République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed El-Bachir El-Ibrahimi - Bordj Bou Arreridj

Faculté des Sciences et de la technologie
Département d'Electronique
Laboratoire d'Electronique et Télécommunications Avancées (ETA)

THÈSE

EN VUE DE L'OBTENTION DU DIPLOME DE DOCTORAT

Domaine : Sciences et Technologies Filière : Télécommunications
Spécialité : Télécommunications et Intelligence Artificielle

Présentée par
Hamza ROUBHI

Thème

Sélection de paramètres en grande dimension pour la reconnaissance d'émotions

Soutenue le: 21/01/2026 **Devant le Jury composé de :**

Nom et Prénom	Grade		
Mr Idris MESSAOUDENE	**MCA**	Univ. de BBA	Président
Mr Abdenour HACINE-GHARBI	**MCA**	Univ. de BBA	Rapporteur
Mr Abdelmalik OUAMANE	**Professeur**	Univ. de Biskra	Examinateur
Mr Ammar CHOUCHANE	**MCA**	Univ. de Barika	Examinateur
Mr Salah Eddine MEZAACHE	**MCA**	Univ. de BBA	Examinateur
Mr Khaled ROUABAH	**Professeur**	Univ. de M'sila	Invité

Année Universitaire: 2025 /2026

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University Mohamed El Bachir El Ibrahimi of Bordj Bou Arreridj

Faculty of Technology
Electrical Engineering Department
Laboratory of Advanced Electronics and Telecommunications (ETA)

THESIS

WITH A VIEW TOWARDS OBTAINING THE DIPLOMA OF DOCTORATE

Domain: Science and Technology Branch: Telecommunications
Speciality: Telecommunications and Artificial Intelligence

Presented by
Hamza ROUBHI

Title

Feature Selection in High Dimension for Emotion Recognition

Defended on: 21/01/2026 **In front of the Jury composed of:**

Last and first name	Grade		
Mr Idris MESSAOUDENE	**MCA**	Univ. of BBA	President
Mr Abdenour HACINE-GHARBI	**MCA**	Univ. of BBA	Supervisor
Mr Abdelmalik OUAMANE	**Professor**	Univ. of Biskra	Examiner
Mr Ammar CHOUCHANE	**MCA**	Univ. of Barika	Examiner
Mr Salah Eddine MEZAACHE	**MCA**	Univ. of BBA	Examiner
Mr Khaled ROUABAH	**Professor**	Univ. of M'sila	Invited

University year: 2025 /2026

ACKNOWLEDGEMENT

First and foremost, I express my deepest gratitude to **Allah** for granting me the faith, patience, perseverance, and courage to embark upon and complete this challenging yet profoundly rewarding journey. All praise and thanks are due to Him alone.

The successful completion of this doctoral work has been shaped and enriched by the guidance, support, and encouragement of many exceptional individuals, to whom I owe my deepest appreciation.

My heartfelt thanks go first and foremost to my thesis advisor, **Dr. Abdenour Hacine-Gharbi**, for his unwavering guidance, profound intellectual insight, and steadfast belief in my potential. His mentorship illuminated complex concepts and inspired a deep passion for the subject, transforming what could have been a demanding task into an intellectually stimulating and fulfilling experience. His continuous encouragement has been instrumental to both my academic progress and personal growth.

I also wish to express my heartfelt gratitude to **Prof. Khaled Rouabah**, **Dr. Thameur Dhieb**, and **Prof. Philippe Ravier** for their valuable collaboration and guidance during the preparation and publication of my research articles. Their constructive feedback and scientific expertise have greatly enhanced the quality and impact of my work.

I would also like to extend my sincere appreciation to the distinguished members of my defense committee: **Dr. Idris Messaoudene**, for graciously agreeing to chair the jury, and **Prof. Abdelmalik Ouamane**, **Dr. Ammar Chouchane**, and **Dr. Salah Eddine Mezaache**, for devoting their valuable time to review my thesis and for their insightful comments and constructive feedback, which significantly improved its quality.

My profound thanks also go to all the professors of the **Higher School of Telecommunications (ENTTIC)** and the **ETA Laboratory**, whose inspiring lectures and deep knowledge nurtured my academic curiosity.

Above all, I am deeply grateful to my **beloved parents**, whose unconditional love, endless sacrifices, and steadfast encouragement have been my greatest source of strength. Their faith in me has been the foundation of all my achievements. I am equally grateful to my **family and friends** for their patience, understanding, and unwavering support throughout this long journey.

To everyone who offered a kind word, a smile, or a prayer along the way, your gestures, whether big or small, have been deeply appreciated and sincerely valued.

DEDICATION

إهداء

اللهم لك الحمد كما ينبغي لجلال وجهك وعظيم سلطانك،

لك الحمد على ما أنعمت، ولك الشكر على ما وفقت ويسّرت.

اللهم كما أعنتني على إتمام هذا العمل، فاجعل فيه الخير والبركة، ووفقني لما تحب وترضى.

إلى أمي الغالية، التي منحتني دفء حبها، واحتوتني بكرمها، فكانت سندي في الحياة وأملي وقت الشدة.

إلى أبي العزيز، الذي لم يبخل عليّ بجهده وتعبه وسهره وتضحياته من أجلي.

إلى أختي الغالية سلمى، وإخوتي الأعزاء عبد القادر، عبد الرحيم، ويونس، الذين كانوا دومًا دعمي وسندي في كل خطوة من رحلتي.

إلى أقاربي وأصدقائي، لكل من ساندني وكان له أثر في رحلتي، أقدّم خالص الشكر والتقدير.

أهدي هذا العمل المتواضع إلى كل من وقف إلى جانبي في رحلتي، مع أصدق مشاعر الامتنان والعرفان للجميع.

LIST OF SCIENTIFIC CONTRIBUTIONS

International Publications

[1] **H. Roubhi,** A. Hacine-Gharbi, T. Dhieb, K. Rouabah, and P. Ravier, "A Novel Approach to Enhancing Performance in 1D-CNN-Based Speech Emotion Recognition Using Mutual Information-Based Feature Selection," Journal of Engineering Science and Technology Review, vol. 18, pp. 104–112, Aug. 2025, doi: 10.25103/jestr.184.15.

[2] **H. Roubhi,** A. Hacine-Gharbi, K. Rouabah, and P. Ravier, "Mutual Information-based Feature Selection Strategy for Speech Emotion Recognition using Machine Learning Algorithms Combined with the Voting Rules Method," Engineering, Technology and Applied Science Research, vol. 15, pp. 19207–19213, Nov. 2024, doi: 10.48084/etasr.9066.

[3] **H. Roubhi,** A. Hacine-Gharbi, K. Rouabah, "Real-Time Facial Expression Recognition Using 1D-CNN and Mutual Information-Based Feature Selection " The revision has been submitted to Signal, Image and Video Processing Journal.

International Communications

[4] **H. Roubhi,** A. Hacine-Gharbi, R. Touahria, and K. Rouabah "Mutual Information-Based Feature Selection for Improved AlexNet-Based Facial Expression Recognition " The Paper has been submitted to The 7th International Conference on Computing Systems and Applications, Algiers, ALGERIA

National Communications

[5] **H. Roubhi,** A. Hacine-Gharbi, and R. Touahria, Facial Expression Recognition Using VGG16-Based Feature Extraction with KNN Classifier. 2025.

[6] **H. Roubhi,** A. Hacine-Gharbi, and F. Ghazali, Alex Net-Based Feature Extraction Combined with a KNN Classifier for Facial Expression Recognition. 2024.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS AND ABBREVIATIONS

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

1D-CNN – One-Dimensional Convolutional Neural Network

2D-CNN – Two-Dimensional Convolutional Neural Network

AAM – Active Appearance Model

AER – Automatic Emotion Recognition

AI – Artificial Intelligence

ANN – Artificial Neural Network

BC – Box Constraint

BFCC – Bark Frequency Cepstral Coefficients

BSIF – Binarized Statistical Image Features

CIFE – Conditional Infomax Feature Extraction

CK+ – Extended Cohn–Kanade Dataset

CLAHE – Contrast Limited Adaptive Histogram Equalization

CLCM – Custom Lightweight CNN-based Model

CNN – Convolutional Neural Network

CO – Contempt (facial expression class)

D-CNN – Deep Convolutional Neural Network

DCT – Discrete Cosine Transform

DFT – Discrete Fourier Transform

DI – Disgust (facial expression class)

DISFA – Denver Intensity of Spontaneous Facial Actions Database

DWT – Discrete Wavelet Transform

EEG – Electroencephalogram

EMO-DB – Berlin Database of Emotional Speech

FC6 / FC7 / FC8 – Fully Connected Layers 6, 7, and 8 in AlexNet

FE – Fear (facial expression class)

FER – Facial Expression Recognition

FFT – Fast Fourier Transform

GMM – Gaussian Mixture Model

GSR – Galvanic Skin Response

HA – Happiness (facial expression class)

HMM – Hidden Markov Model

HOG – Histogram of Oriented Gradients

HTK – Hidden Markov Model Toolkit

ICAP – Interaction Capping

JAFFE – Japanese Female Facial Expression Database

JMI – Joint Mutual Information

KNN – k-Nearest Neighbor

LBP – Local Binary Pattern

LDA – Linear Discriminant Analysis

LPCC – Linear Predictive Cepstral Coefficients

LPQ – Local Phase Quantization

MAP – Maximum A Posteriori

MFCC – Mel-Frequency Cepstral Coefficients

MI – Mutual Information

ML – Machine Learning

MLP – Multilayer Perceptron

MMI – MMI Facial Expression Database

mRMR – Minimum Redundancy Maximum Relevance

MUG – Multimedia Understanding Group Facial Expression Database

NLP – Natural Language Processing

PLP – Perceptual Linear Prediction

RBF – Radial Basis Function

RR – Recognition Rate

RRF – Reduction Rate of Features

RRT – Reduction Rate of Processing Time

RNN – Recurrent Neural Network

SC1 / SC2 – Stopping Criteria 1 and 2

SER – Speech Emotion Recognition

SVM – Support Vector Machine

ViT – Vision Transformer

VGG16 – Visual Geometry Group 16-Layer Network

VR – Virtual Reality

Abstract

In recent years, the rapid growth of human–machine interaction systems has created an increasing need for machines that can perceive and interpret human emotions. Emotion recognition systems, based on physiological signals, facial expressions, and speech, have become an essential component in designing intelligent and responsive systems. One of the major challenges in emotion recognition lies in the high dimensionality of extracted features, often caused by the large variety of descriptors used in emotion modeling. This high dimensionality can lead to significant computational cost, overfitting, and reduced classification accuracy due to the curse of dimensionality. Therefore, effective feature selection becomes crucial to retain only the most relevant features while improving recognition accuracy and reducing complexity.

The main objective of this doctoral work is to select the most relevant features from high-dimensional data to develop an efficient emotion recognition system using mutual information measures. Our first contribution proposes a baseline speech emotion recognition (SER) system using a large set of acoustic features (MFCC, LPCC, and PLP), combined with MI-based selection strategies (mRMR, ICAP, CIFE, and JMI). These features are classified using k-NN, SVM, and GMM algorithms, combined with a voting rule strategy. Experimental results on the EmoDB dataset show that applying the ICAP strategy reduces the feature vector size by 62.2% while improving the recognition accuracy to 82.94%. Moreover, the introduction of a 1D-CNN with a stopping criterion guided by classification accuracy achieved superior performance, where applying CIFE resulted in a 73.87% reduction of the feature vector with only 0.39% loss in accuracy.

The second contribution extends this framework to facial expression recognition (FER) using MI-based feature selection strategies. Experiments conducted on HOG descriptors extracted from the CK+ and MUG datasets achieved 100% accuracy with over 60% feature reduction when using ICAP combined with a 1D-CNN. Furthermore, applying mRMR to deep features extracted through transfer learning with AlexNet and classifying them with k-NN yielded 97% accuracy using only 9 features out of 1000.

The findings of this thesis demonstrate the effectiveness of MI-based feature selection strategies in addressing the challenges posed by high dimensionality, offering an optimal balance between recognition accuracy and computational complexity in emotion recognition systems.

Keywords: Speech emotion recognition, facial expression recognition, high-dimension, feature selection, mutual information, 1DCNN, machine learning.

Résumé

Ces dernières années, la croissance rapide des systèmes d'interaction homme–machine a engendré un besoin croissant de machines capables de percevoir et d'interpréter les émotions humaines. Les systèmes de reconnaissance des émotions, basés sur les signaux physiologiques, les expressions faciales et la parole, sont devenus un élément essentiel pour la conception de systèmes intelligents et réactifs. L'un des principaux défis de la reconnaissance des émotions réside dans la haute dimensionnalité des caractéristiques extraites, souvent due à la diversité de descripteurs utilisés dans la modélisation des émotions. Cette dimensionnalité élevée peut entraîner un coût computationnel important, un surapprentissage, ainsi qu'une baisse de la précision de classification en raison de la malédiction de la dimension. Ainsi, la sélection efficace des paramètres devient cruciale afin de ne retenir que les plus pertinents, tout en améliorant la précision et en réduisant la complexité.

L'objectif principal de ce travail doctoral est de sélectionner les paramètres pertinents à partir de données de grande dimension afin de développer un système de reconnaissance des émotions performant, basé sur des mesures d'information mutuelle. Notre première contribution consiste à proposer un système de reconnaissance acoustique des émotions, en utilisant un large ensemble de paramètres acoustiques (MFCC, LPCC et PLP) combinés à des stratégies de sélection basées sur l'information mutuelle (mRMR, ICAP, CIFE et JMI). Ces paramètres sont classés à l'aide des algorithmes k-NN, SVM et GMM, associés à la stratégie de la règle de vote. Les résultats expérimentaux sur la base de données EmoDB montrent que l'application de la stratégie de sélection ICAP réduit la taille du vecteur des paramètres de 62,2% tout en améliorant la précision à 82,94%. De plus, l'introduction d'un 1D-CNN avec un critère d'arrêt guidé par la précision de classification a permis d'obtenir des performances supérieures, où l'application de CIFE a entraîné une réduction de 73,87 % du vecteur de paramètres avec seulement 0,39 % de perte de précision.

La deuxième contribution étend ce cadre à la reconnaissance des expressions faciales en utilisant des stratégies de sélection de paramètres basée sur l'information mutuelle. Des expériences menées premièrement sur les paramètres HOG extraits des bases de données CK+ et MUG montrent une précision de 100 % avec plus de 60 % de réduction des paramètres en utilisant la stratégie ICAP

combinée avec le 1D-CNN. Deuxièmement, l'application de la stratégie MRMR aux paramètres extraits via l'apprentissage par transfert avec AlexNet et leur classification par l'algorithme KNN a permis d'obtenir de bonnes performances avec 97 % de précision en utilisant seulement 9 paramètres sur 1000.

Les résultats de cette thèse montrent l'efficacité des stratégies de sélection de paramètres basées sur l'information mutuelle pour relever les défis liés à la haute dimensionnalité, offrant un compromis optimal entre la précision de reconnaissance et la complexité computationnelle dans les systèmes de reconnaissance des émotions.

Mots-clés : Reconnaissance des émotions dans la parole, reconnaissance des expressions faciales, haute dimension, sélection de paramètres, information mutuelle, 1D-CNN, apprentissage automatique.

الملخص

في السنوات الأخيرة، أدى النمو السريع لأنظمة التفاعل بين الإنسان والآلة إلى تزايد الحاجة إلى آلات قادرة على إدراك المشاعر الإنسانية وتفسيرها. لقد أصبحت أنظمة التعرف على المشاعر، المعتمدة على الإشارات الفسيولوجية وتعبيرات الوجه والكلام، مكوّناً أساسياً في تصميم الأنظمة الذكية والتفاعلية. ومن أبرز التحديات في هذا المجال الارتفاع الكبير في بُعدية الميزات المستخرجة، الناتج غالباً عن التنوع الكبير في الواصفات المستخدمة في نمذجة المشاعر. هذا الارتفاع في البُعدية قد يؤدي إلى زيادة التكلفة الحسابية، وفرط التعلّم(Overfitting) ، وانخفاض دقة التصنيف نتيجة "مشكلة البُعدية". لذلك تصبح عملية اختيار الميزات الأكثر صلة أمراً ضرورياً للاحتفاظ فقط بما هو فعّال، مع تحسين الدقة وتقليل التعقيد.

الهدف الرئيسي من هذه الأطروحة هو اختيار الميزات الأكثر صلة من بيانات عالية البُعد لتطوير نظام فعّال للتعرف على المشاعر، وذلك بالاعتماد على قياسات المعلومة المتبادلة .(Mutual Information) المساهمة الأولى تتمثل في اقتراح نظام أساسي للتعرف على المشاعر في الكلام (SER) باستخدام مجموعة واسعة من الميزات الصوتية(MFCC, LPCC et PLP) ، المدمجة مع استراتيجيات اختيار الميزات القائمة على المعلومات المتبادلة (mRMR, ICAP, CIFE et JMI) .تم تصنيف هذه الميزات باستخدام خوارزمياتk-NN ، وSVM ، وGMM ، مع دمجها عبر استراتيجية التصويت. أظهرت النتائج التجريبية على قاعدة بيانات EmoDB أن تطبيق ICAP أدى إلى تقليص حجم الميزات بنسبة 62.2% مع تحسين دقة التعرف إلى 82.94%. علاوة على ذلك، فإن إدخال شبكة عصبية التفافية أحادية البُعد (1D-CNN) مع معيار إيقاف موجّه بالدقة أتاح أداءً متقدماً، حيث أدى تطبيق CIFE إلى تقليص الميزات بنسبة 73.87% مع فقدان طفيف لا يتجاوز 0.39% في الدقة.

المساهمة الثانية توسّع هذا العمل نحو التعرف على تعابير الوجه (FER) باستخدام استراتيجيات اختيار الميزات بالاعتماد على المعلومة المتبادلة. أظهرت التجارب على واصفات HOG المستخرجة من قواعد البيانات +CK و MUG تحقيق دقة بلغت 100% مع تقليص يتجاوز 60% عند استخدام ICAP مع شبكة 1-D-CNN .كما أن تطبيق mRMR على الميزات العميقة المستخرجة عبر التعلم بالنقل باستخدام AlexNet ، وتصنيفها بخوارزميةk-NN ، حقق دقة بلغت 97% باستخدام 9 ميزات فقط من أصل 1000.

تُبرز نتائج هذه الأطروحة فعالية استراتيجيات اختيار الميزات المعتمدة على المعلوم المتبادلة في مواجهة تحديات ارتفاع البُعدية، وذلك من خلال توفير توازن أمثل بين دقة التعرف والتعقيد الحسابي في أنظمة التعرف على المشاعر.

الكلمات المفتاحية :التعرف على المشاعر في الكلام، التعرف على تعابير الوجه، الأبعاد العالية، اختيار الميزات، المعلومة المتبادلة، الشبكات العصبية الالتفافية D-CNN1، التعلم الآلي.

General Introduction

In recent years, artificial intelligence (AI) has been increasingly integrated into almost every aspect of human life, ranging from healthcare and education to security, entertainment, and human computer interaction. One of the emerging fields within AI is automatic emotion recognition (AER), which aims to give machines the ability to perceive, analyze, and interpret human emotions [1]. Such systems have attracted significant attention due to their potential applications in diverse domains, including intelligent tutoring systems, affective computing, mental health monitoring, social robotics and customer service [2].

Despite this progress, emotion recognition systems still face several challenges. Among the most critical is the issue of high-dimensional feature spaces generated during the feature extraction stage. Modern AER systems typically rely on a large number of features, whether acoustic, visual, or physiological modalities, in order to represent the complex patterns that define emotional states. However, not all extracted features are equally relevant or informative for this task. The presence of redundant, irrelevant, or noisy features not only increases computational cost and memory consumption but also worsens the so-called "curse of dimensionality", which often leads to degraded classification accuracy [3].

Feature selection therefore plays a central role in the design of efficient emotion recognition systems. By identifying an optimal subset of the most relevant features, feature selection not only reduces system complexity but also improves interpretability and classification accuracy. Among existing approaches, mutual information (MI)-based feature selection methods have gained particular attention due to their strong theoretical foundations in information theory and their ability to quantify nonlinear dependencies between features and class labels [4]. Nevertheless, the effectiveness of MI-based methods strongly depends on the search strategy and, critically, on the choice of an appropriate stopping criterion, which remains an open research issue in high-dimensional emotion recognition tasks.

The main objective of this PhD thesis is to address these challenges by investigating and enhancing MI-based feature selection strategies for high-dimensional automatic emotion recognition systems. The proposed work aims to balance recognition accuracy with computational efficiency, making emotion recognition systems more suitable for real-world and resource-constrained environments. Specifically, this thesis seeks to achieve three core objectives:

1. Enhance processing efficiency by reducing the feature set and thereby lowering computational overhead.
2. Reduce memory usage through compact feature representations.
3. Mitigate the curse of dimensionality to improve the accuracy of emotion recognition models.

This research focuses on speech and facial expression modalities, which are among the most widely studied and practically deployable sources of emotional information. These modalities are non-intrusive, easily accessible, and well suited for real-world applications compared to physiological signals.

Building upon this foundation, this thesis also explores the design of robust classification systems that combine machine learning algorithms with voting rules as well as lightweight 1D convolutional neural networks (1D-CNNs). These approaches are chosen to balance efficiency with high recognition accuracy, making them particularly suitable for practical applications.

Within this context, the main scientific contributions of this thesis are presented as follows:

The first contribution consists in the design of a comprehensive speech emotion recognition (SER) system based on the extraction of a large and diverse set of acoustic features, including MFCC, LPCC, and PLP. To mitigate the effects of high dimensionality in the extracted feature space, several MI–based feature selection methods, namely mRMR, ICAP, CIFE, and JMI are systematically applied. Emotion classification is performed using multiple machine learning algorithms such as KNN, SVM, and GMM, combined with voting rule strategy. In addition, a lightweight 1D-CNN architecture is proposed, incorporating a stopping criterion guided by classification accuracy to automatically determine an optimal feature subset and further enhance system performance.

The second contribution extends the proposed framework to facial expression recognition (FER). The FER system uses both handcrafted texture descriptors, such as LBP, HOG, LPQ, and BSIF, and deep feature representations extracted from a pretrained AlexNet model. MI–based feature selection strategies (mRMR, ICAP, CIFE, and JMI) are applied using different stopping criteria, while classification is carried out using KNN and 1D-CNN models. This contribution demonstrates the effectiveness and generality of MI-based feature selection across both traditional handcrafted features and deep learned representations.

This PhD thesis is structured into four main chapters, in addition to the general introduction and conclusion:

- **Chapter I** presents a general description of automatic emotion recognition. It introduces the concept of emotion and its relevance to AI, followed by an overview of different modalities (speech, facial expressions, physiological signals) and applications. The chapter also outlines the architecture of AER systems, reviews the literature, and discusses feature extraction and classification approaches.
- **Chapter II** forms the core of the thesis, focusing on feature selection. It provides a theoretical background on feature selection approaches, the information-theoretic foundations of mutual information, and the challenges posed by high-dimensional feature spaces. Different search strategies for feature selection and the role of stopping criteria are also discussed.
- **Chapter III** addresses the application of feature selection to Speech Emotion Recognition (SER). The chapter presents the proposed contributions, including the combination of machine learning classifiers with voting rules, the design of a 1D CNN architecture for SER, and the implementation of mutual information-based feature selection with greedy forward search and stopping criteria. Extensive experimental results and evaluations are reported.
- **Chapter IV** extends the application of feature selection to Facial Expression Recognition (FER). This chapter covers both handcrafted features and deep features extracted through transfer learning. The proposed system leverages MI-based feature selection and advanced classification strategies, with experiments and results validating the effectiveness of the approach.

Finally, the general conclusion summarizes the findings of the thesis, highlights the contributions, and outlines potential future research directions.

I

GENERALITIES OF AUTOMATIC EMOTION RECOGNITION

I.1. INTRODUCTION

Emotion recognition is a crucial field within artificial intelligence that enables machines to interpret human affective states and thereby enhance human–computer interaction. Its importance lies in making interactions more natural and adaptive by allowing systems to recognize and respond to human emotions. Emotions can be conveyed through different modalities, including speech, facial expressions, gestures, and physiological signals, each providing complementary information [5]. The ability to recognize emotions has led to a wide range of applications in healthcare, education, security, marketing, and entertainment. To achieve these applications, a typical emotion recognition system is composed of several stages, including feature extraction, feature selection, and classification [6].

In this chapter, we introduce the fundamental concepts of emotion recognition, focusing on its definition, applications, and modalities. We then present the state of the art of existing systems and methods in emotion recognition, with a particular focus on speech and facial expressions. This is followed by an overview of emotion recognition systems, and a detailed discussion of the feature extraction and classification techniques used in this study. Together, these sections provide a solid foundation for the subsequent chapters, where more advanced approaches and research contributions will be examined.

I.2. OVERVIEW OF EMOTION

The concept and definition of emotion serve as the cornerstone for the field of emotion recognition. Paul Ekman, in the 1970s, introduced the foundational concept of emotion, which has profoundly influenced subsequent research and understanding in this domain [7].

I.2.1. Defining Emotion

This subsection presents different perspectives on the definition of emotion. The fundamental concepts are summarized as follows:

Definition 1: According to the Merriam-Webster Dictionary [8], emotion is described as a conscious mental response, such as anger or fear, that is subjectively experienced as an intense feeling, generally oriented toward a particular object and accompanied by characteristic physiological and behavioral alterations.

Definition 2: As stated in [9], Emotion is a state that reflects evaluative judgments of the environment, self, and others in relation to an organism's goals and beliefs, motivating and coordinating adaptive behavior.

Definition 3: Emotion [10] is a mental state linked to thoughts, feelings, and behavior.

Definition 4: Emotion [11] is an affective state triggered by interpersonal or external events, involving an appraisal of the situation and expressed through physiological signals or actions.

Definition 5: Emotions [12] are states of the central nervous system associated with a range of behavioral, cognitive, somatic, and physiological responses that have evolved to fulfill survival needs.

I.3. EMOTION RECOGNITION

Emotion recognition is the ability to identify and understand emotions expressed by oneself and others, playing a crucial role in social interactions and various fields such as psychology, neuroscience, and artificial intelligence. This capacity is not only essential for effective communication but also has significant implications for mental health treatment and human-computer interaction. As technology advances, the methodologies for emotion recognition have expanded to include various modalities, most notably facial expressions, speech, and physiological signals, each providing unique insights into emotional states and enhancing the accuracy of emotion detection systems [13].

I.4. MODALITIES FOR EMOTION RECOGNITION

Emotion recognition systems use various modalities to identify and interpret human emotions effectively. These modalities encompass different types of data inputs, including

visual, auditory, and textual information, each contributing unique insights into emotional states [14].

Visual Modality

Visual modality primarily relies on facial expressions and gestures to determine emotions. This approach often involves the use of facial recognition technology, which analyzes images or video frames to extract features indicative of specific emotional states. Recent advancements have incorporated deep learning techniques, such as CNNs, to enhance accuracy in recognizing emotions from facial sequences [15].

Auditory Modality

The auditory modality analyzes voice features such as tone, pitch, and intonation to detect emotions. Speech processing techniques, including Fast Fourier Transform (FFT) and Mel-Frequency Cepstral Coefficients (MFCCs), are commonly used to extract these features from audio data. This modality is widely used because vocal expressions naturally convey emotional information [16].

Textual Modality

The textual modality examines written or transcribed language to identify emotions. Techniques from Natural Language Processing (NLP), such as sentiment analysis and linguistic feature extraction, are applied to capture emotional content in text. Textual analysis is useful for understanding expressed or implied emotions in communication [17].

Physiological Modality

Physiological signals, such as ECG, GSR, and EEG, provide direct measurements of the body's emotional responses. These signals capture subtle changes in heart rate, skin conductance, and brain activity that reflect different emotional states. When combined with deep learning models, including Transformers, physiological data can enhance the accuracy, stability, and robustness of emotion recognition systems [18].

Multimodal Approaches

Recent trends in emotion recognition have favored multimodal approaches, which combine data from various modalities to enhance overall performance. By integrating visual, auditory, and textual information, these systems can leverage the strengths of each modality while compensating for individual weaknesses [19].

