Analysis of ToN-IoT, UNW-NB15, and Edge-IIoT datasets using DL in cybersecurity for IoT," *Applied Sciences*, vol. 12, 2022.

[63] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[64] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *arXiv preprint arXiv:2104.02395*, 2022.

[65] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.

[66] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias-variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[67] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[68] U. Ishtiaq, E. R. M. F. Abdullah, and Z. Ishtiaque, "A hybrid technique for diabetic retinopathy detection based on ensemble-optimized cnn and texture features," *Diagnostics*, vol. 13, no. 10, p. 1816, 2023.

[69] M. Alshahrani, M. Al-Jabbar, E. M. Senan, I. A. Ahmed, and Jamil, "Hybrid methods for fundus image analysis for diagnosis of diabetic retinopathy development stages based on fusion features," *Diagnostics*, vol. 13, no. 17, pp. 2783–2783, 2023.

[70] Y.-C. Lin, H.-C. Chou, and H.-y. Lee, "Mitigating subgroup disparities in multi-label speech emotion recognition: A pseudo-labeling and unsupervised learning approach," 2025.

[71] F.-A. Croitoru, N.-C. Ristea, R. T. Ionescu, and N. Sebe, "Learning rate curriculum," *International Journal of Computer Vision*, pp. 1–24, 2024.

[72] N.-C. Ristea, R. T. Ionescu, and F. S. Khan, "SepTr: Separable transformer for audio spectrogram processing," *arXiv preprint arXiv:2203.09581*, 2022.

[73] N.-C. Ristea and R. T. Ionescu, "Self-paced ensemble learning for speech and audio classification," *arXiv preprint arXiv:2103.11988*, 2021.

[74] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[75] Z. Kahhoul, N. Terki, I. Benaissa, and Z.-E. Baarir, "Spectoresnet: Advancing speech emotion recognition through deep learning and data augmentation on the crema-d dataset," in *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2024.