In this thesis, the focus is placed on speech and facial modalities, as they represent the most natural and expressive channels for conveying human emotions in everyday interactions. Speech provides dynamic information through prosody and vocal tone, while facial expressions offer rich visual cues that are often considered universal indicators of affect. Together, these two modalities complement each other, enhancing robustness and reliability in automatic emotion recognition, especially in scenarios where one modality may be noisy or incomplete.

I.5. APPLICATIONS OF EMOTION RECOGNITION

In recent years, emotion recognition has gained significant attention due to its wide range of applications in many domains. By enabling machines to interpret human emotions, this technique enhances human-computer interaction, improves mental health monitoring, and facilitates progress in areas such as security, education and entertainment. The ability to automatically identify emotions from facial expressions, speech and physiological signals has led to the development of innovative solutions with the aim of improving user experience, decision making and overall well-being. Some of the major applications of emotion recognition include [20]:

Healthcare and Mental Health Monitoring

Emotion recognition plays an important role in healthcare, especially in mental health monitoring and diagnosis. By analyzing facial expressions, speech patterns and physiological signals, the emotional recognition system can help detect stress [21], anxiety [22] and depression [23] in an early stage. Furthermore, integrating emotion recognition into digital health platforms, mobile applications, and telemedicine services enables real-time mental health monitoring outside clinical settings. This capability allows for proactive intervention, reducing the burden on healthcare professionals and enhancing patient outcomes [24].

Education and E-learning

Emotion recognition has become an essential tool in modern education and e-learning, revolutionizing the way students interact with digital learning environments [25]. By analyzing facial expressions, the emotion-aware system can assess students' engagement, inspiration and mental effort in real time. This capacity allows teachers and adaptive teaching platforms to personalize the instructions, ensuring that students receive materials suited to their emotional and cognitive states [26].

Security and Surveillance

Detection and prevention of security threats is a long -term concern for government agencies [27]. Research has detected the ability to analyze physiological and facial reactions to identify signs of deception during interviews [28]. Such technology has been considered for implementation in high-protection environments including airports to increase the assessment of danger. Additionally, the recognition of emotions can be integrated into surveillance systems to monitor the crowd [29], detect unusual emotional patterns, and identify potential risks. By improving situational awareness and enabling active safety measures, this technique enhances public safety and supports emergency response efforts.

Virtual Reality

Virtual Reality (VR) facilitates the presentation of realistic and dynamic stimuli in social contexts, making it a valuable tool for assessing and training emotion recognition [30]. Studies have shown that VR can enhance both the evaluation and development of emotion recognition skills, particularly in individuals with severe mental illness [31].

Marketing and Customer Experience

In marketing, understanding customer emotions has become essential for predicting consumer engagement and purchase decisions [32]. In customer service, real-time monitoring of emotional states allows companies to detect frustration or satisfaction, personalize interactions, and strengthen loyalty [33]. Contact centers also benefit from emotion-aware systems that help managers identify dissatisfaction early and provide targeted employee training [34].

I.6. EMOTION RECOGNITION SYSTEM OVERVIEW

An AER system, whether based on speech or facial expressions, generally follows a structured process consisting of several key stages: preprocessing, feature extraction, feature selection, and classification [20]. This process aims to transform raw input data into a predicted emotional class. It begins with data acquisition, where emotional speech signals or facial images are collected from annotated databases. The preprocessing stage prepares the data by removing noise and standardizing input formats. In speech, this may involve silence removal, normalization, and segmentation; for facial data, it often includes face detection, grayscale conversion, and alignment. Following preprocessing, the system extracts meaningful features that capture the emotional content. In SER, these features may include prosodic features such

as pitch and energy, spectral features such as MFCC, PLP, and LPCC, as well as voice quality features. In FER, the system may use handcrafted features such as HOG, LBP, LPG, and BSIF, or deep features obtained from CNNs such as VGG16, VGG19, or AlexNet. Given the high dimensional nature of the extracted features especially in image based or deep learning approaches, feature selection plays a crucial role. It reduces dimensionality by identifying the most relevant and non-redundant features, thereby enhancing classification performance and minimizing computational cost. In this context, mutual information-based techniques are particularly effective, as they evaluate the dependency between features and emotion labels, allowing the system to retain only the most optimal features. The selected features are then used by a classifier to determine the corresponding emotional class. Classification techniques range from traditional machine learning models such as SVM, GMM, and k-NN, to deep learning architectures like CNNs and recurrent neural networks (RNNs). It is also important to note that the system is typically divided into two phases: a training phase, where the model learns from labeled data, and a testing phase, where its performance is evaluated on unseen data. This overall process applies to both speech and face based emotion recognition systems [35]. Within the context of this thesis, the feature selection stage is of particular importance, as it directly addresses the challenges associated with high-dimensional features. All these stages are illustrated in *Figure I-1* below, which provides a global overview of the emotion recognition process adopted in this work.

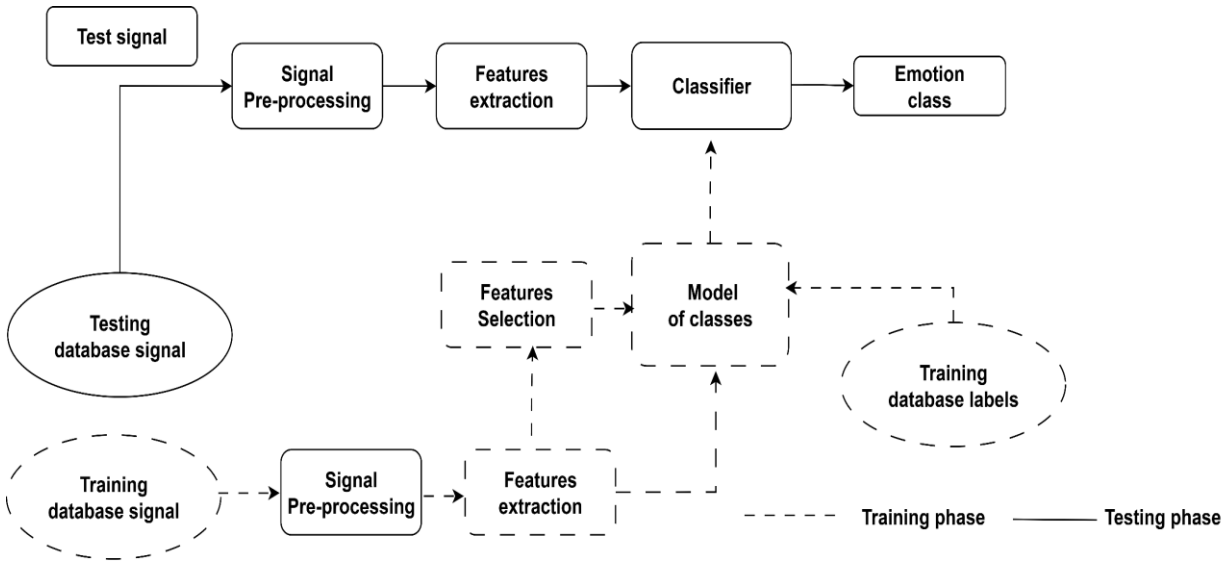


Figure I-1Overview of the Emotion Recognition System Architecture [35]

I.7. CHALLENGES IN AUTOMATIC EMOTION RECOGNITION

Despite significant progress, AER systems continue to face multiple challenges that limit their performance and generalization in real world conditions. These challenges arise mainly from the high dimensionality of input data, system complexity, data variability, and limitations in feature extraction and selection techniques.

One of the major issues is high-dimensional feature data, especially in modalities such as speech, facial expressions, and EEG signals. Extracting informative features from these high-dimensional sources often leads to redundancy, irrelevant information, and an increased risk of overfitting, particularly when the amount of data is limited. Studies such as [36] and [37] have emphasized the need for efficient feature selection strategies to reduce dimensionality while preserving emotion-relevant features.

Another persistent challenge is system complexity. Deep learning-based AER models, though powerful in automatic feature extraction, require extensive computational resources and are prone to overfitting due to their large number of parameters and dependence on large, well-balanced datasets. Works like [38] underline the difficulty of balancing model accuracy and computational efficiency, especially in resource-constrained or real-time applications.

Furthermore, variations in data and recording conditions such as differences in speakers, recording conditions, lighting, and background noise pose significant barriers to achieving robust emotion recognition. In multimodal systems, feature fusion introduces additional complexity, as shown in [39] and [19], where combining high-dimensional data from multiple modalities (speech, text, face, EEG) increases computational load and complicates optimization.

Finally, interpretability and explainability remain crucial challenges. As highlighted in [40], many deep and hybrid AER models function as “black boxes,” making it difficult to understand the contribution of individual features or network layers to the final classification decision.

I.8. LITERATURE REVIEW ON EMOTION RECOGNITION

The study of SER and FER has evolved significantly over the years, reflecting advancements in technology and methodology. Recent literature highlights various approaches and challenges associated with these fields.

I.8.1. Speech Emotion Recognition

Speech Emotion Recognition focuses on extracting meaningful features from audio signals to classify emotional states. Over the years, methods have evolved from classical feature-based approaches, such as prosodic and MFCC features combined with SVM or HMM classifiers, to deep learning techniques using CNNs, LSTMs, and Transformers. This subsection highlights representative works in this field.

Bhangale et al. [41] proposed a 1D Deep CNN that uses 39 MFCC coefficients along with 715 additional acoustic features to enhance feature distinctiveness. Experiments on the EmoDB database reported accuracies of 91.28% and 93.31%, respectively demonstrating the effectiveness of the approach. Nevertheless, the direct application of this model entails increased computational time and memory consumption due to the high dimensionality of the extracted features.

Zhao et al. [42] introduced a hybrid CNN framework that combines a 1D-CNN for raw speech processing with a 2D-CNN designed to capture spectral patterns from log-mel spectrograms. This dual-path design aimed to extract complementary high-level features from both temporal and spectral domains. To enhance training efficiency, transfer learning was employed. On the EmoDB database, the model achieved accuracies of 91.78% for Speaker Independent (SI) and 92.71% for Speaker Dependent (SD). Despite these promising results, the approach suffers from computational inefficiency due to the complex multi-layer CNN architecture.

Issa et al. [43] explored a diverse set of features of speech signals by extracting MFCCs, Mel spectrograms, chromagrams, spectral contrast, and Tonnetz features. These handcrafted features were passed through 1D convolutional and pooling layers to capture deeper features, followed by a fully connected layer for classification. On the EmoDB dataset, the model obtained an accuracy of 86.10%. Nevertheless, despite the diverse feature set, the architecture

remains relatively complex and its recognition performance is modest, which limits its applicability in real-world scenarios.

In [44], the authors applied a pre-trained D-CNN for feature extraction, followed by correlation-based feature selection to keep the most relevant features. Several classifiers were tested, including SVM, KNN, Random Forest, and MLP. Using the EmoDB database, MLP achieved 90.05% in SI experiments, while SVM reached 95.10% in SD settings. Despite these strong results, the reliance on correlation-based selection may exclude nonlinear features that are important for accurate emotion recognition.

The representative studies discussed above are summarized in *Table I-1*, which provides an overview of the datasets, features, classifiers, and reported performance for each approach in SER.

Table I-1:Overview of Studies on SER: Datasets, Features, Feature Selection, Classifiers, and Performance

Dataset	Features	Classifier / Model	Dimension	Feature Selection	Performance	Reference
EmoDB	39 MFCCs, 715 acoustic features	1D D-CNN	Low/ High	None	91.28%, 93.31%	[41]
EmoDB	Raw speech, Log-mel spectrograms	Hybrid 1D-CNN + 2D-CNN (Transfer Learning)	High	None	91.78% (SI), 92.71% (SD)	[42]
EmoDB	MFCC, Mel spectrogram, chromagram, spectral contrast, Tonnetz	1D-CNN + FC layer	High	None	86.10%	[43]
EmoDB	Pre-trained D-CNN features	SVM, KNN, RF, MLP	Medium	Correlation-based	90.05% (SI, MLP), 95.10% (SD, SVM)	[44]

I.8.2. Facial Expression Recognition

Facial Expression Recognition aims to classify emotions from visual cues, often relying on deep learning architectures due to the high-dimensionality of image data. Methods have progressed from handcrafted features like LBP or HOG with classical classifiers to deep CNNs and Transformer-based models. This subsection highlights representative works in this field [45].

Qin et al. [46] proposed an FER method combining Gabor wavelet transform with a two-channel CNN. After preprocessing and key frame extraction, Gabor features were used to train the CNN, achieving 96.81% accuracy on CK+.

The authors in [47] introduced a Custom Lightweight CNN-based model (CLCM) derived from MobileNetV2 for FER. Evaluated on FER-2013, RAF-DB, AffectNet, and CK+ datasets, it achieved competitive performance (63% on FER-2013, 84% on RAF-DB) while using only 2.3M parameters, fewer than MobileNetV2 (3.5M) and ShuffleNetV2 (3.9M).

Bendjillali et al. [48] developed a system that applies the Viola–Jones algorithm for face detection, CLAHE for image enhancement, and DWT for feature extraction, followed by CNN classification. This approach reached 96.46% on CK+ and 98.43% on JAFFE.

In [49], the authors proposed a three-stage FER framework: (1) feature extraction using three DWT-based descriptors, (2) wrapper-based feature selection, and (3) SVM classification. The method achieved 87.76% on CK+ and 89.66% on JAFFE.

The work in [50] presented a video-based FER framework using feature point movement and block texture variability. Twenty-four feature points from the AAM were analyzed, and expressive features were extracted and classified with a 1D-CNN. Reported accuracies were 95.2% on BHU, 96.5% on MMI, and 97% on the merged dataset.

In [51], a genetic programming-based FER framework was proposed for feature selection and fusion. It combined geometric and textural features through a tree-based genetic algorithm, tested on DISFA+, CK+, and MUG datasets, achieving 94.2%, 98.0%, and 97.2%, respectively.

Boukhobza et al. [52] integrated a wrapper-based feature selection approach with localized facial region analysis (mouth, eyes, eyebrows). On the MUG dataset, their system

achieved 100% accuracy with LDA while reducing the feature set by 50%. Eyebrow features were found to be highly discriminative.

The studies reviewed above for FER are summarized in *Table I-2*, highlighting the datasets, feature extraction methods, classification models, and their reported performance.

Table I-2:Overview of Studies on FER: Datasets, Features, Feature Selection, Classifiers, and Performance

Dataset	Features	Classifier / Model	Dimension	Feature Selection	Performance	Reference
CK+	Gabor wavelet features + 2-channel CNN	CNN	High	None	96.81%	[46]
FER-2013, RAF-DB, AffectNet, CK+	Raw images, MobileNetV2-inspired lightweight CNN (CLCM)	CLCM	High	None	63% (FER-2013), 84% (RAF-DB)	[47]
CK+, JAFFE	DWT features + CNN (with Viola–Jones + CLAHE preprocessing	CNN	High	None	96.46% (CK+), 98.43% (JAFFE)	[48]
CK+, JAFFE	3 DWT-based descriptors	SVM	Medium	Wrapper	87.76% (CK+), 89.66% (JAFFE)	[49]
BHU, MMI, BHU+MMI	AAM feature points + block texture	1D-CNN	High	None	95.2% (BHU), 96.5% (MMI), 97% (Merged)	[50]
DISFA+, CK+, MUG	Geometric + textural features, genetic	Binary classifiers (pairwise)	Medium	Genetic programming	94.2% (DISFA+), 98.0% (CK+), 97.2% (MUG)	[51]
MUG	Local facial regions (mouth, eyes, eyebrows)	LDA	Medium	Wrapper	100% (50% feature reduction)	[52]

By reviewing representative studies in both speech and facial emotion recognition, several trends and challenges become evident. SER research increasingly integrates temporal modeling and deep embeddings, while FER research relies on high-dimensional visual features and deep learning architectures. These observations motivate the need for robust feature selection strategies, particularly in high-dimensional scenarios, which is the focus of this thesis.

I.9. FEATURE EXTRACTION

Feature extraction is a crucial step in any AER system. Its purpose is to transform raw signals, whether speech or facial images into informative representations suitable for classification. For speech signals, this typically involves preprocessing, segmenting the signal into short-time analysis windows, and extracting features that capture spectral and temporal characteristics. For facial images, preprocessing may include normalization and alignment, followed by extraction of texture or shape descriptors. By converting raw data into discriminative feature vectors, this step enables the system to efficiently recognize emotions from complex inputs [53].

I.9.1. Speech Feature Extraction

I.9.1.1. Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are a popular feature extraction technique used primarily in speech and audio signal processing. They provide a representation of the short-term power spectrum of an audio signal, based on a nonlinear Mel scale of frequency, which closely resembles the human auditory system's response to sound [54].

The main stages involved in the MFCC extraction process are summarized in *Figure I-2*

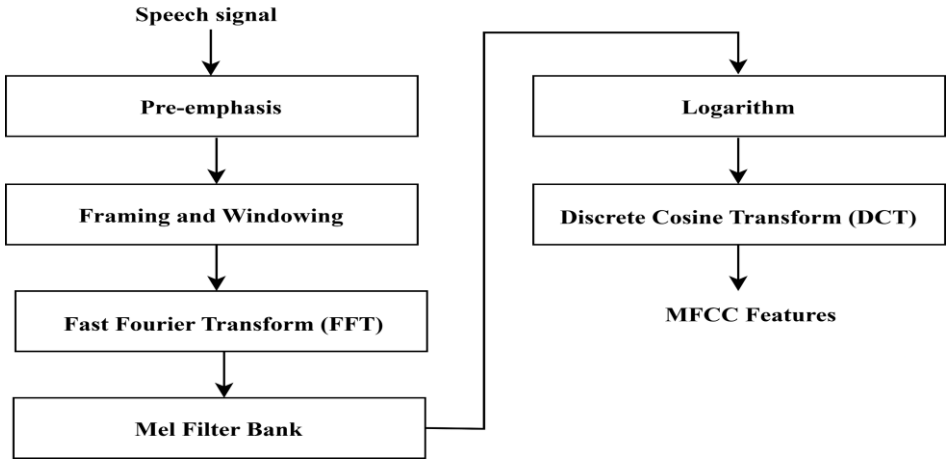


Figure I-2:MFCC Feature Extraction Diagram [55]

Pre-emphasis: This step boosts high-frequency components in the audio signal to enhance the overall spectral characteristics.

Windowing: The pre-emphasized signal is segmented into overlapping frames, typically 20-40 ms in duration. Each frame is multiplied by a windowing function, such as the Hamming or Hanning window, to minimize discontinuities at the edges of the frames [56].

Fast Fourier Transform (FFT): The windowed signal is converted to the frequency domain using the FFT algorithm, which facilitates the analysis of its frequency components.

Mel-Filter Bank: The FFT output is filtered using a bank of triangular filters spaced according to the Mel scale, which aligns more closely with human perception of sound frequencies

Log Power Spectrum: The output from the Mel-frequency filtering is transformed into a log power spectrum.

Discrete Cosine Transform (DCT): Finally, the log power spectrum is subjected to the DCT to obtain the MFCCs, which serve as the feature set for further analysis. The Mel scale itself is a perceptual scale of frequencies that reflects the human auditory system's sensitivity to different frequency ranges. This characteristic makes MFCCs particularly effective in distinguishing various emotions conveyed through speech [57].

I.9.1.2. Perceptual Linear Prediction (PLP)

The PLP was introduced by Hermansky [58] and has since been widely adopted due to its ability to enhance robustness against noise, reverberation, and echo, thereby improving overall system performance. Its feature extraction process is based on three fundamental psychoacoustic principles that approximate the human auditory system to estimate the auditory spectrum and derive the PLP coefficients: (1) critical-band spectral resolution, (2) the equal-loudness curve, and (3) the intensity-to-loudness power law, also known as cubic-root compression [58]. The complete block diagram illustrating the extraction process of PLP features from the speech signal is presented in *Figure I-3.*

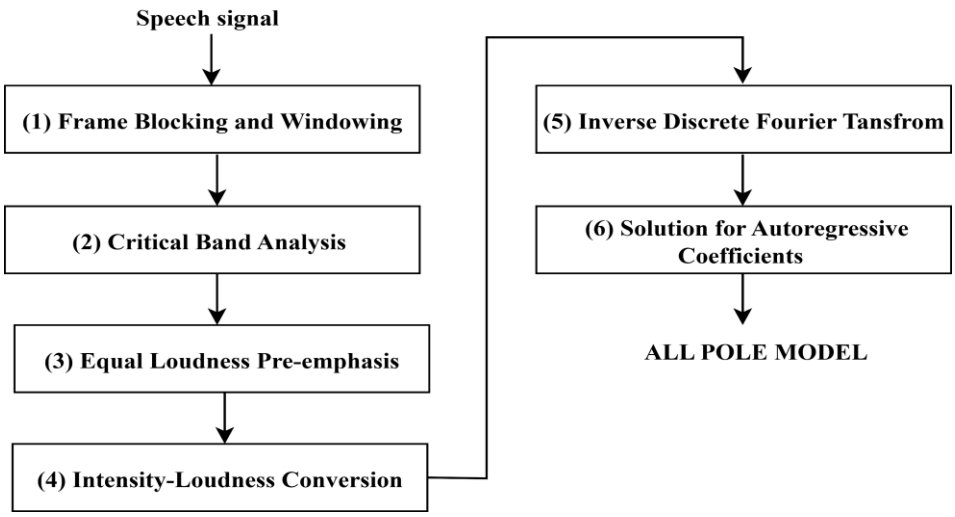


Figure I-3: PLP Feature Extraction Diagram [58]

I.9.1.3. Linear Predictive Cepstral Coefficients (LPCC)

LPCCs are a type of cepstral feature derived from Linear Predictive Coding (LPC) techniques. They are widely used in speech and audio signal processing. The core principle of LPCC lies in the assumption that a current speech sample can be approximated as a linear combination of its preceding samples [59]. The algorithmic flow of the LPCC extraction process is illustrated in *Figure I-4*. The first step consists of framing and windowing the input speech signal using a Hamming window. Next, linear predictive analysis is performed, based on the assumption that the configuration of the vocal tract primarily determines the characteristics of the produced speech signal. To model this behavior, the vocal tract is typically represented using a digital all-pole filter, whose transfer function in the z-domain is expressed as:

$$H(z) = \frac{\mathrm{G}}{1 - \sum_{\mathrm{k}=1}^{\mathrm{p}} a_k z^{-k}} \qquad (I\text{-}1)$$

Where $H(z)$ represents the transfer function of the vocal tract, G denotes the gain of the filter, a_k is the set of autoregressive coefficients known as the Linear Prediction Coefficients (LPCs), and p indicates the order of the all-pole filter. An efficient technique commonly used to estimate both the LPC coefficients and the filter gain is the autocorrelation method.

The final stage of the algorithm involves cepstral analysis, which refers to the process of extracting the cepstrum from a speech signal. There are two primary approaches to cepstral analysis: the Fast Fourier Transform (FFT)-based cepstrum and the LPC-based cepstrum [60].

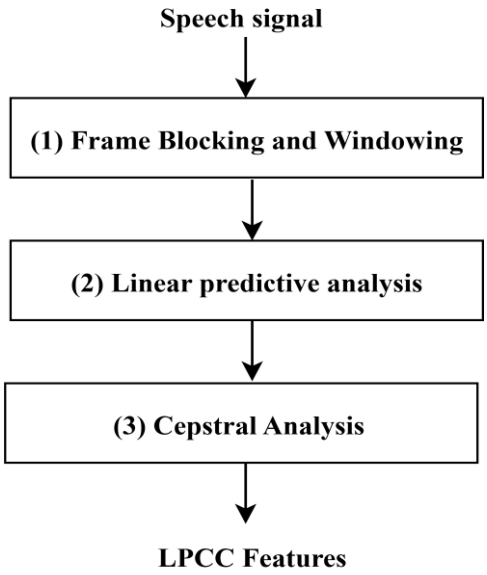


Figure I-4: LPCC Feature Extraction Diagram [60]

I.9.2. Facial Feature Extraction

I.9.2.1. Histogram of Oriented Gradients (HoG)

The HoG is a well-known descriptor proposed by Dalal and Triggs [61]. It is a widely used feature descriptor in various computer vision tasks, including human detection, FER, and pedestrian identification. The underlying principle of the HoG descriptor is that the local appearance and shape of objects can be effectively characterized by the distribution of intensity gradients and edge orientations within an image.

HoG feature extraction involves dividing the image into small spatial regions, referred to as cells. Within each cell, a one-dimensional histogram of gradient orientations is computed using the local pixel intensities. These histograms are subsequently concatenated to form a comprehensive representation of the image. To enhance robustness to illumination changes and shadows, local contrast normalization is applied. This is achieved by grouping adjacent cells into larger regions called blocks and normalizing the histograms within each block. This normalization process significantly improves the descriptor's invariance to variations in lighting conditions.

I.9.2.2. Local Binary Patterns (LBP)

The LBP is a widely recognized gray-scale texture descriptor commonly employed in image processing and computer vision tasks [62]. The LBP operator assigns a binary label to

each pixel by thresholding the intensity values of its neighboring pixels against the value of the central pixel. The resulting binary pattern is then converted into a decimal value, which serves as the LBP code for that pixel.

To construct the feature vector, the image is typically divided into blocks, and histograms of the LBP codes (with $2^8 = 256$ possible values for a 3×3 neighborhood) are computed for each block. These histograms are then concatenated to form the final feature representation.

The original LBP operator is based on a 3×3 neighborhood, but it can be extended to neighborhoods of different sizes. This is achieved by using a circular neighborhood and applying bilinear interpolation when the neighboring points fall between pixel locations. The extended version is expressed as $LBP(P,R)$, where P represents the number of neighboring points and R is the radius of the circular neighborhood. This generalization allows the LBP to capture texture information at multiple spatial resolutions. The LBP code is then computed for each pixel located at coordinates $(\mathrm{x_c}, \mathrm{y_c})$ as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \qquad (I\text{-}2)$$

Here, g_c is the gray value of the center pixel $(\mathrm{x_c}, \mathrm{y_c})$, g_p represents the gray values of Pequally spaced pixels on a circle with radius R, and $s(x)$ is a thresholding function defined as follows:

$$s(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ otherwise} \end{cases} \qquad (I\text{-}3)$$

I.9.2.3. Local Phase Quantization (LPQ)

To address the limitation of LBP's sensitivity to image blur, Ojansivu and Heikkilä proposed the LPQ descriptor [63]. LPQ enhances robustness to blur by quantizing the phase information obtained from the local Fourier transform, which is less affected by blurring than raw intensity values.

In this method, the 2-D local Fourier transform is computed within a $(2R+1) \times (2R+1)$ window centered around each pixel in an image of size $n \times n$. From this transform, four complex coefficients are retained, corresponding to spatial frequencies $\mathrm{v1} = [\mathrm{a}, 0]$, $\mathrm{v2} = [0, \mathrm{a}]$, $\mathrm{v3} = [\mathrm{a}, \mathrm{a}]$, $\mathrm{v4} = [-\mathrm{a}, \mathrm{a}]$, where $a = \frac{1}{2R+1}$.

The real and imaginary parts of these coefficients are then combined into an 8-dimensional vector for each pixel, resulting in a matrix of size $8 \times n^2$. This matrix is decorrelated using a whitening operation, assuming a correlation of 0.95 between neighboring pixels and a Gaussian distribution. The resulting values are then binarized: positive values are set to 1, and negative values to 0. Each 8-bit column is converted into a decimal value (0–255), creating a 256-dimensional histogram used for classification tasks.

I.9.2.4. Binarized Statistical Image Features (BSIF)

The BSIF descriptor, proposed by Kannala and Rahtu [64], learns filters directly from a small set of natural images, distinguishing it from descriptors like LBP and LPQ that rely on manually defined filters. For each pixel, BSIF generates a binary code by analyzing the local intensity pattern within its neighborhood [65].

Given an image patch X of size $l \times l$ pixels and a linear filter W_i of the same dimensions, the filter response S_i is computed as follows:

$$s_i = \sum_{u,v} W_i(u,v)X(u,v) = w_i^T x \qquad (I\text{-}4)$$

The vectors w_i and x contain the pixel values of the filter W_i and the image patch X, respectively. The binarized feature b_i is defined as $b_i = 1$ if $s_i > 0$, and $b_i = 0$ otherwise. The filters W_i are derived through Independent Component Analysis (ICA).

I.9.2.5. Feature extraction-based transfer learning

Transfer learning, also known as knowledge transfer, is a machine learning technique in which a model trained on one task can be refined or reused for another task [66]. Instead of training a model from scratch, which is often time-consuming and requires large datasets, transfer learning leverages pre-trained models, thereby addressing the challenges of training deep learning algorithms when only limited data are available [67] [68].

Pre-trained models can be used in two main approaches: as feature extractors or as end-to-end classifiers [67]. In the feature extraction approach, the pre-trained model serves as a feature extractor, where the learned parameters are transferred to the new task without modification [69]. Alternatively, in fine-tuning, certain parameters are adjusted to better adapt to the target task. In the end-to-end approach, the entire model is trained on the target dataset, starting from pre-trained weights.

In this study, we adopt the feature extraction strategy by using pre-trained models such as AlexNet and VGG16 as feature extractors for emotion recognition.

I.9.2.5.1. AlexNet

AlexNet is a convolutional neural network developed by Alex Krizhevsky et al [70], it was trained on the ImageNet dataset and has a total of eight layers containing 62.3 million learnable parameters. The model architecture consists of five convolutional layers with max pooling, followed by three fully connected layers and two dropout layers. The ReLU activation function is applied to all layers except the final output layer, which uses the Softmax activation function, as illustrated in *Figure I-5.*

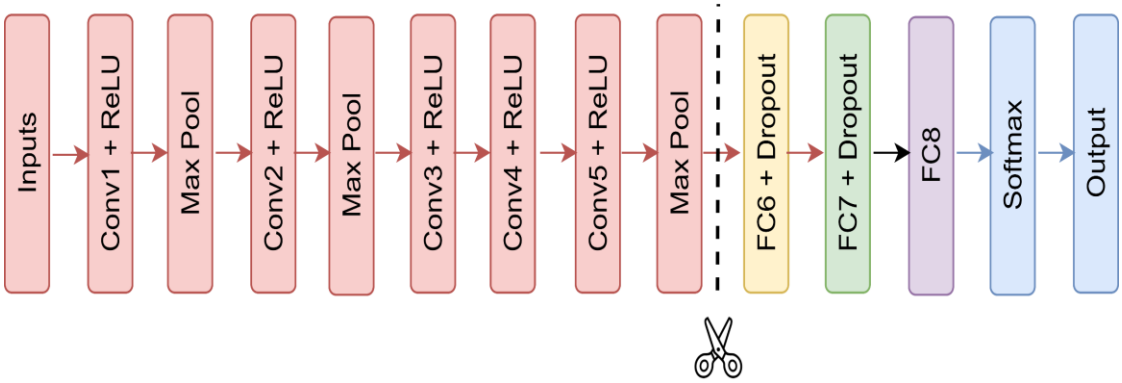


Figure I-5: AlexNet Feature Extraction [70]

I.9.2.5.2. VGG16

VGG16, similar to AlexNet, is a convolutional neural network introduced by Simonyan and Zisserman [71], it was also trained on the ImageNet dataset and has a total of 16 layers with 13 convolution layers and 3 fully connected. VGG16 is widely used for image classification because it applies multiple 3×3 filters with a stride of one in each convolutional layer. As shown in *Figure I-6*, the 13 convolutional layers extract hierarchical features from the input, while the fully connected layers perform classification. These layers are divided into five blocks, each followed by a max-pooling layer. The model takes an input image of size 224×224 and outputs the predicted object label [72].

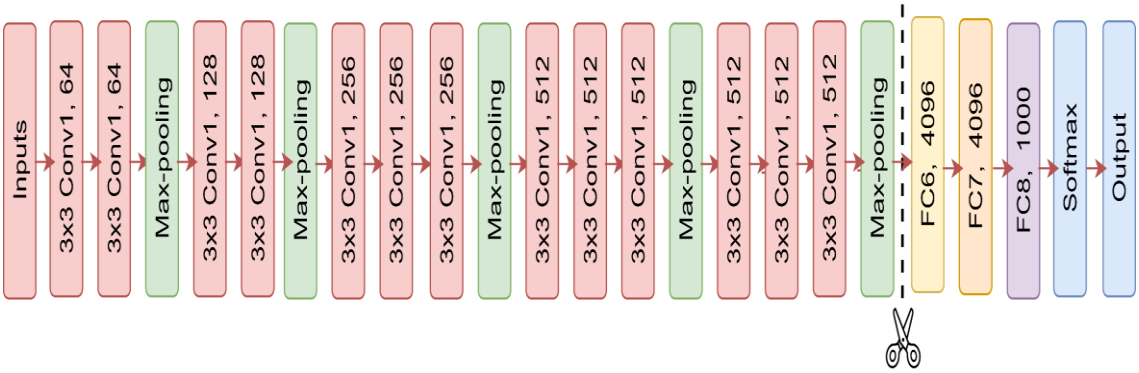


Figure I-6:VGG 16 Feature Extraction [71]

In this study, AlexNet and VGG16 were used in the feature extraction phase. By leveraging the transfer learning approach, it was not necessary to retrain these models from scratch, saving significant computational time. The fully connected layers, specifically Fc6, Fc7, and Fc8, were used to extract features, and these layers were combined to generate a comprehensive feature representation [73]. The details of the extraction process and the number of features will be presented in *Chapter IV.*

I.10. CLASSIFICATION APPROACHES

In an AER system, the classification stage is generally divided into two primary phases: model learning during training and decision making during testing. During the training phase, the classifier learns to represent each emotion from labeled feature vectors and constructs statistical models or discriminative boundaries that distinguish classes in the feature space. The structure of the learned model depends on the selected approach, including probabilistic modeling such as Gaussian Mixture Models, instance-based metric learning like k-Nearest Neighbors, margin-based optimization such as Support Vector Machines, or hierarchical feature learning using deep neural networks [74].

During the testing phase, the trained classifier applies the learned model to previously unseen feature vectors and assigns them to the most likely class using a decision rule based on likelihood, distance, similarity, or posterior probability. In this work, we use both machine learning and deep learning approaches to perform emotion classification, allowing the system to effectively learn from complex feature representations and achieve robust recognition across varying conditions.

I.10.1. Classical Machine Learning Classifiers

I.10.1.1. Gaussian Mixture Models (GMM)

A GMM is a probabilistic model that expresses a probability distribution as a weighted sum of multiple Gaussian component densities [75]. It has been extensively employed in biometric applications, particularly in emotion recognition, due to its capability to model data distributions with high flexibility. Each GMM is defined by a set of parameters, including mean vectors, covariance matrices, and mixture weights, which collectively describe how the data is distributed. Several variants of GMMs exist, and the selection of a specific model often depends on the available data and the type of application. Parameter estimation is typically performed using the Maximum Likelihood (ML) approach, most commonly implemented through the Expectation-Maximization (EM) algorithm, an iterative procedure that refines parameter estimates to maximize the likelihood of the observed data. When adaptation from a pre-trained model is needed, Maximum A Posteriori (MAP) estimation is used to update the parameters using a Universal Background Model (UBM). The Gaussian distribution used in GMMs is mathematically defined as follows:

$$N(x \mid \mu, \sum) = \frac{1}{(2\pi)^{d/2}\sqrt{|\sum|}} exp\left(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)\right) \quad (I\text{-}5)$$

where d represents the dimension of the feature vector, μ is the mean, and Σ denotes the covariance matrix [76].

I.10.1.2. K-Nearest Neighbors (KNN)

The k-NN algorithm is a non-parametric classification method that assigns a class to a data point based on the classes of its k nearest neighbors in the training dataset [77] [78]. Unlike parametric models, k-NN makes no assumptions about the underlying data distribution and relies solely on the features and labels of the stored samples. When classifying a new instance, the algorithm determines the *k* nearest neighbors of a given data point using one of several distance measures, namely Euclidean, Cityblock, Cosine, or Correlation and assigns the class label that occurs most frequently among these neighbors. The parameter k controls the number of neighbors considered and thereby influences the classification outcome. The basic K-NN classifier procedure is outlined as follows:

- **Calculate the distance** between the unknown test feature vectors $A\ (A_1, A_2, \dots A_d)$ and all the known training feature vectors $\mathrm{B}_j = \left(B_{j1}, B_{j2}, \dots, B_{jd}\right)$ using the four following metric distances:

$$\mathrm{distance}\left(A, B_j\right) = \begin{cases} \sqrt{\sum_{i=1}^{d} \left(A_i - B_{ji}\right)^2} & \text{for Euclidean} \\ \sum_{i=1}^{d} \left|A_i - B_{ji}\right| & \text{for Cityblock} \\ 1 - \dfrac{\sum_{i=1}^{d} A_i B_{ji}}{\sqrt{\sum_{i=1}^{d} A_i^2}\sqrt{\sum_{i=1}^{d} B_{ji}^2}} & \text{for Cosine} \\ \dfrac{1}{2}\left(1 - \dfrac{\sum_{i=1}^{d} \left(A_i - \bar{A}\right)\left(B_{ji} - \bar{B}_j\right)}{\sqrt{\sum_{i=1}^{d} \left(A_i - \bar{A}\right)^2}\sqrt{\sum_{i=1}^{d} \left(B_{ji} - \bar{B}_j\right)^2}}\right) & \text{for Correlation} \end{cases} \quad (I\text{-}6)$$

where d denotes the number of features, and $\bar{A}$ and $\bar{B}$ are the means of A_i and B_i respectively.

- Select the shortest k distances from the unknown test feature vector.
- Determine the most common class label among these k neighbors using a majority voting approach [79].

I.10.1.3. Support Vector Machines (SVM)

The SVM classifier is a supervised machine learning algorithm introduced by Vladimir Vapnik and his collaborators [80], [81]. Rooted in statistical learning theory, it was originally developed for binary classification problems and later extended to handle multi-class classification. The algorithm separates classes by finding an optimal hyperplane with the maximum margin, which ensures improved generalization performance.

Initially, SVM was limited to linear classification. However, with the introduction of kernel methods, SVMs gained the ability to perform nonlinear classification by implicitly mapping the original data into a higher-dimensional feature space. In this transformed space, the objective is to find a maximum-margin linear decision boundary. This effectively converts a nonlinear problem into a linearly separable one while preserving the convexity of the optimization.

A simple example is illustrated in *Figure I-7*, where one-dimensional data samples x_i are mapped to two dimensions using the transformation (x_i, x_i^2). This corresponds to using a polynomial kernel of degree 2, defined as:

$$K(x_i, x_j) = (x_i \cdot x_j + c)^2 \qquad (I\text{-}7)$$

which captures quadratic relationships without explicitly computing the transformation [72].

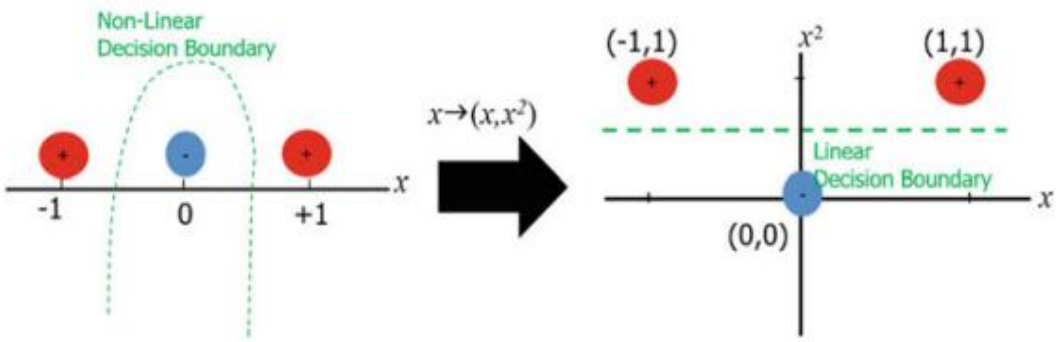


Figure I-7: Transforming data from a nonlinear space to a higher-dimensional linear space [82]

In practice, several kernel functions are commonly used, as summarized in *Table I-3* For the experiments in this thesis, the Radial Basis Function (RBF) kernel was employed due to its effectiveness in handling complex nonlinear relationships.

Table I-3:Common Kernel Functions Used in SVM

Kernel Name	Mathematical Expression
Polynomial kernel	$K(x_1, x_2) = ({x_1}^T, x_2)^n$
Radial basis function kernel	$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\delta^2}\right)$
Laplace kernel	$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|}{\delta}\right)$
Sigmoid kernel	$K(x_1, x_2) = \tanh\left[a({x_1}^T x_2) - b\right] a, b > 0$

I.10.2. Deep Learning Approaches

Deep learning is a subfield of machine learning within artificial intelligence, uses multi layered neural networks to automatically learn hierarchical feature representations from raw data. Deep learning models consist of interconnected layers of neurons that progressively

transform inputs into more abstract and informative representations through nonlinear mappings. CNN constitute a prominent class of deep learning architectures that excel at learning spatial or temporal patterns by applying convolutional operations. Originally devised for image recognition and computer vision, CNNs have been successfully adapted to one-dimensional data, such as time series and sequential signals, yielding models known as 1D-CNNs [83], [84].

I.10.2.1. 1D Convolutional Neural Networks (1D-CNN)

Overview of 1D CNN Architecture

The architecture of a 1D CNN primarily consists of sequential layers that process 1D data inputs through various stages. The fundamental components include convolutional layers, pooling layers, and fully connected layers, each contributing to the overall performance of the network in tasks such as data classification and feature extraction [85].

Convolutional Layers

The key component of a 1D CNN is the 1D convolutional layer, where filters or kernels slide over the input data in a single dimension. This layer is responsible for extracting local features from the input data, which are essential for understanding the underlying patterns [86].

Mathematically, the output of a one-dimensional convolution at position i in the L^{th} layer can be formulated as:

$$x_j^{(l)} = f\left(\sum_{i=1}^{M} x_i^{(l-1)} * k_{ij}^{(l)} + b_j^{(l)}\right) \quad (I\text{-}8)$$

where:

- $x_i^{(l-1)}$: the i^{th}input feature map from the previous layer,
- $k_{ij}^{(l)}$: the convolution kernel that connects input i to output j,
- $b_j^{(l)}$: the bias term associated with the j^{th}output,
- $*$: the one-dimensional convolution operator,
- $f(\cdot)$: the nonlinear activation function (such as ReLU, tanh), and
- M: the number of inputs from the previous layer.

The convolutional layers in a typical 1D CNN architecture are often configured in modules that may also include max-pooling layers to aid in down-sampling and feature abstraction.

Pooling Layers

Pooling layers are employed to reduce the dimensionality of the feature maps produced by the convolutional layers. By aggregating information over localized regions, pooling helps in reducing the number of parameters and computational load while also controlling overfitting

Common pooling methods used in 1D CNNs include max pooling and average pooling, which serve to retain the most significant features while discarding less informative data [87].

Activation Functions

Activation functions, such as the Rectified Linear Unit (ReLU), are utilized throughout the layers of a 1D CNN to introduce non-linearity into the model. ReLU is defined as:

$$f(z) = \max(0, z) \qquad (I\text{-}9)$$

Negative inputs are set to zero and positive inputs are passed unchanged, allowing the network to focus on relevant features while mitigating the vanishing gradient problem.

This non-linearity enables the network to learn complex patterns and relationships within the data. Each layer's activation output feeds into the subsequent layer, progressively refining the feature representation before final classification [88].

Fully Connected Layers

Following the convolutional and pooling layers, fully connected layers play a critical role in the final decision-making process of the network. The output from the convolutional and pooling layers is flattened and passed through one or more fully connected layers, where each neuron is connected to every neuron in the previous layer. These layers combine the high-level features extracted by the earlier layers to perform classification tasks [89].

The output of a neuron in a fully connected layer can be expressed as:

$$z = f(W \cdot x + b) \qquad (I\text{-}10)$$

where W is the weight matrix connecting the inputs to the neuron, x is the input vector from the previous layer, b is the bias term, and $f(\cdot)$ is the activation function (ReLU).

For the final output layer in a classification task, the softmax function is typically used to produce class probabilities:

$$\hat{y}_j = \frac{e^{z_j}}{\sum_k e^{z_k}} \quad (I\text{-}11)$$

where $\hat{y}_j$ is the predicted probability for class j.

I.11. DECISION FUSION USING VOTING RULES IN EMOTION RECOGNITION SYSTEMS

In AER, speech signals are processed in short time frames, generating a sequence of feature vectors that describe the emotional content over time. Two main strategies can be used for classification: the whole sequence can be classified directly with a temporal model such as an LSTM or HMM [90], or each frame can be classified independently using traditional classifiers like KNN, GMM, or SVM [91].

However, when classifying frame by frame, the predicted labels may change from one frame to another because of noise or speaker variations. To obtain a final and reliable decision for the whole signal, we therefore need a fusion method that combines all frame level predictions into a single result [92].

Among fusion techniques, voting rules have proven to be effective and widely adopted [93] [94] [95]. By considering the predictions across all frames, voting selects the most frequent class, thereby reducing the impact of misclassified frames and enhancing overall system stability.

In our work [96], we adopt the second approach: each feature vector is classified independently using traditional machine learning classifiers, and a voting rule is applied to produce the final emotion label for the signal. This strategy ensures consistent decisions across the temporal sequence while avoiding the complexity associated with temporal models.

The following section presents the main voting strategies commonly used in emotion recognition and discusses their contribution to improving decision reliability.

I.11.1. Definition of Voting Rule Strategy

A voting rule strategy is a technique used at the decision level to aggregate multiple individual predictions into a single, final decision[97] [98]. In the context of emotion recognition, it is particularly useful when multiple classifiers or predictions are available, for

example, from different segments of a signal or from various feature representations. The goal of voting is to improve system performance, enhance reliability, and reduce classification errors by leveraging agreement among predictions to determine the final output [99].

I.11.2. Types of Voting Rules

I.11.2.1. Max Voting

The first and most widely used voting method is max voting [100], also commonly known as majority voting or hard voting. This method operates by collecting the predicted class labels and selecting the one that receives the highest number of votes, as illustrated in *Equation (I-12)*. In this thesis, majority voting is adopted as the primary decision fusion strategy to determine the final emotion label from multiple frame-level predictions.

In our approach, we apply majority voting not across different classifiers, but across multiple predictions generated from the feature vectors of the same input sequence. For example, if a sequence yields the predicted labels [0,0,1], the final decision corresponds to the most frequent label, which is 0. This technique, also known as max voting or hard voting, selects the class that occurs most often among the individual predictions.

Formally, let $\{y_1, y_2, \dots, y_T\}$ be the set of predicted class labels corresponding to the T feature vectors extracted from a given input sequence. The final predicted label $\hat{y}$ is obtained using:

$$\hat{y} = \text{mode}(y_1, y_2, \dots, y_T) \qquad (I\text{-}12)$$

where mode(.) returns the class label that appears most frequently among the predictions.

Unlike soft voting, which averages or aggregates predicted probability distributions and is more computationally intensive, majority voting operates on discrete labels only, making it both simple and efficient in terms of implementation and memory usage. Despite its limitations in cases of uncertain or evenly split predictions, this method is well-suited to our sequence-based classification framework due to its low complexity and robustness.

I.11.2.2. Averaging Voting

The second technique is the *averaging voting* method [101]. In this approach, predictions generated by multiple models are combined by computing their average to produce the final

output. The final prediction $\hat{y}$ is calculated as the arithmetic mean of the individual model predictions:

$$\hat{y} = \frac{1}{M}\sum_{i=1}^{M} y_i \qquad (I\text{-}13)$$

where M is the total number of models and $y_i$represents the prediction from the i^{th}model.

I.11.2.3. A weighted Average Voting

The third voting strategy is weighted average voting, which represents a refined variant of the standard averaging voting approach [102]. This method assigns distinct weights to each base learner, reflecting the relative contribution of each model to the final prediction. The weighted average for each class is computed by multiplying the prediction of each classifier by its corresponding weight, summing these weighted predictions, and then dividing by the total sum of the classifier weights, as expressed in the following equation:

$$\hat{y} = \frac{\sum_{i=1}^{M} w_i\, y_i}{\sum_{i=1}^{M} w_i} \qquad (I\text{-}14)$$

where $\hat{y}$is the final predicted value, $M$is the total number of base learners, $y_i$is the prediction of the i-th base learner, and $w_i$is the weight assigned to that learner.

I.12. CONCLUSION

In this chapter, we presented an overview of emotion recognition, highlighting its applications and the different modalities through which it can be achieved. Our focus was placed on speech and facial expression, given their wide use and effectiveness in emotion recognition tasks. We reviewed the state of the art in both domains and introduced a general overview of classification systems. Furthermore, we discussed feature extraction techniques and classification algorithms, including both traditional machine learning and deep learning approaches. Finally, we introduced the concept of combining machine learning algorithms with voting rules to enhance classification performance.

II

FEATURE SELECTION

II.1. INTRODUCTION

In recent years, the explosion of high-dimensional data has presented both opportunities and challenges for machine learning, deep learning and pattern recognition applications. One of the key challenges lies in the presence of redundant or irrelevant features, which can degrade model performance, increase computational cost and memory usage. This is particularly evident in automatic emotion recognition systems, where datasets often contain a large number of features extracted from speech, facial expressions, or physiological signals. In such contexts, feature selection becomes a vital preprocessing step to select the most relevant features that contribute to accurate and robust classification.

Feature selection refers to the process of selecting a subset of relevant features from the original set without altering the original representation of the data. Unlike feature extraction methods, which transform data into a new space, feature selection retains the original semantics of the features, making the models more interpretable and often more efficient. The goals of feature selection are manifold: to reduce dimensionality, lower computational cost and memory usage, and enhance system performance.

In the context of emotion recognition, the need for efficient feature selection is amplified by the curse of dimensionality. Selecting the right subset of features can significantly improve system performance. Various feature selection techniques have been proposed in literature, broadly categorized into filter methods, wrapper methods, and embedded methods, each with its advantages and limitations.

This chapter presents an in-depth overview of feature selection techniques, with a particular focus on MI-based approaches. MI offers a powerful criterion for measuring the statistical dependence between features and class labels, allowing the selection of features that are both relevant and non-redundant. It has demonstrated promising performance in balancing relevance and redundancy.

II.2. FEATURE REDUNDANCY AND IRRELEVANCE IN HIGH DIMENSIONS

In high dimensional data spaces, such as those commonly faced in emotion recognition, not all features contribute equally to the classification task. The presence of irrelevant features and redundant features often hinders the classification accuracy.

Irrelevant features are those that carry little or no discriminative information with respect to the target classes. Their inclusion increases the dimensionality of the feature space without improving the classification accuracy, which can result in longer training times, higher computational costs and memory usage [103].

Redundant features, on the other hand, may contain information that is already captured by other features. While they may still have some relationship with the target variable, their contribution is largely overlapping. This redundancy increases the dimensionality unnecessarily, making the feature space more complex without yielding new or complementary information [104].

The presence of irrelevant and redundant features is particularly problematic in high-dimensional settings such as speech and facial, where feature vectors often comprise hundreds or even thousands of dimensions. In such cases, the curse of dimensionality exacerbates the issue, as dimensionality increases, data points become sparse, distance metrics lose discriminative power, and the risk of overfitting grows [105].

To address these challenges, feature selection techniques aim to select a subset of features that maximizes relevance while minimizing redundancy. Approaches based on MI are particularly well suited for this task, as they measure both the dependency between features and class labels (relevance) and the dependency among features themselves (redundancy) [106]. By filtering out irrelevant and redundant dimensions, such methods not only improve recognition accuracy but also enhance model interpretability and reduce computational burden.

II.3. CURSE OF DIMENSIONALITY

The term *curse of dimensionality*, also referred to as the *peaking phenomenon*, was first introduced by Richard E. Bellman [107] to describe the exponential increase in data volume and complexity as the dimensionality of the feature space grows. From a theoretical perspective,

adding more features should provide additional information, which in turn could improve the performance of learning algorithms. However, in practice, this assumption does not always hold. As dimensionality increases, the data often becomes sparse, leading to redundancy and noise that can negatively impact the learning process [52].

In machine learning and pattern recognition, this phenomenon manifests as a performance degradation when too many features are introduced [108]. Initially, as the number of features increases, recognition performance tends to improve since informative features contribute positively to discrimination. Nevertheless, beyond a certain point, the performance curve reaches a peak and subsequently declines [109]. This decline results from the inclusion of irrelevant or redundant features, which complicate the model, increase computational costs, and often cause overfitting.

II.4. FEATURE SELECTION PROCESS

In scientific literature, the concept of feature selection has been defined in multiple ways. While many authors propose similar views, a few definitions highlight distinct conceptual aspects. Some of the most representative definitions are outlined below:

- Kira et al. [110] define feature selection as *"the task of identifying the minimal subset of features that is both necessary and sufficient for representing the target concept."*
- Narendra et al. [111] describe it as *"the process of selecting a subset of M features from an initial set of N features* $(where\ M < N)$*, in which a criterion function is optimized across all possible subsets of size* M*."*
- Koller et al. [112] emphasize dimensionality reduction, defining feature selection as the task of selecting a subset of features that reduces the dimensionality of the original set while preserving comparable classification accuracy.
- Kohavi et al. [113] formulate feature selection as the task of identifying a compact subset of features such that the resulting class distribution, based on the selected features, remains as close as possible to the distribution obtained when using the full feature set.

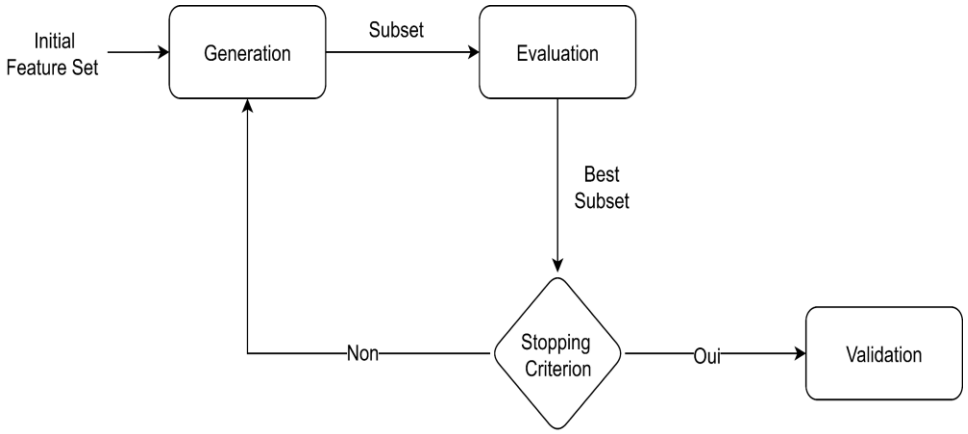


Figure II-1:General Feature Selection Process [114]

Building on these conceptual definitions, a typical feature selection procedure can be described as a systematic process involving several key steps, as illustrated in *Figure II-1*. These steps are summarized as follows:

- **Generation**: This step defines how the search space is explored to generate candidate subsets of features.
- **Evaluation**: This step measures the quality or relevance of each feature.
- **Stopping criterion**: This specifies when the selection process should terminate.
- **Validation**: This step verifies whether the selected subset meets the intended objective.

II.4.1. Generation Procedures

The generation procedure consists of generating all possible subsets from an initial feature set of size N, which results in a search space of 2^N combinations. This exhaustive strategy provides complete coverage and ensures that the globally optimal subset is identifiable. However, the exponential growth of the search space with respect to the number of features makes this approach computationally impractical, even for moderately sized feature sets. To address this limitation, alternative generation strategies have been introduced, including complete generation, heuristic generation, and random generation [115].

II.4.1.1. Complete

Complete generation procedures perform a full search to identify the optimal subset of features according to an evaluation measure function. A method is regarded as complete if it guarantees to always return the optimal subset. It is important to distinguish between a complete search

and an exhaustive search. Exhaustive search is always complete because it evaluates every possible subset, ensuring that the best one is identified. However, a complete search is not exhaustive, specifically when the evaluation measure is monotonic, it is unnecessary to examine all subsets to identify the optimal one. Although complete search remains computationally expensive, it generally evaluates fewer subsets than exhaustive search [114].

II.4.1.2. Random

This search method begins with a randomly selected subset and can proceed in two main ways. The first is to generate candidate subsets in a random manner, referred to as the Las Vegas algorithm [116]. The second integrates randomness into sequential search procedures. The role of randomness is to help escape local optima within the search space. The computational complexity of these approaches is generally in the order of $O(N^2)$.

II.4.1.3. Heuristic

Heuristic search methods, also known as sequential or greedy approaches, sacrifice the guarantee of finding the optimal subset and therefore risk overlooking the best solution. These approaches are generally grouped into three types: forward selection, backward elimination, and bidirectional selection. At each iteration of the generation procedure, these methods add or remove features. With a search space complexity on the order of $O(N^2)$, heuristic methods are simple to implement and fast in generating feature subsets [117].

II.4.1.3.1. Forward

Forward selection begins with an empty feature set and iteratively adds features one by one. At each step, the feature whose addition provides the greatest improvement with respect to the evaluation criterion is selected. This process continues until adding new features no longer improves performance or until a stopping criterion is reached.

II.4.1.3.2. Backward

On the other hand, backward elimination starts with the full set of features and progressively removes them. At each iteration, the feature that contributes the least to the evaluation criterion is eliminated. This procedure continues until no further improvement is possible or until the subset reaches a predefined K features.

II.4.1.3.3. Bidirectional

Bidirectional selection combines both forward selection and backward elimination. The process can begin either with an empty set or with the full feature set, and features are both added and removed during the search.

II.4.2. Feature Evaluation

An optimal feature subset depends on the evaluation function used, meaning a subset considered optimal by one function may not be optimal by another. Generally, evaluation functions aim to measure the discriminative ability of a feature or subset to separate different class labels. According to the literature, evaluation functions can now be broadly divided into five categories: distance or divergence measures, information measures, dependence measures, consistency measures, and classifier error rate measures. Each category is explained in the following subsections [114].

II.4.2.1. Distance Measures

This evaluation function is also referred to as a divergence, separability, or discrimination measure. In the case of a two-class problem, a feature X is considered more discriminative than a feature Y if X induces a greater distinction between the conditional probability distributions of the two classes. Conversely, if this difference equals zero, then X and Y are regarded as indistinguishable. A typical example of this type of measure is the Euclidean distance [118].

II.4.2.2. Information Measures

This measure relies on the information gain of the features. The information gain of a feature is defined as the reduction in uncertainty, calculated as the difference between prior uncertainty and the expected posterior uncertainty. Information gain reaches its maximum when the classes are equally probable, while uncertainty is minimized under this condition [119].

II.4.2.3. Dependence Measures

Dependence measures, also known as correlation or similarity measures, evaluate how much the value of one variable can be predicted from another. In the context of feature selection for classification tasks, the correlation coefficient is a classical measure of dependence and is often used to quantify the relationship between a feature and a class. If the correlation of feature X with class C exceeds that of feature Y with C, then X is considered more relevant than Y. A

related approach evaluates the dependence of a feature on other features, providing an indication of its redundancy [120].

II.4.2.4. Consistency Measures

Consistency measures rely strongly on class label information and use the Min-Features bias, aiming to select the smallest subset of features that separates classes as effectively as the full set. An inconsistency arises when instances share identical feature values but differ in class labels [121].

II.4.2.5. Classifier Error Rate Measures

Evaluation functions based on classifier error rates are commonly referred to as wrapper methods, in which the classifier itself serves as the evaluation function. In this approach, feature selection is directly guided by the performance of the classifier, which then uses the selected features to predict the class labels of unseen data. This strategy typically yields high classification accuracy; however, it is associated with significant computational cost.

II.4.3. Stopping Criterion

The stopping criterion (SC) determines when the feature selection process should terminate. Since the optimal number of features is usually unknown, several approaches have been proposed.

A first approach is the fixed threshold, where the procedure stops once a predefined number of features, iterations, or computational budget is reached. This method is easy to apply but may result in suboptimal subsets [122].

Another common approach is performance stabilization, which terminates the search when adding or removing features no longer improves the classification performance [123].

For small subsets, exhaustive exploration can be used, where the search is completed only after all subsets have been evaluated. While this ensures optimality, it is computationally expensive in high dimensions.

More specific approaches are based on the recognition rate (RR). The first, SC1 (Relative RR Criterion), considers a subset S as optimal if:

$$RR(S) \geq RR(F) \qquad (II\text{-}1)$$

where F is the full feature set. This criterion is efficient since the process stops once the target accuracy is achieved, often yielding smaller subsets [124].

The second, SC2 (Maximum RR Criterion), defines the optimal subset S^* as:

$$RR(S^*) = \max_{S \subseteq F} RR(S) \quad (II\text{-}2)$$

This approach ensures the highest recognition rate but requires more computation as more subsets must be examined [125] [126].

In this thesis, SC1 and SC2 are used to guide the feature selection process, ensuring both efficiency and optimal recognition performance.

II.4.4. Validation

Two main validation strategies are commonly used for feature selection methods: (i) evaluation using artificial datasets and (ii) evaluation using real-world datasets. Artificial datasets are designed with known relevant and irrelevant features, allowing direct comparison between the selected subset and known features. In contrast, real-world datasets do not provide such prior knowledge. Here, validation is performed by evaluating the impact of the selected subset on the performance of a learning algorithm, for example by comparing classification error rates obtained with the full feature set and with the reduced subset.

II.5. FEATURE SELECTION APPROACHES

This section provides an overview of information theory as well as the different feature selection strategies.

Feature selection plays a pivotal role in machine learning and pattern recognition, as it focuses on reducing dimensionality by identifying and selecting the most relevant subset of features from an initial set. This process enhances computational efficiency, minimizes memory requirements, and can potentially improve system accuracy by mitigating the effects of the curse of dimensionality [127].

Feature selection approaches are generally categorized into three main groups: Wrapper approaches [128], Filters approaches and embedded approaches.

II.5.1. Filter approaches

Filter approaches select relevant features by relying on a performance metric, independently of the classification system's performance. Once the optimal features are selected, they can be used by classification algorithms. Broadly, feature filtering measures can be categorized into distance [118], consistency [121], dependency [120], and information measures [119]. They offer several advantages, including efficient scalability to high-dimensional datasets, computational simplicity and speed, and independence from the classification algorithm [129].

Univariate filter methods generally evaluate and rank individual features in isolation, while multivariate filter methods assess the joint contribution of feature subsets, thereby capturing interactions and redundancies among features.

A wide range of filter-based feature selection methods have been proposed in the literature, each belonging to a specific class depending on the evaluation criterion. *Table II-1* summarizes representative filter methods, their filter class, applicable tasks, and key references.

In this thesis, we focus on feature selection methods based on criterion of mutual information maximization to assess both the relevance and redundancy of features.

Table II-1:Common Filter Methods for Feature Selection [130]

Method	**Filter Class**	**Applicable Task**	**Studied by**
Information Gain	Univariate, Information	Classification	[131]
Gain Ratio	Univariate, Information	Classification	[132]
Symmetrical Uncertainty	Univariate, Information	Classification	[133]
Correlation	Univariate, Statistical	Regression	[133]
Chi-square	Univariate, Statistical	Classification	[132]
Minimum Redundancy Maximum Relevance (mRMR)	Multivariate, Information	Classification, Regression	[134]
JMI (Joint Mutual Information)	Multivariate, Information	Classification, Regression	[135]

CIFE (Conditional Infomax Feature Extraction)	Multivariate, Information	Classification, Regression	[136]
ICAP (Interaction Capping)	Multivariate, Information	Classification, Regression	[137]
Fast Correlation-Based Filter (FCBF)	Multivariate, Information	Classification	[133]
Fisher Score	Univariate, Statistical	Classification	[138]
Relief / ReliefF	Univariate, Distance	Classification, Regression	[139]
Spectral Feature Selection (SPEC) / Laplacian Score	Univariate, Similarity	Classification, Clustering	[140]
Feature selection for sparse clustering	Multivariate, Similarity	Clustering	[141]
Localized Feature Selection Based on Scatter Separability (LFSBSS)	Multivariate, Statistical	Clustering	[142]
Multi-Cluster Feature Selection (MCFS)	Multivariate, Similarity	Clustering	[140]
Feature weighting K-means	Multivariate, Statistical	Clustering	[143]
ReliefC	Univariate, Distance	Clustering	[144]

II.5.2. Wrapper approaches

Wrapper approaches involve selecting a subset of features from the original feature set and evaluating its relevance based on the performance of a classification algorithm. This process requires training and testing for each potential feature subset, making wrapper methods computationally expensive and time-consuming, particularly when dealing with high-dimensional feature sets [145].

II.5.3. Embedded approaches

Embedded approaches integrate feature selection directly into the model training process, allowing for the simultaneous optimization of the model parameters and the selection of relevant features [146].

Although mutual information can, in principle, be employed in wrapper and embedded approaches, it has proven to be most effective in filter-based approaches due to their scalability, computational efficiency, and independence from any specific classifier. For this reason, the following section is dedicated to mutual information-based filter methods in high-dimensional settings.

II.6. INFORMATION-THEORETIC FOUNDATIONS FOR FEATURE SELECTION

II.6.1. Fundamentals of Information Theory

Entropy and mutual information are fundamental concepts established within information theory [147]. Initially introduced in the context of communication theory, information theory aimed to address key questions related to data compression and transmission efficiency [148]. Over time, its theoretical foundations have been widely adopted in the fields of machine learning and feature selection.

II.6.1.1. Definition of Entropy

Entropy is a measure of uncertainty or randomness in a random variable. For a discrete random variable $X$with alphabet $\mathcal{X}$and probability mass function $p(x) = \Pr\{X = x\}$, for $x \in \mathcal{X}$, the entropy $H(X)$is defined as:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x)\log_2\ p(x) \qquad (II\text{-}3)$$

The logarithm is taken to base 2 (default base), and the entropy is therefore expressed in bits. We use the convention $0\log(0) = 0$, which is justified by continuity since $x\log(x) \to 0$ as $x \to 0$. Therefore, adding outcomes with zero probability does not affect the entropy value. If the logarithm is taken in base b, the entropy is denoted as $H_b(X)$. If the base of the logarithm is e, the entropy is expressed in nats.

It should be noted that entropy depends solely on the probability mass function $p(x)$of the random variable $X$and not on the specific values of the alphabet $\mathcal{X}$. Moreover, Shannon entropy is bounded as follows:

$$0 \leq H(X) \leq \log_2 |\mathcal{X}| \quad (II\text{-}4)$$

where $|\mathcal{X}|$denotes the number of elements in the alphabet. Equality is achieved if and only if the distribution $p(x)$is uniform over $\mathcal{X}$, meaning all outcomes are equally likely [147].

Intuitively, a high entropy value indicates that all events occur with approximately equal probability, reflecting maximum uncertainty. In contrast, low entropy implies that event probabilities differ significantly, indicating a more predictable outcome.

II.6.1.2. Joint Entropy and Conditional Entropy

II.6.1.2.1. Joint Entropy

The joint entropy $H(X,Y)$ of two discrete random variables X and Y, quantifies the total uncertainty associated with the simultaneous observation of both variables. It is defined as follows:

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y) \quad (II\text{-}5)$$

where $p(x,y)$ denotes the joint probability of the simultaneous occurrence of $x \in \mathcal{X}$and $y \in \mathcal{Y}$.

The value of joint entropy $H(Y,X)$ lies within the following range:

$$max\{H(X), H(Y)\} \leq H(X,Y) \leq H(X) + H(Y) \quad (II\text{-}6)$$

The upper bound is attained when the variables X and Y are fully independent, indicating that their combined uncertainty equals the sum of their individual uncertainties. In contrast, the lower bound is realized when X is completely determined by Y, reflecting total dependence between the two variables [147].

II.6.1.2.2. Conditional Entropy

Conditional entropy quantifies the amount of uncertainty that remains in a random variable X given that the value of another variable Y is known. In other words, it measures the residual uncertainty in X after incorporating the information provided by Y.

The conditional entropy reaches its minimum value of zero when X is completely determined by Y, meaning that knowing Y entirely eliminates uncertainty about X. Conversely, it reaches its maximum when X and Y are statistically independent, in which case knowledge of Y provides no information about X and does not reduce its uncertainty.

Formally, the conditional entropy is defined as:

$$H(X \mid Y) = -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P(x,y) \log_2 P(x \mid y) \quad \textit{(II-7)}$$

The conditional entropy satisfies the following inequality:

$$0 \leq H(X \mid Y) \leq H(X) \quad \textit{(II-8)}$$

Here, $H(X \mid Y = y_j)$ represents the entropy of the values x_i that are associated with the condition $Y = y_i$. In other words, it quantifies the uncertainty of X given that Y takes the value y_i .

An alternative and equivalent representation of conditional entropy is given by:

$$H(X \mid Y) = H(X,Y) - H(Y) \quad \textit{(II-9)}$$

II.6.1.3. Mutual Information (MI)

MI is a fundamental concept in information theory that quantifies the amount of information shared between two random variables. It measures the reduction in uncertainty of one variable due to the knowledge of the other [147].

For two discrete random variables X and Y , the mutual information is defined as:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)} \right) \quad \textit{(II-10)}$$

Where:

- x and y denote individual outcomes (samples) of the random variables X and Y, respectively.
- $P(x)$ and $P(y)$ represent the marginal probabilities of X and Y.
- $P(x,y)$ denotes the joint probability distribution of X and Y.

Alternatively, it can be expressed in terms of entropy:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (II\text{-}11)$$

When the variables are continuous, probabilities are replaced by probability density functions, and the summations become integrals:

$$I(X;Y) = \int_Y \int_X p(x,y)\log_2 \left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy \quad (II\text{-}12)$$

Figure II-2 illustrates the conceptual relationship between entropy and MI.

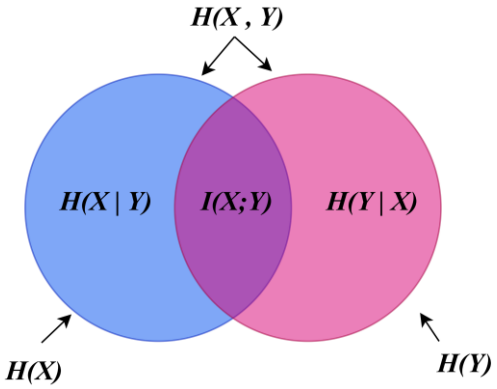


Figure II-2: Venn Diagram Illustrating Mutual Information [136]

II.6.1.4. Multivariate Mutual Information (MMI)

While MI provides a measure of dependency between two random variables, many real-world problems, including feature selection, involve interactions among three or more variables. To capture these higher-order dependencies, MMI also known as interaction information, has been introduced as an extension of MI to multiple variables [149].

For the case of three variables, called triple mutual information [150], the definition differs between discrete and continuous variables.

For discrete random variables X, Y, Z with joint probability mass function $p(x,y,z)$, triple mutual information is defined as:

$$I(X;Y;Z) = \sum_{x,y,z} p(x,y,z)\log \left(\frac{p(x,y)p(x,z)p(y,z)}{p(x,y,z)p(x)p(y)p(z)}\right) \quad (II\text{-}13)$$

For continuous random variables X, Y, Z the triple mutual information is expressed using probability density functions as follows:

$$I(X;Y;Z) = \int_z \int_y \int_x f(x,y,z)\ln \frac{f_{XY}(x,y)f_{XZ}(x,z)f_{YZ}(y,z)}{f(x,y,z)f_X(x)f_Y(y)f_Z(z)} dxdydz \quad (II\text{-}14)$$

where $f(x, y, z)$ is the joint density and $f_X(x), f_Y(y), f_Z(z)$ are the marginal densities. This can also be equivalently written as:

$$I(X;Y;Z) = I(X;Y) - I(X;Y \mid Z) = I(X;Z) - I(X;Z \mid Y) \\ = I(Y;Z) - I(Y;Z \mid X) \qquad (II\text{-}15)$$

Unlike the two-variable MI, the multivariate extension can take positive, negative, or zero values. A positive value reflects redundancy, or overlapping information among variables, while a negative value indicates synergy, where the joint contribution provides more information than the sum of pairwise interactions. In the context of feature selection, this property is particularly useful as it allows one to distinguish between redundant and synergistic feature interactions, thereby enabling the design of more effective selection strategies in high-dimensional data [150].

To better illustrate this concept, *Figure II-3* shows a Venn diagram of three variables, where the central overlapping region corresponds to the multivariate mutual information, highlighting redundancy or synergy depending on its sign.

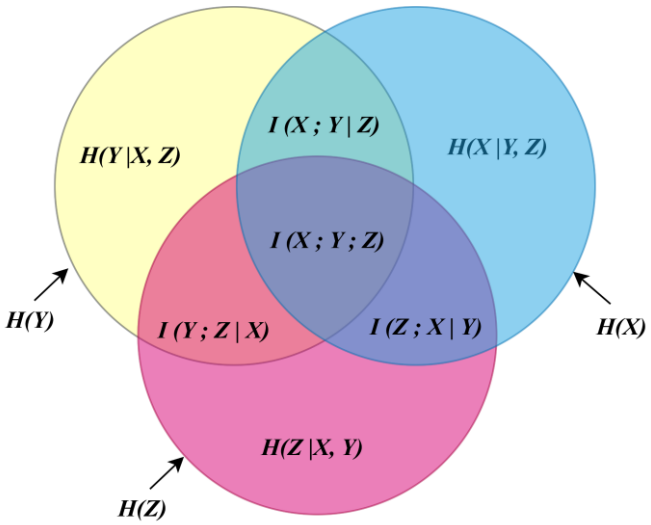


Figure II-3:Venn Diagram Illustrating Triple Mutual Information [147]

II.6.2. Mutual Information Estimation

The computation of MI relies on estimating both marginal and joint probability distributions. In practice, obtaining their exact values is impossible; therefore, approximation methods are required. Regardless of the chosen approach, accurate estimation of these distributions is essential, as it directly affects the precision of the MI value and, consequently, the overall system performance.

A common and straightforward technique for probability estimation is the histogram method, which was adopted in this study due to its simplicity and computational efficiency [151]. Alternative methods include Parzen windows [152], kernel density estimation [153], and the GMM approach [154], which has also been adopted in the literature.

A limitation of the histogram method lies in selecting an appropriate number of bins, k, during the discretization process. To address this, several heuristic rules have been proposed, the most widely used being Sturges' rule, Scott's rule, Freedman–Diaconis rule, and HGR.

Heuristic Rules for Bin Selection

Sturges' Rule [155]: The bin width w is defined as

$$w = \frac{r}{1 + \log_2(n)} \qquad (II\text{-}16)$$

where r is the range of the data and n is the number of observations. The number of bins is then computed as $k = r/w$.

Scott's Rule [156]: The bin width w is given by

$$w = \frac{3.49\sigma}{n^{\frac{1}{3}}} \qquad (II\text{-}17)$$

where σ is the standard deviation of the data set and n is the number of observations.

Freedman–Diaconis Rule [157]: The bin width w is defined as

$$w = \frac{2 \cdot \mathrm{IQR}}{n^{1/3}} \qquad (II\text{-}18)$$

where IQR is the interquartile range (the difference between the 75th and 25th percentiles) and n is the number of observations. The number of bins is then obtained as $k = r/w$.

HGR Rule [158]: The number of bins k used to estimate mutual information with the joint histogram method is given by:

$$k = \mathrm{round}\left\{\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\sqrt{L}}\right\} \qquad (II\text{-}19)$$

$$\text{avec } L = \frac{N\rho^2}{12(1-\rho^2)}\left(\alpha_x^2 + \alpha_y^2\right) \qquad (II\text{-}20)$$

where $\alpha_x$and α_y are constants

ρ is the correlation coefficient

In the case of a Gaussian distribution, the number of bins k for the histogram-based MI estimator is given by:

$$k = \text{round}\left\{\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\sqrt{\frac{6N\hat{\rho}^2}{1-\hat{\rho}^2}}}\right\} \quad (II\text{-}21)$$

where ρ is the estimator of the correlation coefficient.

In this study, Sturges' rule was chosen for bin estimation, as it provides a simple and effective approach for moderate-sized datasets, balancing accuracy and computational efficiency compared to alternative rules such as Scott's and Freedman–Diaconis [159].

II.7. MUTUAL INFORMATION-BASED FILTER METHODS IN HIGH DIMENSIONS

In the process of a feature selection strategy, it is crucial to define an evaluation or relevance criterion capable of quantifying the importance of each feature with respect to the classification task. MI is among the most widely used criterion functions due to two key advantages: (i) it captures both linear and nonlinear dependencies between features and the target, making it more flexible and robust than correlation-based methods that detect only linear relationships [160]; and (ii) it facilitates the selection of features that maximize the shared information with the target variable, often leading to improved model performance, particularly in high-dimensional datasets where many features may be redundant or irrelevant [161].

The $MI(C;X)$ between a feature X and the class label C represents the reduction in the uncertainty of C provided by the knowledge of X. This concept can be readily extended to groups of features. However, when handling a large number of features in a dataset, exploring all possible combinations becomes computationally difficult because the number of possibilities grows exponentially, as shown by the law of combinations.

$$C\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \quad (II\text{-}22)$$

To mitigate this challenge, iterative greedy algorithms such as Sequential Forward Selection (SFS) are commonly used. Originally introduced by Battiti [119], SFS incrementally selects feature subsets by adding one feature at a time according to predefined relevance criteria, thereby significantly reducing computational complexity compared to exhaustive combinatorial

searches. This strategy achieves a balance between computational efficiency and the capacity to identify informative features. In this context, applying a feature selection procedure based on a greedy forward selection approach that uses MI as the relevance criterion has proven to be a highly effective solution [162].

Algorithm II-1 (pseudo-code) illustrates the standard forward greedy search procedure applied in combination with mutual information as the selection criterion:

Algorithm II-1: Forward Greedy Feature Selection with MI

Input:

Features set $F = \{X_1, X_2, \ldots, X_n\}$

Desired number of features k

Output:

Subset of selected features $S = \{X_{p1}, X_{p2}, \ldots, X_{Pk}\}$

Algorithm Steps:

Initialization:

Initialize F with n features

Set S as an empty subset

Calculation of MI:

For each feature X_i *in F, calculate* $MI(C;X_i)$

Select the first feature

Identify X_{p1} *maximizing* $MI(C;X_i)$

Remove X_{p1} *from F and add it to S*

Greedy selection (continue iterating from $j = 2$ ***until*** k ***features reached:***

while *size of* $S \leq k$ ***do***

for *each feature* X_i *in F* ***do***

Calculate $MI(C;S,X_i)$

end for

Select X_{pj} *from F maximizing* $MI(C;S,X_i)$ *at step j*

Remove X_{pj} *from F and add it to S*

Output:

Return subset S containing the selected k features

Calculating MI from data requires estimating the probability density, which cannot be accurate for high-dimensional features. As a result, most algorithms use measures based on at most three variables (two features plus the class index).

We present the most common methods for feature selection based on mutual information criteria, in which the order is limited to a maximum of three variables.

The first strategy, known as Mutual Information Maximization (MIM), establishes the simplest criterion for feature selection at step $j+1$. At this step, the feature $X_{p_{(j+1)}}$ is determined as follows:

$$X_{p_{j+1}} = \arg \max_{X_i \in \mathcal{F} - S_j} MI(C; X_i) \qquad (II\text{-}23)$$

At this stage, $S_j = S_{j-1} \cup \{X_{p_j}\}$ denotes the subset updated to include the new feature X_{p_j} selected at step j. In this procedure, the relevance of each feature X_i is assessed individually, without considering redundancy with the features already included in S_j for predicting the class label C. This may result in the selection of redundant features that provide redundant information about C, an issue that must be addressed. To tackle this problem, several strategies have been proposed to optimize feature relevance while controlling redundancy [104], [151]. In the next subsection, we focus on four such strategies: mRMR, JMI, CIFE, and ICAP.

II.7.1. Feature Selection Based on Mutual Information Maximization Criteria

II.7.1.1. The Maximum-Relevance Minimum Redundancy (MRMR)

mRMR was introduced by Peng [134], [104] to overcome the MIM issue by enhancing relevance with the class variable and minimizing redundancy between the selected features. The mRMR formula is given as follows:

$$X_{p_{j+1}} = \arg \max_{X_i \in \mathcal{F} - S_j} \left[MI(C; X_i) - \frac{1}{|S|} \sum_{k=1}^{j} MI\left(X_i; X_{p_k}\right) \right] \qquad (II\text{-}24)$$

II.7.1.2. Joint Mutual Information strategy (JMI)

JMI was introduced by Yang and Moody [135]. The JMI approach examines relevance and redundancy by computing the mean value while incorporating the class label during MI

calculation. JMI and MRMR share strong similarities, but their key distinction is how they handle conditional redundancy. JMI formula is represented as follows:

$$X_{p_{j+1}} = \arg \max_{X_i \in \mathcal{F}-S_j} \left[\mathrm{MI}(C;X_i) - \frac{1}{|S|} \sum_{k=1}^{j} \left[MI(X_i;X_{pk}) - \mathrm{MI}\left(X_i;X_{pk}|C\right) \right] \right] \quad (II\text{-}25)$$

II.7.1.3. Conditional Infomax Feature Extraction strategy (CIFE)

Lin and Tang [136] introduced the CIFE criterion. This method focuses on maximizing the information relevant to the joint class by explicitly minimizing redundancies between class-relevant features. The criterion can be expressed as follows:

$$X_{p_{j+1}} = \arg \max_{X_i \in \mathcal{F}-S_j} \left[MI(C;X_i) - \sum_{k=1}^{j} \left[MI(X_i;X_{p_k}) - MI(X_i;X_{p_k}|C) \right] \right] \quad (II\text{-}26)$$

II.7.1.4. The Interaction Capping (ICAP)

In the ICAP criterion [137], interaction information is integrated for feature selection. This criterion can be expressed as follows:

$$X_{p_{j+1}} = \arg \max_{X_i \in \mathcal{F}-S_j} \left[MI(C;X_i) - \sum_{k=1}^{j} max\left[MI(X_i;X_{p_k}) - MI(X_i;X_{p_k}|\,C), 0 \right] \right] \quad (II\text{-}27)$$

II.8. CONCLUSION

In this chapter, we review the main feature selection approaches commonly discussed in the literature, namely Wrapper, Filter, and Embedded methods. The Wrapper approach evaluates subsets of features based on the performance of a classifier. While effective, it is computationally expensive and thus more suitable for problems with low-dimensional features. In contrast, the Filter approach operates independently of any classifier, evaluating features based on predefined criteria such as information measures, statistical tests, similarity, or distance metrics. This independence renders Filter methods computationally efficient and

particularly well-suited for high-dimensional features. Embedded methods, on the other hand, perform feature selection during the model training process itself.

In our study, we adopt a Filter-based approach using MI for feature selection in the context of emotion recognition, motivated by the high dimensionality of the features typically generated in this domain. MI is particularly advantageous because, unlike linear correlation measures, it can capture non-linear dependencies between variables, which is crucial for effectively modeling complex emotional patterns.

The standard Filter-based MI method is often combined with forward greedy search, where a predefined number of K features is selected to complete the process. However, this approach suffers from the lack of an effective stopping criterion, which may either lead to over-selection or premature termination of the process. To address this limitation, we introduce a deterioration factor into the forward greedy algorithm. This modification provides a trade-off between the classification accuracy and the number of selected features. The details of this proposed strategy will be discussed thoroughly in *Chapter III.*

III

APPLICATION OF FEATURE SELECTION FOR SPEECH EMOTION RECOGNITION

III.1. INTRODUCTION

In this chapter, we explore the application of feature selection methods to SER systems, with the goal of improving classification accuracy while reducing system complexity. SER systems typically involve extracting a large set of acoustic features from speech signals, such as MFCC, PLP, and LPCC. However, using high-dimensional feature vectors often leads to increased computational cost, memory usage, and potential performance degradation due to the curse of dimensionality. Feature selection thus becomes essential to retain only the most relevant features, thereby optimizing both model performance and resource efficiency.

Many existing approaches in the literature classify entire sequence-level feature vectors directly, often without applying any prior feature selection. This overlooks important factors such as computational time and memory consumption, especially in real-time or embedded contexts.

The work presented in this chapter is based on two published studies. The first focuses on MI-based feature selection combined with lightweight machine learning classifiers, namely SVM, KNN, and GMM, using a frame-level classification approach followed by a voting rule to aggregate decisions at the signal level [96]. The second study extends this methodology to a deep learning context, where a one-dimensional CNN is combined with MI-based feature selection guided by a stopping criterion [163]. In both cases, experiments are conducted on the EMO-DB dataset using various feature sets and selection strategies.

This chapter presents the design of each system, the applied feature selection techniques, and a comparative analysis of their performance and efficiency.

III.2. DATASET DESCRIPTION FOR SER

In the present study, the Berlin Emotional Speech Database (EmoDB) was selected as the benchmark dataset for evaluating the proposed SER system. EmoDB is widely used in the literature because it is balanced across genders and emotional classes, recorded under controlled conditions, and covers a broad range of emotional states. A detailed description of the corpus is provided below.

The database is composed of ten unique German sentences taken from common conversations, divided into two subsets: Set A with five short sentences and Set B with five longer ones. Ten voice actors, evenly split between male and female, pronounced these sentences while emulating seven primary emotions: Anger (Angry), Boredom (Bored), Disgust (Disgust), Fear (Fearful), Happiness (Happy), Sadness (Sad), and a Neutral (Neutral) state. The database, composed of 535 utterances, was initially recorded at 48 kHz and then down-sampled to 16 kHz. For this study, Set A, with 277 utterances, served as the training dataset, while Set B, with 258 utterances, was used for testing. It is essential to point out that the testing sentences differ in content from the training ones, resulting in an SER system that operates in an independent text mode [164]. *Table III-1* provides a comprehensive breakdown of the number of occurrences used during the testing and training phases.

Table III-1:Distributing sentences from the EmoDB database across the 7 emotional states, both for testing and training purposes.

Emotional states	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral
Number	127	81	46	69	71	62	79
Testing	62	40	21	34	33	30	38
Training	65	41	25	35	38	32	41

III.3. FEATURE EXTRACTION

The extraction of features involves converting each raw speech signal into a structured sequence of low-dimensional, informative vectors suitable for the classification stage. This transformation is achieved through a sequence of signal analysis steps, including preprocessing, framing and windowing, and finally feature vector computation. These steps are designed to

retain the key characteristics of the speech signal while reducing noise, redundancy, and variability.

Preprocessing

Each speech utterance first undergoes silence removal to eliminate non-informative parts, usually found at the beginning and end of the recording. The resulting signal is then passed through a high-pass pre-emphasis filter with a coefficient of 0.97. This filter boosts higher frequencies to balance the natural drop in energy at higher frequencies in speech and improve the signal-to-noise ratio [165].

Framing and Windowing

The pre-emphasized speech signal is segmented into overlapping frames of 30 milliseconds, with a 10 millisecond shift between successive frames. This framing preserves the temporal dynamics of the signal while facilitating short-time spectral analysis. Each frame is multiplied by a Hamming window to minimize spectral leakage during the Fourier transformation and to enhance frequency resolution.

Feature Vector Computation

From each windowed frame, three types of feature vectors are extracted: MFCC, LPCC, and PLP. These methods provide different views of the speech spectrum, offering a rich and diverse set of features for emotion classification.

- ***MFCC (Mel-Frequency Cepstral Coefficients):***

MFCCs are computed by applying the Discrete Fourier Transform (DFT), followed by a Mel-scale filter bank that mimics human hearing. The filter outputs are logarithmically scaled and transformed using the Discrete Cosine Transform (DCT), resulting in a compact and decorrelated feature set. In this work, 12 MFCC coefficients were extracted together with the log-energy and their temporal derivatives, resulting in a 39-dimensional feature vector per frame. The MFCC extraction parameters were configured in HTK (Hidden Markov Model Toolkit) as shown in *Figure III-1* [166].

```
SOURCEFORMAT    = WAV
SOURCEKIND      = WAVEFORM
TARGETKIND      = MFCC_E_D_A
SOURCERATE      = 625           # 16 kHz sampling rate
WINDOWSIZE      = 300000.0      # 30 ms window
TARGETRATE      = 100000.0      # 10 ms frame shift
USEHAMMING      = T
PREEMCOEF       = 0.97
NUMCEPS         = 12
CEPLIFTER       = 22
```

Figure III-1: HTK file Configuration [166]

Each parameter in the configuration file contributes to the quality and representativeness of extracted speech features:

- **SOURCEFORMAT = WAV / SOURCEKIND = WAVEFORM** Indicates that the input data consists of raw audio waveforms stored in standard WAV format.
- **SOURCERATE = 625** Specifies the sampling period in 100-ns units. A value of 625 corresponds to a **16 kHz** sampling frequency, which is commonly used in SER to preserve the relevant spectral information.
- **WINDOWSIZE = 300000.0** Represents the analysis window length in 100-ns units. The configured value equals **30 ms**, offering an appropriate trade-off between temporal resolution and capturing the quasi-stationary speech characteristics.
- **TARGETRATE = 100000.0** Sets the frame shift to **10 ms**, ensuring frame-to-frame overlap and maintaining continuity in the speech signal representation.
- **USEHAMMING = T** Enables the Hamming windowing function, which minimizes spectral leakage at frame boundaries before applying the Fourier Transform.
- **PREEMCOEF = 0.97** Defines the pre-emphasis factor applied to the input waveform. This filter enhances high-frequency components to correct the natural spectral drop in speech.
- **NUMCEPS = 12** Determines the number of static MFCC coefficients retained after the DCT.

- **CEPLIFTER = 22** Applies cepstral liftering to improve the representation of cepstral coefficients by reducing the influence of very low and very high quefrency components, thus enhancing robustness to channel variations.
- **TARGETKIND = MFCC_E_D_A** Specifies the final feature vector format, which includes:
    - MFCC static coefficients (12)
    - **E**: Log-energy
    - **D**: Delta (first-order temporal derivatives)
    - **A**: Delta-Delta (second-order temporal derivatives)

This configuration is widely adopted in SER due to its ability to capture both spectral and dynamic information of speech signals.

- ***LPCC (Linear Predictive Cepstral Coefficients):***

The LPCC extraction process begins with autocorrelation followed by Linear Predictive Coding (LPC) to estimate the spectral envelope of the signal. The resulting LPC coefficients are then converted into cepstral coefficients.

In this work, LPCC features were extracted without including the energy term. Each feature vector consisted of 12 static coefficients, along with their delta and delta-delta derivatives, resulting in 36-dimensional vectors. The extraction was implemented using the HTK framework with appropriate configurations that exclude energy-related components.

- ***PLP (Perceptual Linear Prediction):***

PLP extraction incorporates auditory modeling, starting with critical-band analysis via Bark-scale filtering, followed by equal-loudness pre-emphasis and intensity-loudness compression. The resulting spectrum is modeled using autoregressive techniques to generate cepstral coefficients.

Like LPCC, the PLP features were computed without the energy component, resulting in 36-dimensional vectors (12 static + delta + delta-delta coefficients). These features provide a psychoacoustically motivated representation that complements the MFCC and LPCC features.

All features (MFCC, LPCC, and PLP) were extracted using the HTK toolkit. The MFCC configuration was detailed above, while the LPCC and PLP extractions followed similar setups,

differing mainly in the TARGETKIND parameter and the omission of the energy component. This unified approach ensured robustness, reproducibility, and consistency across feature types.

Figure III-2 illustrates the complete feature extraction process. It begins with the common analysis steps described earlier. In this example, a single utterance is transformed into a sequence of feature vectors using the HCopy command from the HTK toolkit. The same procedure is applied to all utterances, with different feature descriptors extracted in parallel namely, MFCC, LPCC, and PLP.

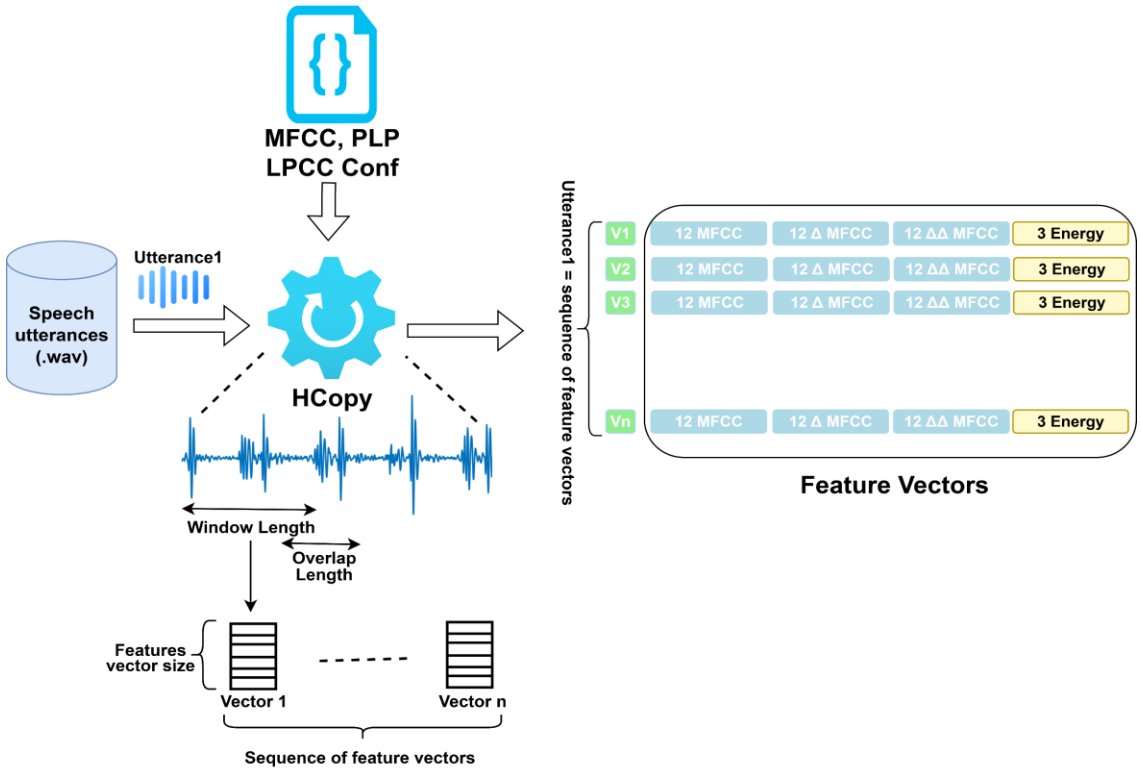


Figure III-2: HCopy Feature Extraction [166]

III.4. SER SYSTEM USING MI-BASED FEATURE SELECTION AND MACHINE LEARNING CLASSIFIERS COMBINED WITH VOTING RULES

III.4.1. Motivation and Objectives

Existing SER systems, such as those using GMM or Hidden Markov Models (HMM), often achieve high accuracy but are computationally expensive. These models require significant processing power and memory, which limits their deployment in real-time or embedded environments. This work aims to simplify the SER architecture by leveraging lightweight classifiers namely SVM, KNN, and GMM alongside a MI based feature selection

strategy and a voting rule method for decision level fusion. This combination targets three main objectives: (i) reducing the dimensionality of the input feature set, (ii) improving system efficiency and speed, and (iii) maintaining or improving recognition performance.

III.4.2. System Architecture

The architecture of the proposed SER system is organized around three central components: the use of machine learning classifiers, a decision mechanism based on voting rules, and the later integration of feature selection techniques.

The process begins with the representation of each utterance as a sequence of feature vectors extracted from the speech signal. These feature vectors, previously introduced in *Section III.3*, provide the acoustic information necessary for modeling emotional classes. In the training phase, they are used to build models with one of the selected classifiers, KNN, SVM, or GMM. In the testing phase, each feature vector of a new utterance is independently classified, producing a sequence of frame level predictions.

To move from frame level to utterance level recognition, the system applies a voting rule mechanism. Instead of relying on isolated predictions, all frame level outputs are aggregated, and the class that receives the majority of votes is assigned as the final emotion label of the utterance. This decision level fusion strategy enhances robustness by reducing the influence of noisy or ambiguous frames and yields more stable global predictions. It is particularly well suited to SER, since emotions are typically consistent throughout an utterance even if individual frames vary acoustically. In this way, the voting rule offers a computationally efficient yet effective alternative to more complex temporal models.

Although the baseline system relies on classifiers and voting rules, a third element, feature selection phase will be used to further enhance performance. By identifying the most informative features and reducing redundancy, feature selection improves both recognition accuracy and computational efficiency, thereby strengthening the system beyond its baseline configuration.

The overall flow of the proposed system is illustrated in *Figure III-3*, which highlights the path from feature vector representation through classification, voting rule fusion, and the later integration of feature selection.

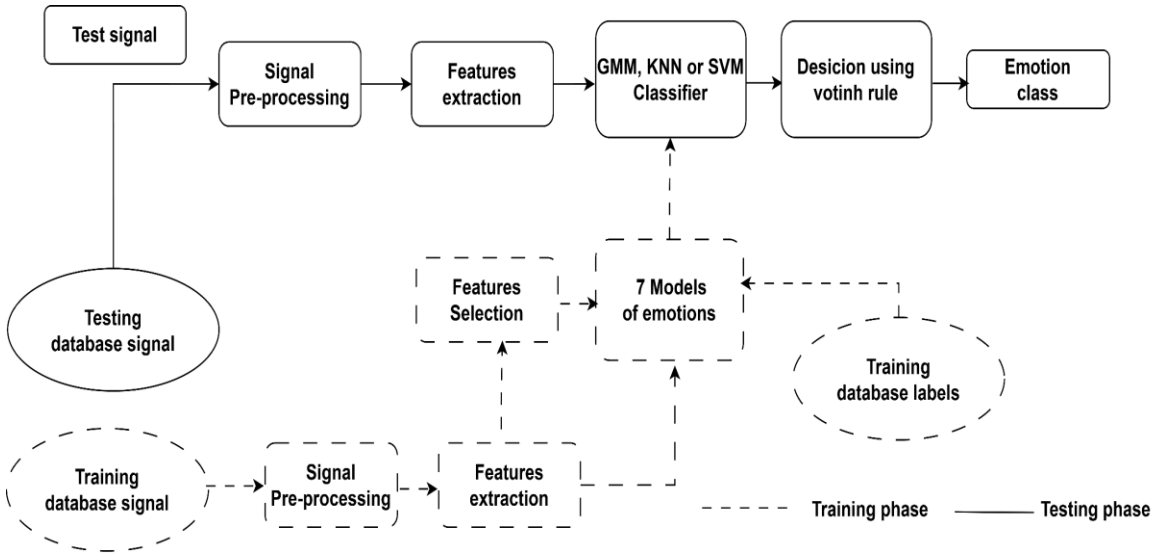


Figure III-3:Flowchart of the proposed SER system, showing feature vectors, classifiers, voting rule fusion

III.4.3. Feature Selection Strategy

The feature selection process aims to identify the most relevant and informative features from two distinct feature sets: a low-dimensional set composed of MFCC features, and a high-dimensional set obtained by combining MFCC, PLP, and LPCC features. These features are extracted for the SER task. To efficiently reduce dimensionality while maintaining relevant information, four mutual information-based feature selection methods are implemented: CIFE, JMI, mRMR, and ICAP.

Each of these strategies seeks to maximize the mutual information between the selected features and the target emotion classes, while minimizing redundancy among the features themselves. In practice, the estimation of mutual information relies on histograms constructed through discretization of the feature values, where the number of bins is determined using Sturges' formula.

Figure III-4 illustrates the overall process of the feature selection framework. The process begins with the input feature vectors, which are processed to select the best subset of features based on mutual information criteria. The selected features are then evaluated using classifiers to measure their performance. If the obtained performance does not meet the desired threshold, the process iteratively refines the selection until the optimal performance is reached. Once the best performance is achieved, the system outputs the optimal subset of features that balance relevance and redundancy most effectively.

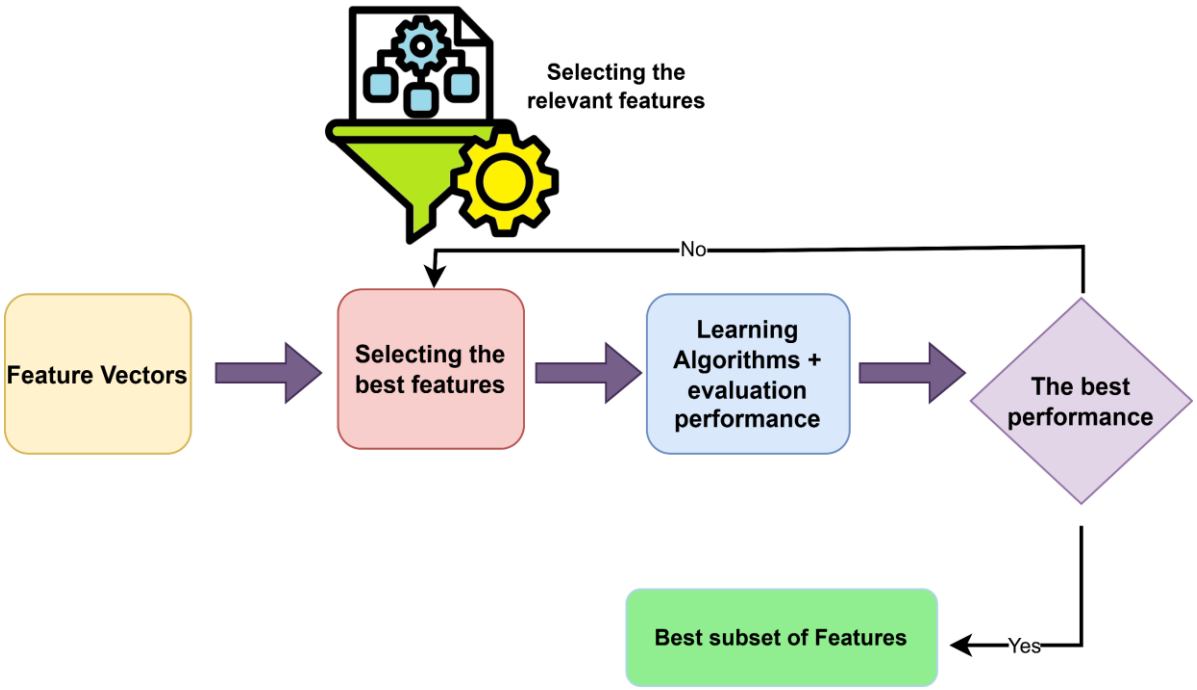


Figure III-4:Overview of the mutual information-based feature selection framework

To determine the optimal number of features, two stopping criteria are applied:

- SC1: The selection process stops as soon as the classification accuracy equals or exceeds the accuracy obtained using the full feature set. This criterion typically yields compact subsets, making it suitable for real-time or resource-constrained applications.
- SC2: The process continues until the highest possible accuracy is achieved, regardless of subset size.

Additionally, a tolerance factor $\alpha \in [0.01,0.05]$ is introduced to allow slight performance degradation compared to the full feature set. This tolerance defines an acceptable trade-off between performance and model complexity, offering a flexible compromise for practical deployment.

Through these selection methods and stopping criteria, the proposed feature selection strategy serves as a performance-enhancing and complexity-reducing component within the overall SER framework, ensuring both computational efficiency and high discriminative capability.

III.4.4. Results and Analysis

III.4.4.1. Classifier Performance with MFCC Features

The initial experiments evaluated the performance of three classifiers using MFCC features. For the KNN classifier, the Euclidean distance metric was used to measure similarity between feature vectors. The parameter k, representing the number of neighbors considered in classification, was varied from 1 to 10 to observe its impact on performance.

Table III-2 reports the accuracy obtained for each value of k. The best accuracy of 76.53% was achieved when k = 2, indicating that considering two nearest neighbors provides the most reliable classification with the MFCC feature set. Increasing k beyond 2 usually reduced accuracy, as more distant neighbors from other classes can confuse the classifier.

Table III-2:Accuracy of the SER system using the KNN classifier as a function of number of neighbors k

Value of K	1	2	3	4	5	6	7	8	9	10
Accuracy (%)	76.36	**76.53**	74.80	75.19	75.19	75.58	74.03	72.88	72.48	70.54

For the SVM classifier, experiments were conducted by varying the Box Constraint (BC) parameter while keeping the kernel scale set to *auto*. *Table III-3* shows the accuracy obtained for different BC values. The optimal performance of 83.72% was achieved for BC = 8, indicating that this value provides the best trade-off between margin maximization and misclassification penalty.

Table III-3:Accuracy of the SER system using the SVM classifier as a function of Box Constraint (BC) parameter

SVM BC	2	4	6	8	10	12	14	16	18
Accuracy (%)	82.17	83.33	83.33	**83.72**	83.33	82.55	82.17	81.78	82.17

For the GMM classifier, experiments were performed by varying the number of Gaussian components. *Table III-4* presents the results, showing that the highest accuracy of 85.27% was obtained with 14 components. This performance surpasses both KNN and SVM classifiers, establishing GMM as the most suitable baseline classifier for the proposed system.

Table III-4:Accuracy of the SER system using the GMM classifier as a function of GMM components

GMM components	2	4	6	8	10	12	14	16	18
Accuracy (%)	77.51	79.84	78.68	82.17	80.23	82.55	**85.27**	77.90	78.68

These results highlight the importance of tuning classifier-specific parameters. For KNN, k = 2 was optimal; for SVM, BC = 8 gave the best performance; and for GMM, 14 components yielded the highest accuracy.

III.4.4.2. Feature Selection with MFCC Features

To further enhance performance, feature selection strategies were applied to the GMM classifier using MFCC features. The results, presented in *Table III-5*, reveal that the JMI strategy significantly improved accuracy to 86.82% while reducing the feature set from 39 to 30 features. The ICAP strategy also achieved dimensionality reduction, maintaining the baseline accuracy of 85.27% with only 30 features. In contrast, CIFE and mRMR did not yield improvements beyond the performance of the full feature set.

Table III-5:Accuracy and Selected Feature Numbers Using CIFE, JMI, mRMR, and ICAP with MFCC Descriptors"

	SC 1		**SC 2**	
	Number of selected features	Accuracy (%)	Number of selected features	Accuracy (%)
CIFE	39	85.27	39	85.27
JMI	30	86.82	30	86.82
mRMR	39	85.27	39	85.27
ICAP	30	85.27	30	85.27

Figure III-5 presents the recognition accuracy as a function of the number of selected features. It can be observed that the performance of all strategies stabilizes when the number of selected features exceeds approximately 15. The JMI strategy demonstrates the most significant improvement, reaching a maximum of 86.82% at 30 features.

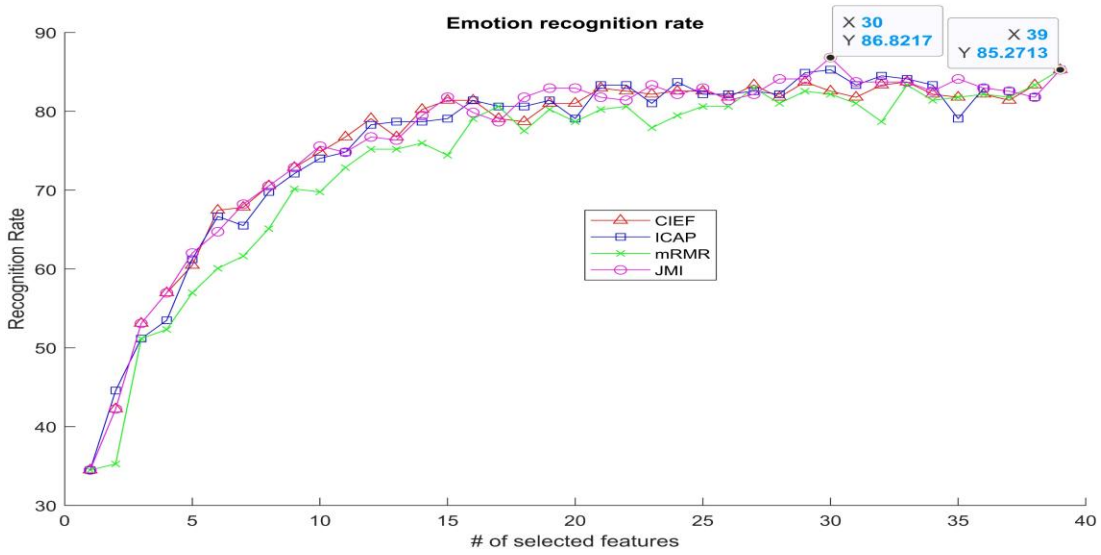


Figure III-5: Recognition accuracy as a function of the number of selected features MFCC

III.4.4.3. Feature Selection with High-Dimensional Features

To investigate the impact of higher-dimensional input, additional descriptors (PLP and LPCC) were combined with MFCC, resulting in a 111-dimensional feature space. The corresponding feature vector is represented as follows:

$$\text{Feature Vector} = \begin{matrix}[\text{MFCC } 1 \ldots, \text{MFCC } 12, \text{E}, \Delta \text{ MFCC } 1, \ldots, \Delta \text{ MFCC } 12, \Delta\text{E}, \\ \Delta\Delta \text{ MFCC } 1 \ldots, \Delta\Delta \text{ MFCC } 12, \Delta\Delta\text{E}, \\ \text{LPCC } 1 \ldots, \text{LPCC } 12, \Delta \text{ LPCC } 1, \ldots, \Delta \text{ LPCC } 12, \Delta\Delta \text{ LPCC } 1 \ldots, \\ \Delta\Delta \text{ PCC12 }, \\ \text{PLP1 } \ldots, \text{PLP12 }, \Delta \text{ PLP1 }, \ldots, \Delta \text{ PLP } 12, \Delta\Delta \text{ PLP } 1 \ldots, \Delta\Delta \text{ PLP12}]\end{matrix} \quad (III\text{-}1)$$

Here, the descriptors are concatenated in a structured order: MFCC features (including log-energy) first, followed by LPCC and then PLP features, with each type including static, delta (Δ), and delta-delta (ΔΔ) coefficients.

The performance of the GMM classifier with varying Gaussian components is reported in *Table III-6*. The best accuracy of 82.55% was achieved with 6 components, which is lower than the performance obtained with MFCC alone. This confirms the detrimental effect of the curse of dimensionality when using high-dimensional descriptors.

Table III-6:Accuracy of the SER system using the GMM classifier as a function of GMM components with 111 features

GMM components	2	4	6	8	10	12	14	16	18
Accuracy (%)	78.29	82.17	82.55	74.80	73.64	72.48	74.03	73.64	73.25

To mitigate the curse of dimensionality, feature selection strategies were applied. The results are summarized in *Table III-7*.

- According to SC1 (favoring the smallest number of features while maintaining good performance), ICAP is the best strategy, selecting 42 features and achieving an accuracy of 82.94%. mRMR follows closely with 48 features and an accuracy of 82.55%, while JMI and CIFE show limited improvement.
- According to SC2 (favoring the highest accuracy even if more features are selected), ICAP again provides the best result with 95 features and an accuracy of 84.49%, followed by mRMR with 104 features achieving 84.10%. JMI slightly improves to 82.94%, whereas CIFE remains at 82.55%.

Table III-7:Accuracy and number of relevant features using CIFE, JMI, mRMR, and ICAP strategies with high-dimensional vectors

	SC1		**SC2**	
	No. of selected features	Accuracy (%)	No. of selected features	Accuracy (%)
CIFE	111	82.55	111	82.55
JMI	68	82.55	88	82.94
mRMR	48	82.55	104	84.10
ICAP	42	82.94	95	84.49

The experimental results indicate that simply concatenating MFCC, LPCC, and PLP descriptors into a high-dimensional vector is not sufficient to improve recognition accuracy; instead, it can degrade performance due to increased redundancy and the curse of dimensionality.

Feature selection significantly mitigates this effect:

- ICAP emerges as the most effective strategy, balancing a reduced number of features with high classification accuracy.
- mRMR is also effective, achieving slightly lower accuracy than ICAP while still providing a significant improvement compared to using the full high-dimensional feature set without selection.
- Strategies like JMI provide only minor improvements, whereas CIFE offers no meaningful gain.

Figure III-6 provides a graphical representation of accuracy as a function of the number of selected features. The results indicate that all strategies converge to stable performance levels once approximately 40 features are retained, suggesting that a relatively small subset of features is sufficient to achieve competitive accuracy while reducing computational costs.

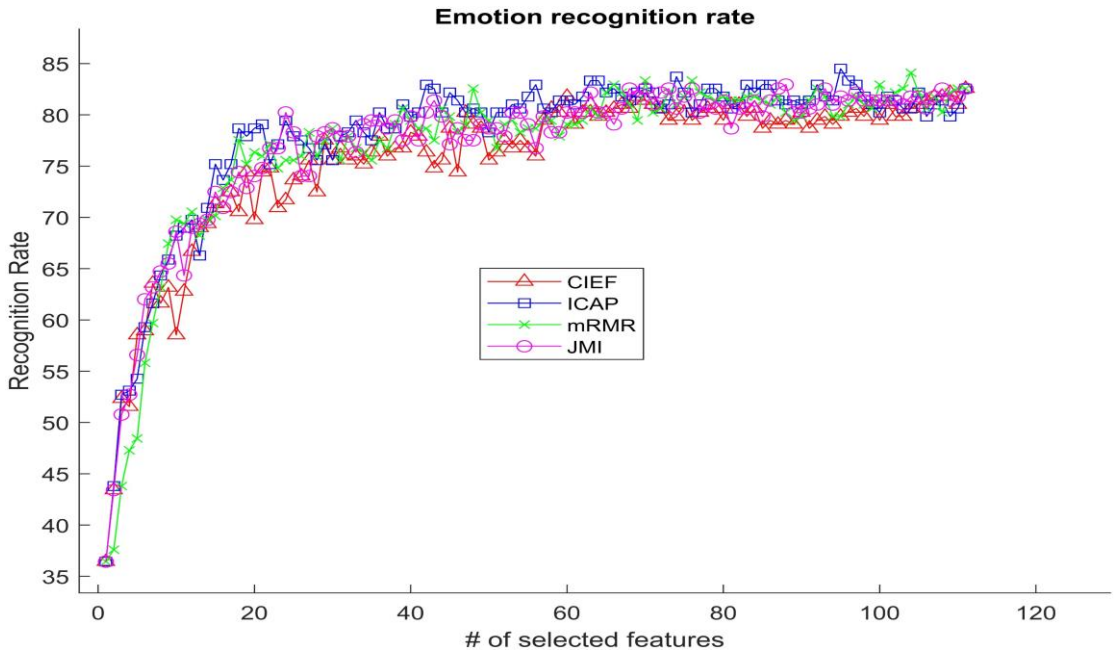


Figure III-6:Variation in accuracy with the number of selected features for CIFE, JMI, mRMR, and ICAP using the GMM classifier.

III.5. SER SYSTEM USING MI-BASED FEATURE SELECTION AND A 1D-CNN CLASSIFIER

III.5.1. Motivation and Challenges

1D-CNNs have proven to be effective for sequential data such as speech. However, as the input dimensionality increases, the complexity of the CNN model rises sharply. This not only increases memory consumption but also leads to overfitting and longer training times. To

address this, we propose integrating a feature selection stage into the CNN model. The objective is to select an optimal subset of features using mutual information, while introducing a stopping criterion to regulate the number of features chosen during the greedy forward selection process.

III.5.2. 1D-CNN Architecture Description

The one-dimensional convolutional neural network proposed in this work is designed to classify emotional states from speech signals by analyzing sequential acoustic features. This architecture was selected for its ability to efficiently model temporal dependencies within one-dimensional input vectors, such as MFCC, LPCC, and PLP features, without requiring large computational resources.

The network consists of two Conv1D layers, each followed by a ReLU activation function and a normalization step to enhance training stability and reduce internal covariate shift. The first convolutional layer uses 260 filters when processing the 39-dimensional MFCC input, and 150 filters when using the extended 111-dimensional feature set that combines MFCC, LPCC, and PLP. This layer applies a kernel size of 3 with stride 1, enabling it to capture short-term dependencies across neighboring frames. The second convolutional layer doubles the number of filters used in the first layer, namely 520 filters for the MFCC configuration and 300 filters for the 111-dimensional input, allowing the model to extract richer and more abstract representations in deeper layers.

After the convolutional blocks, the output is passed through an average pooling layer with a pool size of 2, reducing the temporal dimension of the feature maps and helping to prevent overfitting. The pooled feature maps are flattened and passed to a fully connected layer, followed by a softmax output layer that produces the probability distribution over the target emotion classes. The model is trained using the categorical cross-entropy loss function and optimized with the Adam optimizer. The architecture of the proposed 1D-CNN model is illustrated in *Figure III-7*. It shows the sequential flow from the input acoustic features through the convolutional and pooling layers to the final classification stage.

Two configurations were evaluated during experimentation. The first uses only 39 MFCC features, while the second incorporates a richer, high-dimensional representation with 111 features by combining MFCC, LPCC, and PLP. The 1D-CNN model achieves strong performance in both settings, with 91.09% accuracy using MFCC and 91.47% with the full feature set. Despite the increase in input size, the model remains compact and efficient due to

the limited number of layers and the integration of average pooling and feature selection techniques.

This 1D-CNN architecture strikes a balance between depth and efficiency. It is particularly well suited to speech-based emotion recognition tasks, and when combined with the mutual information-based feature selection and stopping criterion proposed in the next section, it supports scalable and high-performance emotion classification even in high-dimensional settings.

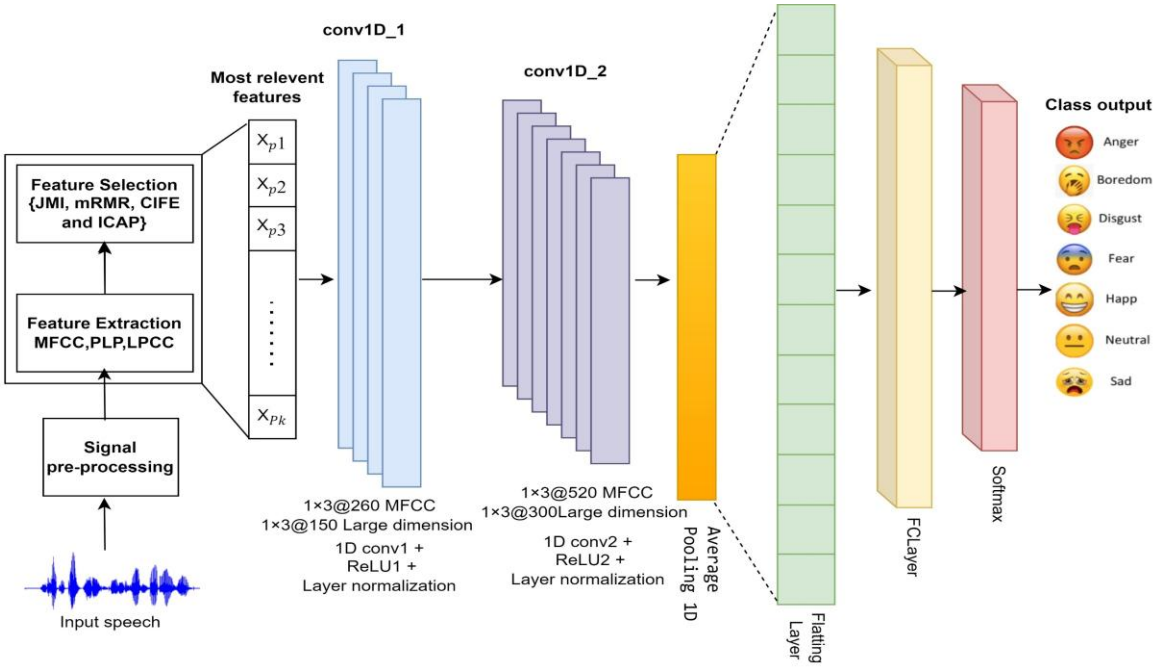


Figure III-7: Proposed SER System

III.5.3. Performance with MFCC Features

In the first experimental configuration, the SER system was trained and evaluated using feature vectors consisting of 39 MFCC coefficients. To ensure reliable performance, a grid search was carried out to identify the most effective combination of hyperparameters for the 1D-CNN model. The parameters investigated included the number of convolutional layers, the number of filters per layer, kernel size, batch size, learning rate, number of epochs, and the optimization algorithm. The corresponding ranges and the selected values are presented in *Table III-8*.

To mitigate overfitting, the network architecture was deliberately kept simple, consisting of two convolutional layers followed by a global average pooling layer. This reduced the number of trainable parameters while preserving the network's ability to learn

discriminative patterns. Additionally, normalization layers were applied to stabilize training and improve generalization.

With this setup, the system achieved an accuracy of 91.09% on the EmoDB database. This result confirms that the 1D-CNN architecture is capable of effectively classifying emotional speech when trained with MFCC descriptors.

Table III-8:Hyperparameters and performance of the 1D-CNN classifier with MFCC features

Parameter Type	**Range of Values**	**Selected Value**
Number of 1D Convolution Layers	2−4	2
Number of filters	8−512	260
Size of kernel	1×1−7×1	3×1
Batch size	8−32	22
Learning rate	0.0005−0.01	0.001
Number of epochs	50−500	190
Optimization Algorithm	Adam - SGDM	Adam

III.5.4. Performance with High-Dimensional Feature Vectors

The second configuration extended the input to a high-dimensional feature vector of 111 components by combining MFCC, LPCC, and PLP descriptors. The rationale was to investigate whether complementary features could provide additional discriminative information beyond MFCC alone. The hyperparameters selected for this configuration are provided in *Table III-9*.

The maximum accuracy obtained with this high-dimensional input was 91.47%, which is only a marginal improvement compared to the MFCC-only case. However, this performance gain came at the cost of increased complexity. The system required more memory, longer training time, and became more susceptible to overfitting due to redundancy among features. This observation illustrates the curse of dimensionality, where increasing the feature space does not necessarily lead to better performance but can instead reduce the system's efficiency.

Table III-9:Hyperparameters and performance of the 1D-CNN classifier with 111-dimensional feature vectors

Parameter Type	**Range of Values**	**Selected Value**
Number of 1D Convolution Layers	2−4	2
Number of filters	8−512	150
Size of kernel	1×1−7×1	3×1
Batch size	8−32	22
Learning rate	0.0005−0.01	0.001
Number of epochs	50−500	160
Optimization Algorithm	Adam - SGDM	Adam

III.5.5. Feature Selection with High-Dimensional Vectors

To address the limitations of high-dimensional input, feature selection was applied using four MI-based strategies: CIFE, JMI, mRMR, and ICAP. These methods are applied to the 111-dimensional combined feature vector described in *Equation (III-1)*, aiming to retain only the most informative parameters while reducing redundancy and improving computational efficiency.

The results showed that MFCC features consistently ranked among the most relevant, reinforcing their importance in SER. Energy related features and certain LPCC and PLP descriptors were also selected, indicating that complementary features can contribute useful information. The top ten selected features for each strategy, along with the corresponding recognition accuracies, are summarized in *Table III-10.*

Table III-10:Top ten selected features and accuracies using CIFE, JMI, mRMR, and ICAP (NF: number of selected features; Acc (%): accuracy)

Features	CIFE		JMI		mRMR		ICAP	
	NF	Acc%	NF	Acc%	NF	Acc%	NF	Acc%
X1	2	51.55	2	51.55	2	51.55	2	51.55
X2	13	58.52	13	58.52	5	53.10	1	61.62
X3	76	67.05	1	67.05	87	67.82	5	60.46
X4	40	70.54	40	62.01	86	64.34	79	63.95
X5	1	70.93	77	68.99	25	59.30	87	69.37
X6	5	72.86	76	65.11	9	70.15	86	75.19

X7	4	75.58	5	69.76	4	76.35	3	82.55
X8	41	76.35	4	76.74	1	77.13	9	83.72
X9	9	76.74	79	72.48	3	82.17	40	82.17
X10	78	81.78	9	77.13	10	81.78	12	79.84

To further illustrate the impact of feature selection, *Figure III-8* presents recognition accuracy as a function of the number of selected features. All methods reached a plateau around 25 features, confirming that a relatively small subset is sufficient to achieve near-optimal performance. This result demonstrates the importance of feature selection as a means of balancing accuracy and computational efficiency.

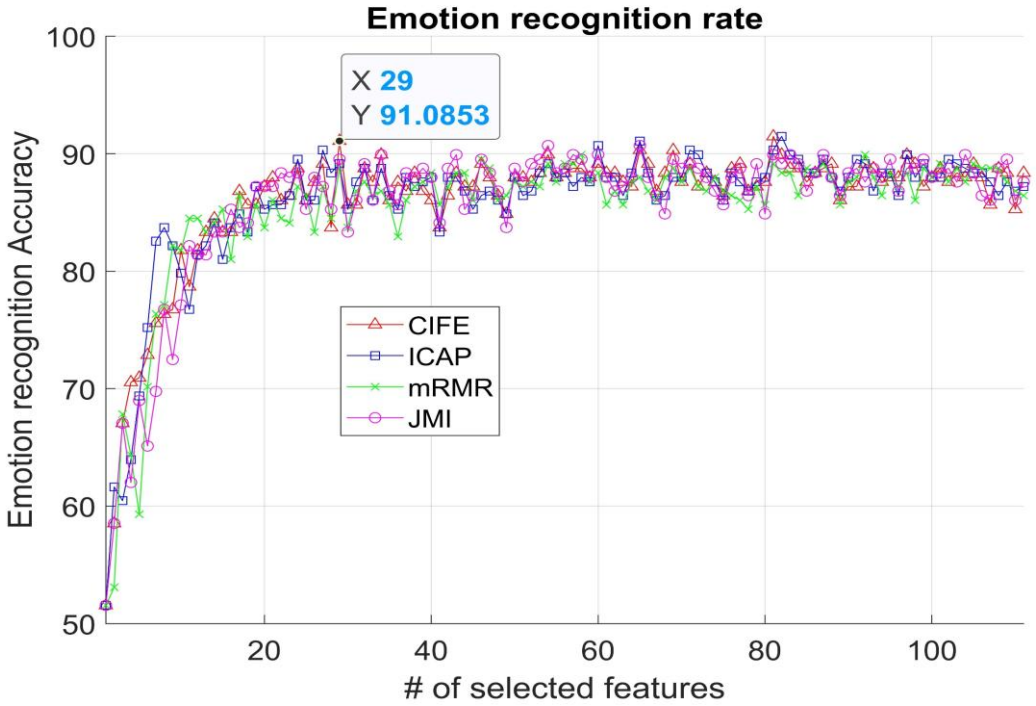


Figure III-8:Accuracy versus number of selected features using MI-based strategies with the 1D-CNN

III.5.6. Proposed Stopping Criterion for feature selection

In the traditional greedy forward selection algorithm, as presented in *Chapter II.7*, the stopping condition is generally determined by a predefined number of features k. However, this approach does not take into account the classification performance or the relevance of the selected subset in relation to the target classes. As a result, it may lead either to the selection of an excessively large set of features that includes redundant or irrelevant information, or to a feature set that is too limited, thereby omitting critical discriminatory features. To overcome

this limitation, we propose an improved stopping criterion that integrates classification accuracy into the feature selection process.

The proposed criterion introduces a tolerance factor $\boldsymbol{\alpha} \in [\mathbf{0.01}, \mathbf{0.05}]$, which defines the allowable degradation in performance relative to the accuracy achieved using the full feature set, denoted as $\boldsymbol{Acc_{All}}$.

The selection process is terminated once the subset of selected features yields an accuracy satisfying:

$$\boldsymbol{Accuracy(S) \geq (1-\alpha) Accuracy(F_{all})}$$

This threshold provides a flexible trade-off between accuracy and dimensionality. A small value of α keeps high accuracy, while a larger value allows selecting fewer features, even if accuracy drops slightly. This way, the system can automatically choose a good number of features without needing to fix k in advance [35].

The proposed stopping criterion integrated into the greedy forward selection algorithm is given as follows:

Algorithm III-1:Greedy Forward Selection with Stopping Criterion

---

Input:

Features set $\boldsymbol{F = \{X_1, X_2, \dots, X_n\}}$

Deterioration factor α

Output:

Subset of selected features $\boldsymbol{S = \{X_{p1}, X_{p2}, \dots, X_{pm}\}}$, $\boldsymbol{with\ m < n}$

Algorithm Steps:

Initialization:

Initialize F with n features

Set S as an empty subset

Calculation of Mutual Information (MI):

For each feature X_i *in F, calculate* $MI(C;X_i)$

Select the first feature:

Identify X_{p1} *maximizing* $MI(C;X_i)$

Remove X_{p1} *from F and add it to S*

(Greedy selection) continue iterating from $j = 2$ ***until*** $\boldsymbol{Accuracy(S) \geq (1-\alpha)Accuracy(F_{all})}$***:***

while $\mathbf{Accuracy(S) < (1-\alpha)\ Accuracy\ (F_{all})}$ ***do***

for *each feature* X_i *in* F ***do***

Calculate $MI(C;S,X_i)$

end for

Select X_{pj} *from* F *maximizing* $MI(C;S,X_i)$ *at step* j

Remove X_{pj} *from* F *and add it to* S

end while

Output:

Return subset S *verifying*

$\mathbf{Accuracy(S) \geq (1-\alpha)\ Accuracy\ (F)}$

III.5.7. Performance Evaluation of the Proposed Stopping Criterion

In this subsection, the results of applying the proposed stopping criterion are presented. The purpose of this criterion is to determine the smallest subset of features that sufficiently represents the target classes while avoiding the need to process the full high-dimensional feature space. By integrating the stopping rule into the feature selection process, the algorithm identifies a reduced feature set that balances recognition accuracy and computational efficiency. The deterioration factor α, ranging from 0.01 to 0.05, was used to control the acceptable reduction in accuracy relative to the baseline accuracy (Acc_All=91.47%) obtained with all 111 features.

The outcomes of these experiments are summarized in *Table III-11*, which reports the number of selected features (#SF), classification accuracy (Acc%), runtime, memory usage, and number of trainable parameters for each value of α. This comprehensive evaluation provides insight into the trade-off between performance and computational requirements achieved by each feature selection strategy when combined with the proposed stopping criterion.

Table III-11:Performance evaluation of the proposed stopping criterion across CIFE, JMI, mRMR, and ICAP strategies

Method	**α**	**#SF**	**Acc %**	**Runtime (s)**	**TP k (thousands)**	**Memory (MB)**
CIFE	0.01	29	91.08	65.93	151.5k	220.96
	0.02	29	91.08	65.93	151.5k	220.96
	0.03	24	88.75	60.20	149.2k	195.50
	0.04	21	87.98	58.20	147.9k	185.00
	0.05	20	87.20	57.80	147.4k	179.80
JMI	0.01	54	90.69	66.80	162.7k	234.45
	0.02	34	89.92	63.00	153.7k	215.10
	0.03	29	89.53	65.93	151.5k	220.96
	0.04	22	88.37	59.00	148.3k	189.70
	0.05	19	87.20	57.20	147k	171.60
mRMR	0.01	92	89.93	77.08	179.8k	409.69
	0.02	58	89.92	68.00	164.5k	241.18
	0.03	29	88.75	65.93	151.5k	220.96
	0.04	29	88.75	65.93	151.5k	220.96
	0.05	24	87.20	60.20	149.2k	195.50
ICAP	0.01	65	91.08	68.61	167.7k	258.38
	0.02	67	90.31	72.83	168.6k	259.32
	0.03	24	89.53	60.20	149.2k	195.50
	0.04	24	89.53	60.20	149.2k	195.50
	0.05	19	87.20	57.20	147k	171.60

The results reveal several key observations. When $\alpha = 0.01$ or α=0.02, the CIFE method selected only 29 features while maintaining an accuracy of 91.08%, which is nearly equivalent to the baseline performance. This reduction corresponds to a 73.87% decrease in dimensionality (eliminating 82 features) and was accompanied by significant computational gains. Specifically, runtime decreased to 65.93 seconds, memory usage dropped to 220.96 MB, and the number of trainable parameters was reduced to 151.5k. These results highlight the efficiency of CIFE at low deterioration levels.

When the deterioration factor was increased to α=0.03 and α=0.04, the ICAP strategy demonstrated favorable results. At these levels, ICAP selected 24 features, achieving an accuracy of 89.53% while reducing runtime to 60.20 seconds, the number of trainable

parameters to 149.2k, and memory usage to 195.50 MB. These reductions illustrate that ICAP provides a more balanced trade-off between accuracy and efficiency compared to mRMR and JMI under similar conditions.

At the highest deterioration factor α=0.05, all strategies converged to an accuracy of 87.20%. In this case, JMI and ICAP produced the most compact representations, each selecting 19 features, a reduction of 82.88% relative to the original 111 features. Among them, ICAP achieved the best computational efficiency, with runtime reduced to 57.20 seconds, 147k trainable parameters, and 171.60 MB of memory usage. Although accuracy was reduced at this level, the system still retained acceptable recognition performance while achieving substantial efficiency gains.

Taken together, these results demonstrate the effectiveness of the proposed stopping criterion in reducing the dimensionality of the feature space while maintaining a controlled balance between accuracy and efficiency. By avoiding arbitrary choices of the number of selected features, this approach ensures that feature selection remains adaptive and performance-driven, thereby supporting the development of SER systems suitable for real-time and resource-constrained environments.

III.6. CONCLUSION

In this chapter, we investigated the application of feature selection methods for SER task, aiming to enhance recognition performance while reducing system complexity. Two main approaches were explored: a machine learning based framework combining lightweight classifiers (KNN, SVM, GMM) with MI-based feature selection and voting rules, and a deep learning approach integrating MI-based feature selection into a 1D-CNN architecture with a proposed stopping criterion.

Experiments conducted on the EmoDB dataset demonstrated several important findings. First, among the baseline classifiers, GMM achieved the highest accuracy, highlighting its suitability for frame level emotion modeling. Applying MI-based feature selection consistently improved or maintained performance while reducing the number of features. When combining MFCC, LPCC, and PLP into a high-dimensional 111-feature vector, ICAP strategy reduces the feature vector size by 62.2% while improving the recognition accuracy to 82.94%, demonstrating that feature selection can effectively mitigate the curse of dimensionality in high-dimensional spaces by enhancing efficiency without compromising performance.

The integration of MI-based feature selection into the 1D-CNN further demonstrated the effectiveness of dimensionality reduction in deep learning contexts. The model achieved high accuracy with both MFCC only (91.09%) and high-dimensional input (91.47%), while the proposed stopping criterion allowed adaptive selection of a minimal subset of features. This mechanism provided a flexible trade-off between accuracy and computational efficiency: in the maximum reduction scenario, ICAP selected only 19 features (an 82.88% reduction) while maintaining acceptable performance and significantly reducing memory usage, runtime, and the number of trainable parameters.

These findings underscore the pivotal role of feature selection in SER, providing an efficient, high-performance framework suitable for both traditional and deep learning approaches.

IV

APPLICATION OF FEATURE SELECTION FOR FACIAL EXPRESSION RECOGNITION

IV.1. INTRODUCTION

FER has developed rapidly in recent years. Earlier FER systems mainly relied on handcrafted features, such as LBP, HOG, BSIF, and DWT, to describe facial textures and structures. With the rise of deep learning, especially CNNs, it became possible to automatically extract highly discriminative features directly from facial images, reducing the need for manual feature design.

However, both handcrafted and deep features often produce high-dimensional feature vectors, which increase computational cost and may affect classification performance. To address this issue, this chapter applies MI-based feature selection to reduce redundancy and retain only the most informative features for emotion classification.

Two approaches are presented in this work:

1. **Handcrafted feature-based approach:** LBP, LPQ, BSIF, and HOG features are extracted and combined with MI-based feature selection, then classified using a 1D-CNN.
2. **Deep feature-based approach:** Features are extracted from AlexNet, a pre-trained CNN model, combined with MI-based feature selection, and classified using a KNN classifier.

Using MI-based feature selection with the two stopping criteria, SC1 and SC2, the most informative and optimal features are selected, ensuring a balance between dimensionality reduction, recognition performance, and processing time.

The results presented in this chapter demonstrate that MI-based feature selection, when applied to both handcrafted and deep features under SC1 and SC2, improves recognition performance and reduces processing time, highlighting its effectiveness in developing efficient and accurate FER systems.

IV.2. DATASET DESCRIPTION FOR FER

We tested the proposed system using two benchmark datasets for FER: the MUG dataset and the CK+ dataset. These datasets include a wide variety of facial image scales and intricate variations. Each dataset was divided into training and testing sets, with 70% allocated for training and 30% for testing the model's accuracy, following the commonly used split ratio.

IV.2.1. MUG Dataset

The MUG [167] dataset comprises 86 participants (51 males and 35 females) aged between 20 and 35 years. However, only 52 participants are available to researchers via the internet. The images in the dataset were captured using a camera and two 300 W light sources, with each participant seated on a chair against a blue background. Images were recorded at a rate of 19 frames per second, with a resolution of 896 × 896 pixels in JPEG format. Each participant exhibited seven distinct facial expressions, with each emotion captured across several image sequences (typically three to five), consisting of 50–160 images per sequence.

IV.2.2. CK+ Dataset

The dataset CK+ [168] is one of the most widely used in FER systems. The dataset comprises 981 images extracted from 327 video sequences involving 118 participants. Each sequence includes 10 to 60 frames, capturing the transition from a neutral to an extreme facial expression. The images are categorized into seven basic expressions with a resolution of 48 × 48 pixels, as detailed in *Table IV*-1. In this table, we present the number for each facial expression across the two datasets used in this study to evaluate the proposed FER system.

Table IV-1:Number of images for each expression in the two datasets. The emotions are denoted by the following abbreviations: Anger (AN), Disgust (DI), Fear (FE), Happiness (HA), Sadness (SA), Surprise (SU), Neutral (NE), and Contempt (CO)

Dataset	**AN**	**DI**	**FE**	**HA**	**NE**	**CO**	**SA**	**SU**	∑
CK +	135	177	75	207	/	54	84	249	981
MUG	260	255	240	260	260	/	245	260	1780

IV.3. FER SYSTEM BASED ON TRADITIONAL FEATURES AND MI-BASED FEATURE SELECTION

IV.3.1. Feature Extraction and Preprocessing

Prior to the feature extraction process, all facial images within the datasets undergo a pre-processing phase, during which the facial region is localized using the Haar cascade classifier proposed by Viola and Jones [169]. Following this, feature vectors are extracted by applying four texture-based descriptors including LBP, LPQ, BSIF, and HOG to the entire facial region. Each resulting feature map is subsequently partitioned into four non-overlapping blocks, and a 256-bin histogram is computed for each block. The histograms obtained from these blocks are then concatenated to form the final feature vector [170]. During experimentation, the parameters of each descriptor were varied within predefined ranges, and only the best-performing parameter values are reported below.

The specific parameter ranges explored and the selected optimal values for each descriptor are reported as follows:

- **LBP descriptor:**

The number of neighboring points was fixed at $P = 8$, while the radius R was varied in the range $1 \leq R \leq 8$. The best performance was obtained at $R = 5$.

- **LPQ descriptor:**

The local window size R was varied in the range $\{3, 5, 7\}$. The optimal results were achieved with a window size of $R = 5$.

- **BSIF descriptor:**

Square filter sizes were varied from 5×5 to 11×11, and the number of bits was varied in the range $\{5,6,7,8\}$. The best configuration employed 7×7patches encoded using 8 bits.

- **HOG descriptor:**

The cell size was varied in the range $\{4 \times 4, 5 \times 5, 8 \times 8\}$, and the number of orientation bins was varied from 6 to 9. The optimal configuration used a cell size of $5 \times 5$with 9 orientation bins.

Table IV-2 presents the dimensionality of the resulting feature vectors for each descriptor using the selected optimal parameters.

Table IV-2:Number of features for each descriptor

Descriptor	Number of blocks per image	Number of features
LBP	4 Blocks of size 50 x 50	4 x 256 =1024
LPQ		4 x 256 =1024
BSIF		4 x 256 =1024
HOG		4 x 9(Bins) x (5 x 5) (Celles) = 900

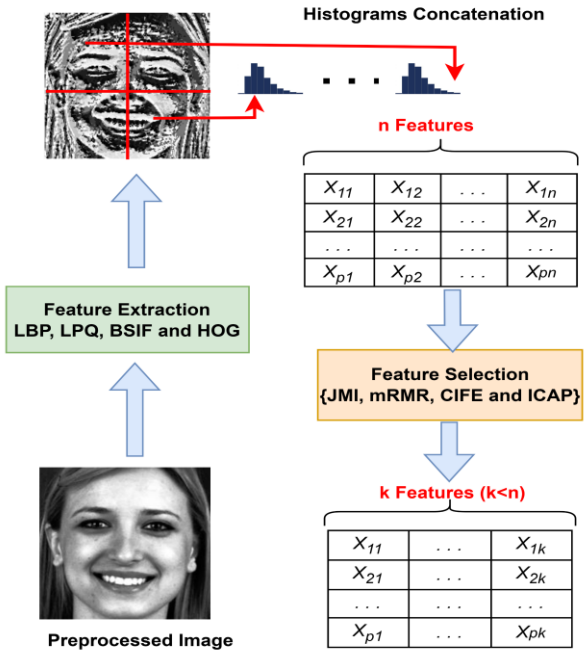


Figure IV-1:Flowchart of Feature Extraction and Selection Process

The facial feature descriptors discussed earlier produce high-dimensional feature vectors. Hence, applying a feature selection technique is essential to retain only the most informative and discriminative features. Dimensionality reduction not only improves computational efficiency and reduces memory requirements but can also enhance recognition accuracy by mitigating the curse of dimensionality. To achieve this objective, several MI–based feature selection methods namely, mRMR, JMI, CIFE and ICAP are used. *Figure IV-1* illustrates the feature extraction and selection process described previously.

IV.3.2. Performance with 1D-CNN Classifier

The architecture of the proposed one-dimensional CNN is illustrated in *Figure IV-2*. The training stage focuses on learning discriminative patterns of facial expressions, while the testing stage evaluates the overall recognition performance of the FER system. Both stages rely on handcrafted feature descriptors, including LBP, LPQ, BSIF, and HOG. These descriptors are subsequently refined through a feature selection process to reduce dimensionality and retain the most informative features.

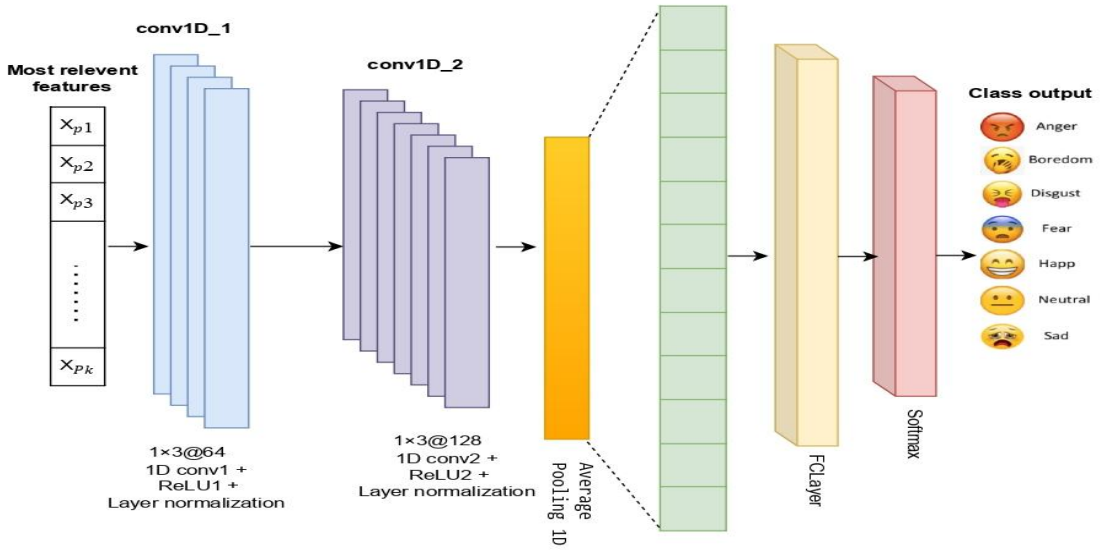


Figure IV-2: Proposed FER System

The designed network consists of two Conv1D layers. The first convolutional layer employs 64 filters, each followed by a Rectified Linear Unit (ReLU) activation function to introduce non-linearity and mitigate the vanishing gradient problem. Normalization layers are also incorporated to ensure consistent data distribution across network layers. The second convolutional layer doubles the number of kernels relative to the first one, further enhancing the network's representational capacity. Both convolutional layers use a kernel size of 3×1 with a stride of 1. Increasing the number of filters in deeper layers contributes to higher recognition accuracy, as reported in [43].

A Randomized Search strategy was employed to identify the optimal set of hyperparameters for the 1D-CNN in combination with the four descriptors. The key

hyperparameters, their search ranges, and the final selected values are summarized in *Table IV-3*.

Table IV-3:Hyperparameters of the proposed 1D-CNN classifier

Parameter Type	**Range of Values**	**Selected Value**
Number of 1D Convolution Layers	2−4	2
Number of filters	8−512	64
Size of kernel	1×1−7×1	3×1
Batch size	8−64	32
Learning rate	0.0005−0.01	0.001
Number of epochs	64−512	128
Optimization Algorithm	Adam - SGDM	Adam

IV.3.3. Optimal Descriptor for FER

This section analyzes the performance of the proposed FER system when combined with different feature descriptors. The evaluation is based on two main criteria: Recognition Rate (RR) and Processing time for full features (PTF) during the classification phase. It is important to note that PTF only refers to the classification stage using 1D-CNN and does not include the feature extraction process.

The recognition rates and processing times obtained with each descriptor on the MUG and CK+ datasets are presented in *Table IV-4 and 5*, respectively.

Table IV-4:Recognition Rate (RR), Processing Time for Full Features (PTF), and Total Number of Features (TNF) for Various Descriptors on the MUG Dataset

Descriptor	**RR (%)**	**PTF (s)**	**TNF**
LBP	99.06	13.69	1024
LPQ	99.81	13.72	1024
BSIF	96.63	13.39	1024
HOG	99.43	13.35	900

Table IV-5:RR, PTF and TNF for Various Feature Descriptors on CK+ Dataset

Descriptor	RR (%)	PTF (s)	TNF
LBP	99.33	9.06	1024
LPQ	98.66	8.91	1024
BSIF	98.99	9.024	1024
HOG	98.99	8.56	900

The results in *Table IV-4* demonstrate that the BSIF descriptor consistently yields the lowest RR (96.63%) across all descriptors on the MUG dataset. Conversely, LPQ achieves the highest RR (99.81%), but with the longest processing time (13.72 s). The HOG descriptor provides a competitive RR (99.43%) with the lowest processing time (13.35 s), making it particularly suitable for real-time applications where computational efficiency is critical.

On the CK+ dataset (*Table IV-5*), LBP achieves the highest RR (99.33%), although it also requires the longest processing time (9.06 s). HOG, while slightly less accurate (98.99%), outperforms other descriptors in terms of efficiency, achieving the shortest processing time (8.56 s).

From these findings, it can be concluded that HOG represents a balanced trade-off between accuracy and computational efficiency, making it the most versatile descriptor for FER applications across multiple datasets. Therefore, in the subsequent analysis of feature selection strategies, HOG is chosen as the primary descriptor to evaluate the influence of dimensionality reduction on both recognition performance and computational efficiency.

IV.3.4. Impact of Feature Selection

IV.3.4.1. Influence on Classification Performance

Table IV-6 and *7* summarize the effect of different feature selection strategies including CIFE, ICAP, JMI, and mRMR on the RR and the number of selected features for four handcrafted descriptors (HOG, LPQ, LBP, and BSIF) on the CK+ and MUG datasets. Two stopping criteria (SC1 and SC2) are used to determine the optimal number of features, allowing an assessment of the trade-off between compact feature subsets and classification performance.

The Tables present the results of feature selection strategies applied to different feature extraction descriptors (HOG, LPQ, LBP and BSIF) and feature selection strategies (CIFE, ICAP, JMI and mRMR) for the CK+ and MUG datasets.

Table IV-6:Evaluation of Feature Selection strategies: Number of Selected Feature and RR for HOG, LPQ, LBP, and BSIF Descriptors on CK+ Dataset using SC1 and SC2

		SC1		SC2	
		No. of selected features	**RR (%)**	**No. of selected features**	**RR (%)**
HOG	CIFE	144	98.99	361	100
	ICAP	304	98.99	692	100
	JMI	170	98.99	361	100
	mRMR	304	98.99	729	100
LPQ	CIFE	361	98.66	679	100
	ICAP	207	98.66	561	100
	JMI	384	98.66	899	100
	mRMR	283	98.66	497	100
LBP	CIFE	379	99.33	1017	100
	ICAP	435	99.33	868	100
	JMI	281	99.33	1007	100
	mRMR	441	99.33	711	99.66
BSIF	CIFE	937	98.99	937	99.33
	ICAP	724	98.99	724	99.33
	JMI	841	98.99	841	98.99
	mRMR	531	98.99	531	98.99

Table IV-7:Evaluation of Feature Selection strategies: Number of Selected Feature and RR for HOG, LPQ, LBP, and BSIF Descriptors on MUG Dataset Using SC1 and SC2

		SC1		SC2	
		No. of selected features	**RR (%)**	**No. of selected features**	**RR (%)**
HOG	CIFE	208	99.62	367	100
	ICAP	94	99.43	278	100
	JMI	114	99.43	331	100
	mRMR	120	99.43	376	100
LPQ	CIFE	506	99.81	633	100
	ICAP	465	100	465	100

	JMI	296	99.81	662	100
	mRMR	452	99.81	708	100
LBP	CIFE	478	99.25	720	99.81
	ICAP	258	99.06	586	100
	JMI	112	99.06	359	99.81
	mRMR	205	99.43	481	100
BSIF	CIFE	643	96.63	975	98.31
	ICAP	392	97.19	941	98.87
	JMI	450	96.82	871	98.50
	mRMR	462	96.82	933	98.87

The results presented in *Tables IV-6 and 7* reveal that the RR generally improves or reaches 100% as the stopping criterion shifts from SC1 to SC2, indicating SC2's ability to select a more comprehensive feature set that optimizes classification performance. On the CK+ dataset (*Table IV-6*), most descriptors achieve 100% RR under SC2, except for BSIF, which shows a slight decrease. Similarly, on the MUG dataset (*Table IV-7*), all descriptors except BSIF reach 100% RR under SC2, highlighting SC2's effectiveness.

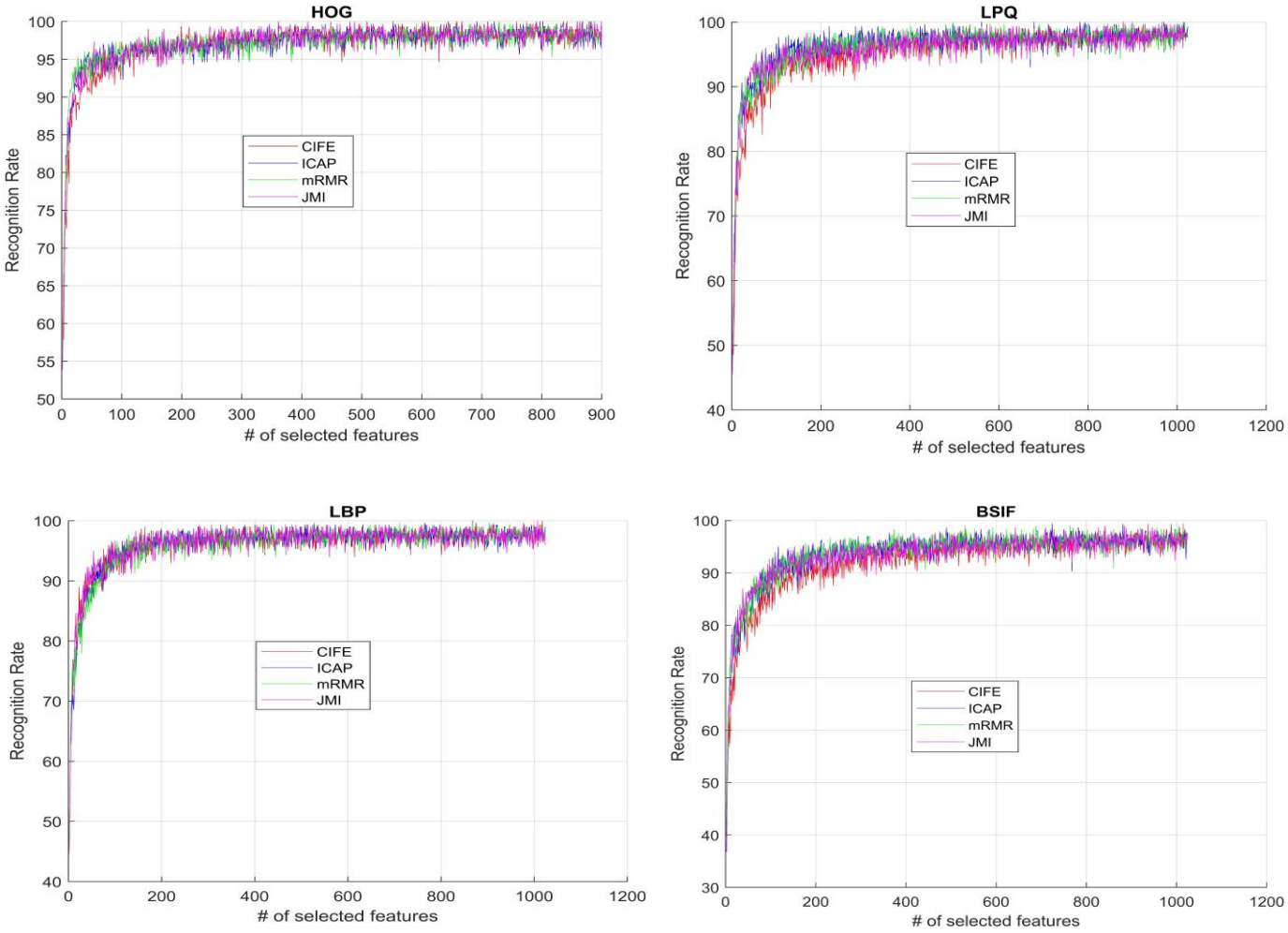
The number of selected features varies significantly across feature selection methods and descriptors. SC1 typically selects fewer features, prioritizing computational efficiency, as seen with HOG on MUG, where ICAP selects only 94 features compared to 208 by CIFE. In contrast, SC2 prioritizes higher RRs by selecting more features, but this requires more computation. For example, on the CK+ dataset, LBP under SC1 selects 379 features for CIFE strategy, which increases to 1017 under SC2.

Figure IV-3 (a) and *(b)* illustrate the results achieved by the four feature selection methods (CIFE, ICAP, JMI, and mRMR) applied to all descriptors on the CK+ and MUG datasets, respectively.

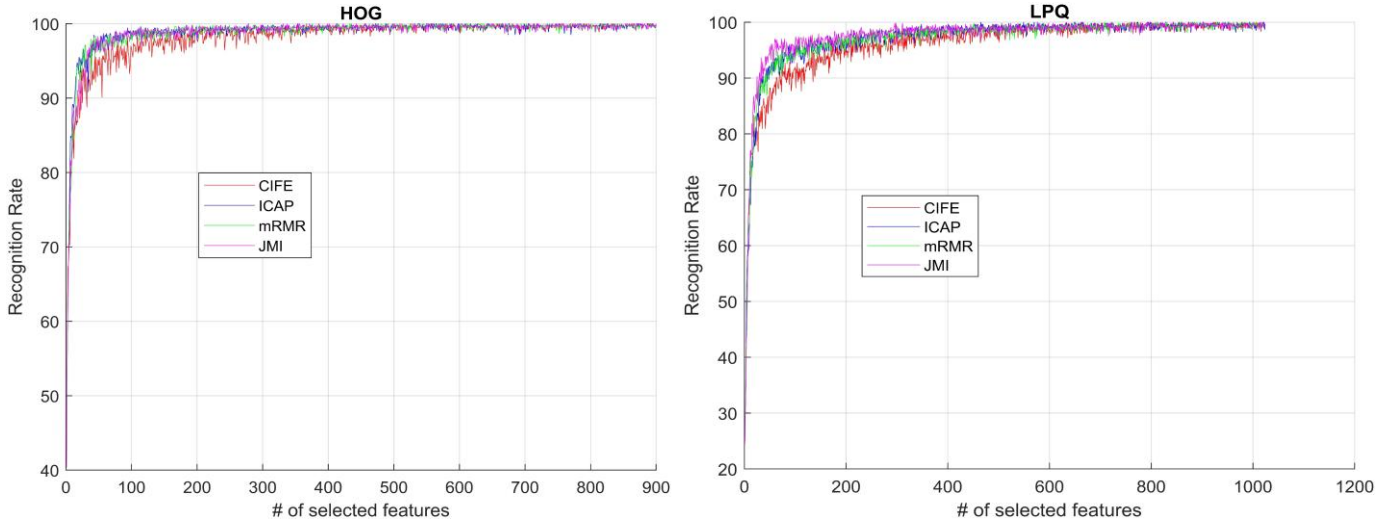
It is worth noting that all the curves exhibit an approximate plateau once 20% of the total features are selected by any of the feature selection strategies. This means that dimensionality reduction can be effectively done for both datasets to reduce the processing time and lower memory consumption.

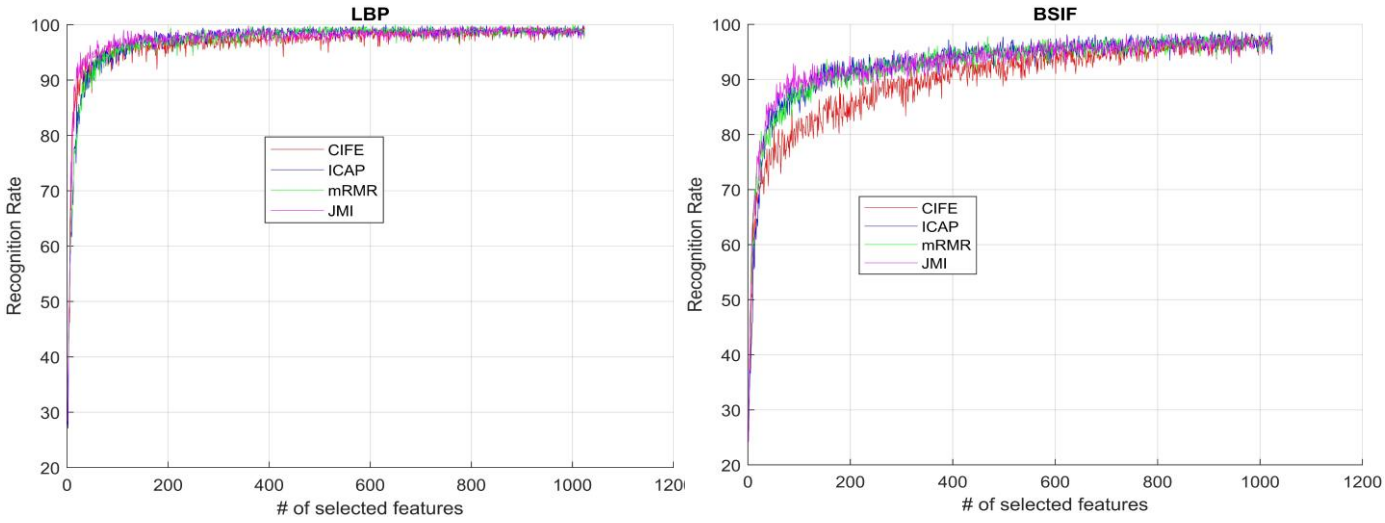
The results achieved using LPQ features are similar to those obtained with HoG and LBP, as noted in the previous study, where BSIF demonstrated less impressive results.

Among the feature selection strategies, CIFE performs slightly worse than the others on both datasets.



(a) CK+ Dataset





(b) MUG Dataset

Figure IV-3:graphical representation of feature selection strategies with various descriptors

IV.3.4.2. Influence on Computational Efficiency

In this section, we assess the advantages of the feature selection process with respect to system complexity, particularly in terms of processing time, and its impact on the system's RR. The HOG descriptor, combined with various feature selection strategies, was chosen for this experiment as it demonstrated strong performance in the previous section in terms of both accuracy and the number of selected features.

Tables IV-8 and 9 present the number of selected features using various feature selection strategies, along with their corresponding RR, Reduction Rate of Processing Time (RRT) and Reduction Rate of Features (RRF) given as follows:

$$\mathrm{RRT}(\%) = \frac{PTF - PTS}{\mathrm{PTF}} \times 100 \qquad (IV\text{-}1)$$

$$\mathrm{RRF}\ (\%) = \frac{TNF - NSF}{\mathrm{TNF}} \times 100 \qquad (IV\text{-}2)$$

Where:

- **PTF** is the processing time for full features.
- **PTS** is the processing time for the selected features.
- **TNF** is the total number of features.
- **NSF** is the number of selected features.

Table IV-8:Number of Selected Features, RR, RRF, and RRT for Feature Selection Strategies on the CK+ Dataset

	No. of selected features	RR (%)	RRF (%)	RRT (%)
CIFE	361	100	59.88	62.03
ICAP	692	100	23.11	17.40
JMI	361	100	59.88	62.03
mRMR	729	100	19.00	13.20

From *Table IV-8*, we can conclude several key observations regarding the impact of different feature selection strategies (CIFE, ICAP, JMI, and mRMR):

CIFE and JMI select 361 features from the full set of 900, showing greater efficiency compared to ICAP and mRMR, which select 692 and 729 features, respectively.

CIFE and JMI achieve higher processing time reduction rates (62.03%) compared to ICAP (17.40%) and mRMR (13.20%).

All strategies enhance the system's RR to 100%, demonstrating the crucial role of feature selection process in improving the performance of SER system.

Table IV-9:Number of Selected Features, RR, RRF, and RRT for Feature Selection Strategies on the MUG Dataset

	No. of selected features	RR (%)	RRF (%)	RRT (%)
CIFE	367	100	59.22	61.42
ICAP	278	100	69.11	63.82
JMI	331	100	63.22	62.62
mRMR	376	100	58.22	59.85

Table IV-9 provides several important insights into the effects of various feature selection strategies:

ICAP achieves the highest RRT of 63.82%, reflecting the highest efficiency in reducing computation time compared with the full feature set. JMI and CIFE also show considerable

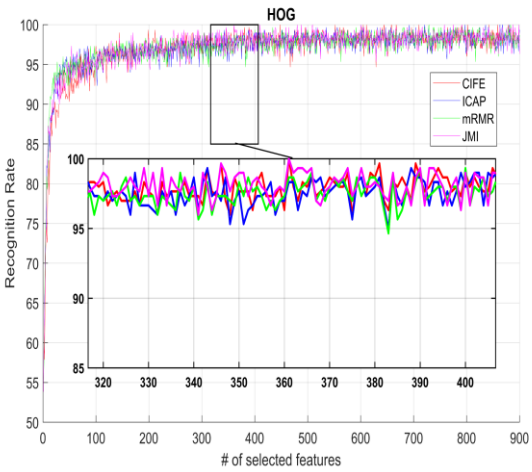
reduction rates of 62.62% and 61.42%, respectively, while mRMR has the lowest RRT equal to 59.85%. Thus, ICAP proves the highest computational performance.

ICAP is the most efficient method, with the best RRF (69.11%) compared to other strategies.

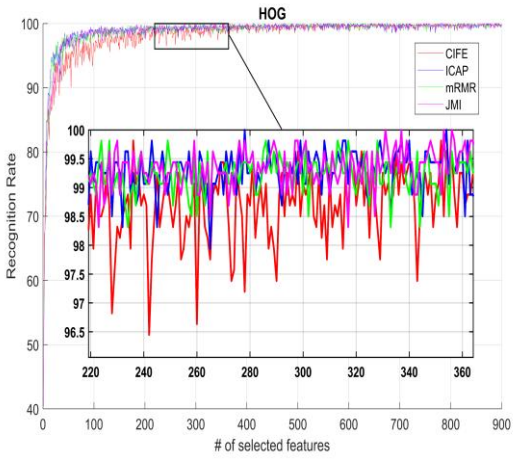
All methods achieve a perfect RR of 100%, highlighting the importance of feature selection strategies.

Both tables highlight the selection of a compact feature set, which significantly reduces processing time. This optimization not only accelerates computations but also enhances suitability for real-time applications and environments with limited computational resources.

Figure IV-4 provides a detailed (zoomed) visualization of the relationship between the number of selected features and the corresponding RR obtained using the HOG descriptor under various MI-based selection strategies, namely CIFE, ICAP, JMI, and mRMR. As illustrated in *Figure IV-4(a)*, both the JMI and CIFE criteria achieve the maximum recognition accuracy of 100% when approximately 361 features are selected, highlighting their strong capability to identify the most discriminative subset of features. Similarly, *Figure IV-4(b)* demonstrates that the ICAP strategy attains an equivalent 100% accuracy with only 278 selected features, confirming its superior efficiency in reducing dimensionality while preserving full classification performance. These results emphasize the effectiveness of MI-based feature selection in achieving an optimal balance between compact feature representation and high recognition accuracy.



(a) CK+ HOG Feature Selection



(b) MUG HOG Feature Selection

Figure IV-4:Zoomed View of Selected Features and RR Using Different Feature Selection Strategies with the HoG Descriptor

IV.4. FER SYSTEM BASED ON DEEP FEATURES AND MI-BASED FEATURE SELECTION

IV.4.1. Performance with KNN using Deep Features

In this study, AlexNet was employed in the feature extraction phase of the FER system. By leveraging a transfer learning approach, the need to re-train AlexNet was eliminated within the FER task, resulting in significant time savings. The fully connected layers of the AlexNet model (FC6, FC7, and FC8) were employed for feature extraction, as they provide compact and discriminative representations of facial expressions. Furthermore, these three layers were also combined in order to evaluate whether aggregating multiple layers could enhance classification performance.

Table IV-10 presents the dimensionality of the feature vectors extracted from each fully connected layer, as well as the combined representation.

Table IV-10:Number of Features for Fully Connected Layers in AlexNet

Fully Connected Layers	Number of Features
FC6	4096
FC7	4096
FC8	1000
Combined Layers	9192

The subsequent experiments focus on the optimal configuration of the KNN classifier by determining the best number of neighbors (K) and the most effective distance metric (Euclidean, Cityblock, Correlation, and Cosine). Performance was evaluated on the features extracted from each fully connected layer (FC6, FC7, and FC8), as well as their combination.

IV.4.1.1. KNN Optimization with AlexNet FC6 Features

The FC6 layer provides a vector with 4096 features. *Table IV-11* reports the recognition rates for different values of K (from 1 to 10) across the four distance metrics.

Table IV-11:Performance of KNN Classifier Using Different Distance Metrics with AlexNet FC6 Features

K	Euclidean	Cityblock	Correlation	Cosine
1	94.31	93.97	94.31	94.31
2	79.93	79.93	80.93	80.93
3	73.91	74.58	71.57	71.90
4	67.22	67.55	63.54	64.21
5	66.22	66.88	63.21	63.54
6	67.55	67.89	66.55	67.89
7	72.90	72.57	69.89	69.56
8	73.24	73.91	71.57	71.57
9	74.91	75.25	71.90	71.90
10	76.58	77.59	73.24	71.57

The highest recognition rate (94.31%) is obtained at K=1 with all distances except Cityblock. Recognition rates decrease sharply for K=2 to K=5, dropping to approximately 66%, indicating that larger K reduces the influence of the nearest neighbor. Beyond K=5, recognition rates slightly improve, reaching 76.58% at K=10 (Euclidean).

IV.4.1.2. KNN Optimization with AlexNet FC7 Features

The FC7 layer also produces 4096-dimensional feature vectors. *Table IV-12* summarizes the recognition rates.

Table IV-12:Performance of KNN Classifier Using Different Distance Metrics with AlexNet FC7 Features

K	Euclidean	Cityblock	Correlation	Cosine
1	95.32	95.31	95.31	96.32
2	79.59	79.59	79.93	79.93
3	73.91	73.91	71.90	71.23
4	63.54	63.87	65.55	65.21
5	62.20	62.54	62.20	61.20
6	63.21	63.21	65.88	65.21
7	67.89	66.88	69.23	68.56
8	69.56	68.56	70.56	70.23
9	69.56	69.89	70.56	70.23
10	69.23	68.89	68.89	71.90

The best performance is achieved with K=1 and Cosine distance (96.32%). Correlation and Cosine yield more stable results compared to Euclidean and Cityblock as K increases, suggesting that they better capture higher-level dependencies in FC7 features.

IV.4.1.3. KNN Optimization with AlexNet FC8 Features

The FC8 layer produces a lower-dimensional vector (1000 features). *Table IV-13* summarizes the results.

Table IV-13:Performance of KNN Classifier Using Different Distance Metrics with AlexNet FC8 Features

K	Euclidean	Cityblock	Correlation	Cosine
1	95.65	95.31	95.65	95.65
2	80.93	80.93	81.60	81.60
3	75.25	74.58	74.91	74.91
4	66.88	65.55	70.23	70.23
5	62.87	63.87	67.22	67.22
6	62.54	63.21	68.56	68.56
7	62.21	65.88	67.22	67.22
8	65.88	67.22	67.89	67.89
9	66.22	65.88	65.55	65.55
10	65.21	66.22	66.22	66.22

At K=1, the classifier achieves the best recognition rate (95.65%) across all metrics except Cityblock (95.31%). Accuracy decreases with higher K, reflecting the drawback of considering too many neighbors.

IV.4.1.4. KNN Optimization with Combined AlexNet Features

Table IV-14 presents results when combining FC6, FC7, and FC8 features (9192 dimensions).

Table IV-14:Performance of KNN Classifier Using Combined AlexNet Features (FC6 + FC7 + FC8)

K	Euclidean	Cityblock	Correlation	Cosine
1	94.31	94.31	94.31	94.31
2	79.59	79.93	80.60	80.60
3	74.58	73.91	71.57	72.57
4	65.88	64.88	64.21	64.88

5	66.55	66.55	64.21	63.87
6	69.23	69.56	65.55	65.55
7	72.57	70.56	70.23	69.89
8	73.57	72.57	70.90	71.57
9	75.58	72.57	70.23	70.23
10	77.59	74.24	70.56	70.56

The highest accuracy (94.31%) is observed at K=1, consistent across all distance metrics. However, this configuration does not outperform FC8 alone, which achieved 95.65%. Therefore, FC8 features are selected for the subsequent feature selection experiments.

Although the FC7 layer achieves the highest recognition rate (96.32%) when using K=1 and Cosine distance, it produces a high-dimensional feature vector of 4096 features. In contrast, the FC8 layer provides a much lower-dimensional feature set with only 1000 features while maintaining a comparable recognition rate (95.65%). Despite a slight reduction in accuracy compared to FC7, FC8 is selected because the large decrease in the number of features makes it more suitable for subsequent feature selection, improves classifier efficiency, and reduces overall computational complexity. Furthermore, combining FC6, FC7, and FC8 features results in a very high-dimensional space (9192 features) without achieving better performance than FC8 alone. Therefore, FC8 features offer a better balance between recognition performance and feature dimensionality and are selected for the feature selection experiments presented in the next section.

IV.4.2. Impact of Feature Selection on FER System Performance

This section evaluates the effect of feature selection applied to FC8 features. *Table IV-15* reports the number of selected features and the corresponding RR obtained using JMI, ICAP, CIFE, and mRMR under two stopping criteria: SC1 (RR ≥ RR(end)) and ST2 (RR = max(RR)).

Table IV-15:Number of Relevant Features and Recognition Rates (RR) with Different Feature Selection Strategies

	SC1		SC2	
	# of relevant features	RR (%)	# of relevant features	RR (%)
JMI	18	95.65	196	97.00
ICAP	11	95.65	118	96.32
CIFE	980	95.65	980	95.65
mRMR	9	97.00	9	97.00

Compared with the full feature set of 1000 features (95.65%), feature selection not only reduces dimensionality but also improves performance. Notably, mRMR achieved the highest RR (97.00%) with only 9 features, demonstrating superior compactness and discriminative power. Similarly, JMI and ICAP required only a small subset of features (18 and 11, respectively) to reach or surpass the baseline. By contrast, CIFE selected nearly the entire set (980 features) without significant improvement, indicating limited effectiveness.

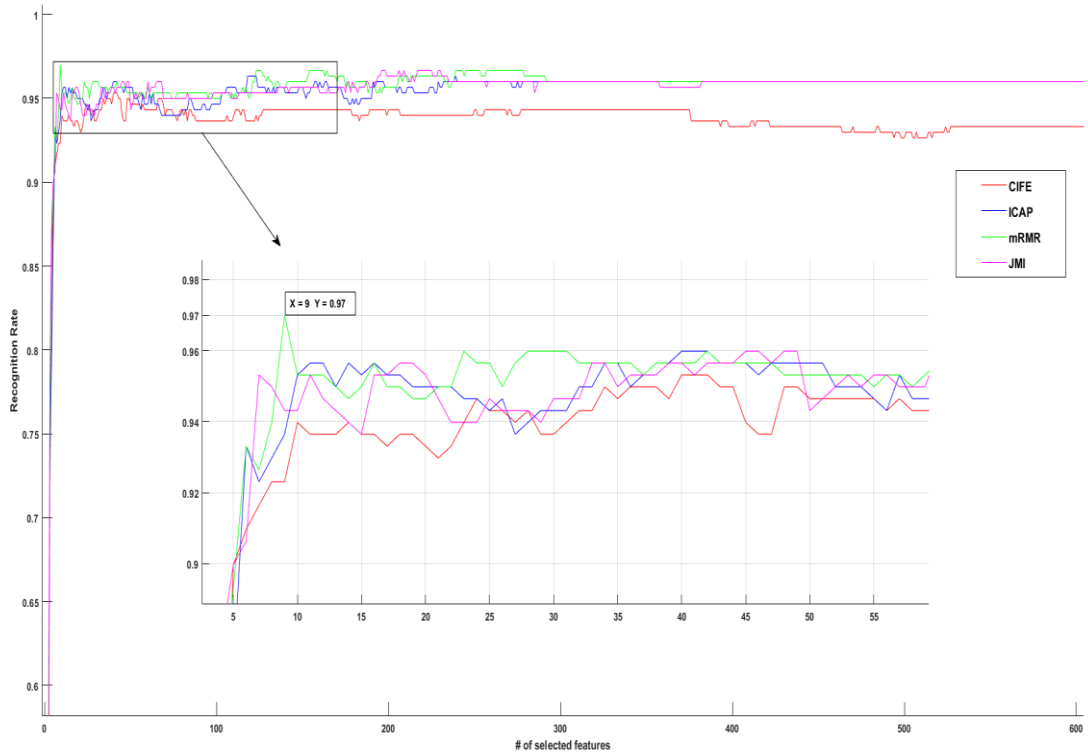


Figure IV-5: AlexNet Feature selection

Figure IV-5 illustrates the evolution of recognition rates as a function of the number of selected features. It can be observed that only a small number of features, about 10, using mRMR is sufficient to achieve top performance, while larger subsets offer diminishing returns. This behavior highlights the advantage of MI–based methods in overcoming the curse of dimensionality, where excessive features may degrade performance.

IV.5. CONCLUSION

This chapter presented the design and evaluation of an FER system based on both traditional and deep feature representations, combined with MI-based feature selection strategies. Two main approaches were investigated: (i) the use of handcrafted descriptors, specifically HOG, followed by feature selection; and (ii) the exploitation of deep features extracted from AlexNet through a transfer learning, also refined by MI-based feature selection.

In the first part, HOG descriptors combined with MI-based feature selection proved highly effective in reducing dimensionality and improving efficiency. On CK+, the system achieved 100% recognition with only 361 features (62.03% reduction in processing time) using CIFE or JMI, while on MUG, 278 features (63.82% reduction) with ICAP yielded comparable performance. These results confirm the importance of feature selection in overcoming the curse of dimensionality while maintaining high accuracy.

In the second part, deep features extracted from AlexNet were evaluated using a KNN classifier. Among the fully connected layers, FC8 achieved the best performance (95.65%), while combining features from multiple layers did not yield further improvements. To enhance efficiency, MI-based feature selection was applied to FC8 features. The results showed that methods such as JMI and ICAP reduced dimensionality while preserving accuracy, whereas mRMR achieved the highest recognition rate (97.00%) with only 9 features out of 1000, showing a significant reduction in dimensionality with superior performance.

These findings highlight that integrating feature selection with both traditional and deep features not only enhances recognition accuracy but also significantly improves computational efficiency, establishing a strong foundation for developing robust FER systems suitable for real-time applications.

CONCLUSION AND PERSPECTIVES

This thesis focused on the challenge of high dimensionality in emotion recognition, a domain where the rapid growth of data with large features often increases computational complexity and can ultimately degrade classification accuracy caused by the phenomenon of curse of dimensionality. The main objective of this thesis is to propose and evaluate MI-based feature selection strategies that effectively reduce dimensionality while maintaining, or even improving, recognition accuracy. Practically, we explored two of the most widely studied modalities for emotion recognition, speech and facial expressions, as they represent practical and less constrained approaches.

For SER, we investigated the application of several machine learning algorithms, including KNN, SVM, and GMM combined with voting rules, together with MI-based feature selection strategies such as CIFE, ICAP, mRMR, and JMI. Various acoustic feature extraction methods were employed, including MFCC, PLP, and LPCC. Results showed that the GMM classifier achieved an accuracy of 85.27% with 39 MFCC features, compared to 82.55% with a high-dimensional feature vector of 111 features. Applying the JMI strategy to MFCC features reduced the feature set by 23.07%, while improving accuracy to 86.82%. In high-dimensional settings, the ICAP strategy achieved 82.94% accuracy with a substantial 62.2% reduction in dimensionality.

We further introduced a 1D-CNN-based approach and proposed a stopping criterion for feature selection guided by classification accuracy. The results demonstrated that 1D-CNN outperformed traditional GMM classifier, achieving 91.09% with 39 selected MFCC features and 91.47% with 111 features. Using the CIFE strategy with our stopping criterion resulted in a 73.87% reduction in the feature vector size, with only a 0.39% decrease in accuracy, thereby ensuring an optimal trade-off between complexity and performance.

For FER, we applied both handcrafted feature extraction methods (HoG, LBP, LPQ, BSIF) and deep feature extraction with AlexNet. The same MI-based feature selection strategies were used as in SER. The proposed 1D-CNN system achieved a 100% recognition rate using only 361 HoG features out of 1024, corresponding to a 62.03% reduction in processing time on the CK+ dataset when applying CIFE or JMI. On the MUG dataset, 278 HoG features were sufficient to achieve a 63.82% reduction with ICAP. Furthermore, AlexNet FC8 features

classified with KNN achieved 95.65%, while the mRMR strategy improved performance to 97% using only 9 features out of 1000, representing a 99.1% reduction.

These findings demonstrate that mutual information-based feature selection strategies, particularly when combined with deep learning models, significantly enhance emotion recognition performance while reducing computational complexity.

This research provides significant contributions to the domain of emotion recognition, offering both methodological advances and empirical validation. Nonetheless, several avenues remain open for further exploration:

1. Multimodal Emotion Recognition:

   Future work should investigate the fusion of multiple modalities such as speech, facial expressions, physiological signals, and body gestures. While multimodal systems typically yield higher recognition rates, they introduce considerable complexity. Feature selection will therefore play a key role in reducing redundancy and ensuring tractable system design.

2. Integration of Vision Transformers (ViTs):

   Recently, ViTs have gained prominence as powerful feature extractors. Unlike CNNs, ViTs use self-attention to capture long-range dependencies, making them well suited for tasks like emotion recognition. Future research could explore the integration of feature selection with ViTs in order to reduce feature dimensionality and enhance system performance.

3. Exploration of Alternative Feature Selection Techniques:

   While this thesis primarily focused on MI-based feature selection, other advanced approaches, such as ReliefF, LASSO and genetic algorithms, could be explored and systematically compared to MI-based methods. Such work would contribute to identifying the most effective strategies across different modalities and datasets.

Through these perspectives, future research can advance towards more accurate, efficient, and scalable emotion recognition systems, further pushing the boundaries of affective computing.

BIBLIOGRAPHY

[1] S. Yu, A. Androsov, H. Yan, and Y. Chen, "Bridging computer and education sciences: A systematic review of automated emotion recognition in online learning environments," *Computers & Education*, vol. 220, p. 105111, Oct. 2024, doi: 10.1016/j.compedu.2024.105111.

[2] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Emotion Recognition and Its Applications," in *Human-Computer Systems Interaction: Backgrounds and Applications 3*, Z. S. Hippe, J. L. Kulikowski, T. Mroczek, and J. Wtorek, Eds., Cham: Springer International Publishing, 2014, pp. 51–62. doi: 10.1007/978-3-319-08491-6_5.

[3] B. H. Krishna *et al.*, "Emotion-net: Automatic emotion recognition system using optimal feature selection-based hidden markov CNN model," *Ain Shams Engineering Journal*, vol. 15, no. 12, p. 103038, Dec. 2024, doi: 10.1016/j.asej.2024.103038.

[4] D. K. Rakesh and P. K. Jana, "A General Framework for Class Label Specific Mutual Information Feature Selection Method," *IEEE Transactions on Information Theory*, vol. 68, no. 12, pp. 7996–8014, Dec. 2022, doi: 10.1109/TIT.2022.3188708.

[5] E. M. G. Younis, S. Mohsen, E. H. Houssein, and O. A. S. Ibrahim, "Machine learning for human emotion recognition: a comprehensive review," *Neural Comput & Applic*, vol. 36, no. 16, pp. 8901–8947, Jun. 2024, doi: 10.1007/s00521-024-09426-2.

[6] R. Pereira *et al.*, "Systematic Review of Emotion Detection with Computer Vision and Deep Learning," *Sensors (Basel)*, vol. 24, no. 11, p. 3484, May 2024, doi: 10.3390/s24113484.

[7] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebraska Symposium on Motivation*, vol. 19, pp. 207–283, 1971.

[8] "Definition of EMOTION." Accessed: Mar. 11, 2025. [Online]. Available: https://www.merriam-webster.com/dictionary/emotion

[9] "Guidelines for Designing Computational Models of Emotions: Computer Science & IT Journal Article | IGI Global Scientific Publishing." Accessed: Mar. 11, 2025. [Online]. Available: https://www.igi-global.com/article/international-journal-synthetic-emotions-ijse/52755

[10] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019, doi: 10.1109/ACCESS.2019.2929050.

[11] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014, doi: 10.1016/j.neuroimage.2013.11.007.

[12] N. A. Puccetti, W. J. Villano, J. P. Fadok, and A. S. Heller, "Temporal dynamics of affect in the brain: Evidence from human imaging and animal models," *Neuroscience & Biobehavioral Reviews*, vol. 133, p. 104491, Feb. 2022, doi: 10.1016/j.neubiorev.2021.12.014.

[13] D. A. Trevisan and E. Birmingham, "Are emotion recognition abilities related to everyday social functioning in ASD? A meta-analysis," *Research in Autism Spectrum Disorders*, vol. 32, pp. 24–42, Dec. 2016, doi: 10.1016/j.rasd.2016.08.004.

[14] N. Ahmed, Z. A. Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, p. 200171, Feb. 2023, doi: 10.1016/j.iswa.2022.200171.

[15] L. Schoneveld and A. Othmani, "Towards a General Deep Feature Extractor for Facial Expression Recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*, Sep. 2021, pp. 2339–2342. doi: 10.1109/ICIP42928.2021.9506025.

[16] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," *Applied Sciences*, vol. 13, no. 8, p. 4750, Jan. 2023, doi: 10.3390/app13084750.

[17] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "A Transformer-based Hybrid Model for Implicit Emotion Recognition in Arabic Text," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23834–23839, Jun. 2025, doi: 10.48084/etasr.10261.

[18] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review," *Applied Sciences*, vol. 14, no. 17, p. 8071, Jan. 2024, doi: 10.3390/app14178071.

[19] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," *Entropy*, vol. 25, no. 10, p. 1440, Oct. 2023, doi: 10.3390/e25101440.

[20] Y. Cai, X. Li, and J. Li, "Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review," *Sensors*, vol. 23, no. 5, p. 2455, Jan. 2023, doi: 10.3390/s23052455.

[21] P. Theerthagiri, "Stress emotion recognition with discrepancy reduction using transfer learning," *Multimed Tools Appl*, vol. 82, no. 4, pp. 5949–5963, Feb. 2023, doi: 10.1007/s11042-022-13593-6.

[22] L. M. Rappaport *et al.*, "Pediatric anxiety associated with altered facial emotion recognition," *Journal of Anxiety Disorders*, vol. 82, p. 102432, Aug. 2021, doi: 10.1016/j.janxdis.2021.102432.

[23] W. Zheng, L. Yan, and F.-Y. Wang, "Two Birds With One Stone: Knowledge-Embedded Temporal Convolutional Transformer for Depression Detection and Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2595–2613, Oct. 2023, doi: 10.1109/TAFFC.2023.3282704.

[24] A. Alshaer, "Improving telemedicine: evaluating emotional recognition for better patient-consultant interaction," *J. Umm Al-Qura Univ. Eng.Archit.*, vol. 16, no. 1, pp. 196–205, Mar. 2025, doi: 10.1007/s43995-025-00101-8.

[25] M. Aly, "Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model," *Multimed Tools Appl*, Jun. 2024, doi: 10.1007/s11042-024-19392-5.

[26] A. O. R. Vistorte, A. Deroncele-Acosta, J. L. M. Ayala, A. Barrasa, C. López-Granero, and M. Martí-González, "Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review," *Front. Psychol.*, vol. 15, Jun. 2024, doi: 10.3389/fpsyg.2024.1387089.

[27] J. Hernandez *et al.*, "Guidelines for Assessing and Minimizing Risks of Emotion Recognition Applications," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2021, pp. 1–8. doi: 10.1109/ACII52823.2021.9597452.

[28] M. Monaro, S. Maldera, C. Scarpazza, G. Sartori, and N. Navarin, "Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models," *Computers in Human Behavior*, vol. 127, p. 107063, Feb. 2022, doi: 10.1016/j.chb.2021.107063.

[29] X. Zhang, X. Yang, W. Zhang, G. Li, and H. Yu, "Crowd emotion evaluation based on fuzzy inference of arousal and valence," *Neurocomputing*, vol. 445, pp. 194–205, Jul. 2021, doi: 10.1016/j.neucom.2021.02.047.

[30] C. N. W. Geraets *et al.*, "Virtual reality facial emotion recognition in social environments: An eye-tracking study," *Internet Interventions*, vol. 25, p. 100432, Sep. 2021, doi: 10.1016/j.invent.2021.100432.

[31] T. Souto, H. Silva, Â. Leite, A. Baptista, C. Queirós, and A. Marques, "Facial Emotion Recognition: Virtual Reality Program for Facial Emotion Recognition—A Trial Program Targeted at Individuals With Schizophrenia," *Rehabilitation Counseling Bulletin*, vol. 63, p. 003435521984728, May 2019, doi: 10.1177/0034355219847284.

[32] Z. Xu and S. Liu, "Decoding consumer purchase decisions: exploring the predictive power of EEG features in online shopping environments using machine learning," *Humanit Soc Sci Commun*, vol. 11, no. 1, p. 1202, Sep. 2024, doi: 10.1057/s41599-024-03691-1.

[33] Y. Guo, Y. Li, D. Liu, and S. X. Xu, "Measuring service quality based on customer emotion: An explainable AI approach," *Decision Support Systems*, vol. 176, p. 114051, Jan. 2024, doi: 10.1016/j.dss.2023.114051.

[34] P. Hajek and M. Munk, "Speech emotion recognition and text sentiment analysis for financial distress prediction," *Neural Comput & Applic*, vol. 35, no. 29, pp. 21463–21477, Oct. 2023, doi: 10.1007/s00521-023-08470-8.

[35] A. Hacine-Gharbi and P. Ravier, "On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, pp. 1074–1083, Jul. 2019, doi: 10.1016/j.jksuci.2019.07.008.

[36] P. Tiwari and A. D. Darji, "Pertinent feature selection techniques for automatic emotion recognition in stressed speech," *Int J Speech Technol*, vol. 25, no. 2, pp. 511–526, Jun. 2022, doi: 10.1007/s10772-022-09978-5.

[37] Z. Liang *et al.*, "EEGFuseNet: Hybrid Unsupervised Deep Feature Characterization and Fusion for High-Dimensional EEG With an Application to Emotion Recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1913–1925, 2021, doi: 10.1109/TNSRE.2021.3111689.

[38] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–31, 2023, doi: 10.1109/TIM.2023.3243661.

[39] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 6, p. e1563, 2024, doi: 10.1002/widm.1563.

[40] A. Nfissi, W. Bouachir, N. Bouguila, and B. L. Mishara, "Iterative Feature Boosting for Explainable Speech Emotion Recognition," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2023, pp. 543–549. doi: 10.1109/ICMLA58977.2023.00081.

[41] K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," *Electronics*, vol. 12, no. 4, Art. no. 4, Jan. 2023, doi: 10.3390/electronics12040839.

[42] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Processing*, vol. 12, no. 6, pp. 713–721, 2018, doi: 10.1049/iet-spr.2017.0320.

[43] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894.

[44] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network," *Sensors*, vol. 20, no. 21, Art. no. 21, Jan. 2020, doi: 10.3390/s20216008.

[45] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets," *Information*, vol. 15, no. 3, p. 135, Mar. 2024, doi: 10.3390/info15030135.

[46] S. Qin, Z. Zhu, Y. Zou, and X. Wang, "Facial expression recognition based on Gabor wavelet transform and 2-channel CNN," *International Journal of Wavelets, Multiresolution and Information Processing*, Nov. 2019, doi: 10.1142/S0219691320500034.

[47] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," *IEEE Access*, vol. 12, pp. 45543–45559, 2024, doi: 10.1109/ACCESS.2024.3380847.

[48] R. I. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Ahmed, "Improved Facial Expression Recognition Based on DWT Feature for Deep CNN," *Electronics*, vol. 8, no. 3, Art. no. 3, Mar. 2019, doi: 10.3390/electronics8030324.

[49] F. Z. Boukhobza, A. Hacine Gharbi, and K. Rouabah, "A New Facial Expression Recognition Algorithm Based on DWT Feature Extraction and Selection," *IAJIT*, vol. 21, no. 4, 2024, doi: 10.34028/iajit/21/4/6.

[50] J. Yi, A. Chen, Z. Cai, Y. Sima, M. Zhou, and X. Wu, "Facial expression recognition of intercepted video sequences based on feature point movement trend and feature block texture variation," *Applied Soft Computing*, vol. 82, p. 105540, Sep. 2019, doi: 10.1016/j.asoc.2019.105540.

[51] H. Ghazouani, "A genetic programming-based feature selection and fusion for facial expression recognition," *Applied Soft Computing*, vol. 103, p. 107173, May 2021, doi: 10.1016/j.asoc.2021.107173.

[52] F. Boukhobza, A. Gharbi, K. Rouabah, and P. Ravier, "LOCAL FEATURE SELECTION USING THE WRAPPER APPROACH FOR FACIAL EXPRESSION RECOGNITION," *JJCIT*, no. 0, p. 1, 2024, doi: 10.5455/jjcit.71-1713081709.

[53] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," *Sensors*, vol. 23, no. 12, p. 5475, Jan. 2023, doi: 10.3390/s23125475.

[54] M. M. R. Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," *PLOS ONE*, vol. 18, no. 11, p. e0291500, Nov. 2023, doi: 10.1371/journal.pone.0291500.

[55] S. A, A. Thomas, and D. Mathew, "Study of MFCC and IHC Feature Extraction Methods With Probabilistic Acoustic Models for Speaker Biometric Applications," *Procedia Computer Science*, vol. 143, pp. 267–276, Jan. 2018, doi: 10.1016/j.procs.2018.10.395.

[56] M. Mohan, P. Dhanalakshmi, and R. S. Kumar, "Speech Emotion Classification using Ensemble Models with MFCC," *Procedia Computer Science*, vol. 218, pp. 1857–1868, Jan. 2023, doi: 10.1016/j.procs.2023.01.163.

[57] "Improved Speech Emotion Recognition Focusing on High-Level Data Representations and Swift Feature Extraction Calculation," *Computers, Materials and Continua*, vol. 77, no. 3, pp. 2915–2933, Dec. 2023, doi: 10.32604/cmc.2023.044466.

[58] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990, doi: 10.1121/1.399423.

[59] R. Deshmukh, P. Kurzekar, Dr. V. Waghmare, and P. Shrishrimal, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 3, pp. 18006–18016, Dec. 2014, doi: 10.15680/IJIRSET.2014.0312034.

[60] T. Gulzar, A. Singh, and S. Sharma, "Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks," *International Journal of Computer Applications*, vol. 101, pp. 22–27, Sep. 2014, doi: 10.5120/17740-8271.

[61] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Jun. 2005, pp. 886–893 vol. 1. doi: 10.1109/CVPR.2005.177.

[62] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.

[63] V. Ojansivu and J. Heikkilä, "Blur Insensitive Texture Classification Using Local Phase Quantization," in *Image and Signal Processing*, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds., Berlin, Heidelberg: Springer, 2008, pp. 236–243. doi: 10.1007/978-3-540-69905-7_27.

[64] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov. 2012, pp. 1363–1366. Accessed: Feb. 01, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/6460393

[65] M. Bennaoum, A. Hacine-Gharbi, K. Rouabah, A. Souahlia, and H. Roubhi, "Fingerprint Recognition Improvement Using BSIF Features and KNN Classifier," in *2024 International Conference on Telecommunications and Intelligent Systems (ICTIS)*, Dec. 2024, pp. 1–6. doi: 10.1109/ICTIS62692.2024.10894235.

[66] M. Hussain, J. Bird, and D. Resende Faria, *A Study on CNN Transfer Learning for Image Classification*. 2018.

[67] H. S. Mputu, A. Abdel-Mawgood, A. Shimada, and M. S. Sayed, "Tomato Quality Classification Based on Transfer Learning Feature Extraction and Machine Learning Algorithm Classifiers," *IEEE Access*, vol. 12, pp. 8283–8295, 2024, doi: 10.1109/ACCESS.2024.3352745.

[68] E. O. Belabbaci, M. Khammari, A. Chouchane, A. Ouamane, and M. Bessaoudi, "Kinship Verification Using Multiscale Retinex Preprocessing and Integrated 2DSWT-CNN Features," in *2024 8th International Conference on Image and Signal Processing and their Applications (ISPA)*, Apr. 2024, pp. 1–8. doi: 10.1109/ISPA59904.2024.10536858.

[69] A. Chouchane, M. Bessaoudi, H. Kheddar, A. Ouamane, T. Vieira, and M. Hassaballah, "Multilinear subspace learning for Person Re-Identification based fusion of high order tensor features," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107521, Feb. 2024, doi: 10.1016/j.engappai.2023.107521.

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012. Accessed: Nov. 10, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[71] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[72] S. Sharma, K. Guleria, S. Tiwari, and S. Kumar, "A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans," *Measurement: Sensors*, vol. 24, p. 100506, Dec. 2022, doi: 10.1016/j.measen.2022.100506.

[73] D. Albashish, R. Al-Sayyed, A. Abdullah, M. Ryalat, and N. Almansour, *Deep CNN Model based on VGG16 for Breast Cancer Classification*. 2021, p. 810. doi: 10.1109/ICIT52682.2021.9491631.

[74] Mustaqeem and S. Kwon, "CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network," *Mathematics*, vol. 8, no. 12, p. 2133, Dec. 2020, doi: 10.3390/math8122133.

[75] D. Reynolds, "Gaussian Mixture Models," in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds., Boston, MA: Springer US, 2009, pp. 659–663. doi: 10.1007/978-0-387-73003-5_196.

[76] G. H. F. M. Oliveira, L. L. Minku, and A. L. I. Oliveira, "GMM-VRD: A Gaussian Mixture Model for Dealing With Virtual and Real Concept Drifts," in *2019 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8852097.

[77] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989, doi: 10.2307/1403797.

[78] Y. Habchi *et al.*, "AI in Thyroid Cancer Diagnosis: Techniques, Trends, and Future Directions," *Systems*, vol. 11, no. 10, p. 519, Oct. 2023, doi: 10.3390/systems11100519.

[79] H. A. Abu Alfeilat *et al.*, "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review," *Big Data*, vol. 7, no. 4, pp. 221–248, Dec. 2019, doi: 10.1089/big.2018.0175.

[80] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013.

[81] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems*, MIT Press, 1996. Accessed: Aug. 10, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html

[82] M. Razzaghi, S. Shekarpaz, and A. Rajabi, "Solving Ordinary Differential Equations by LS-SVM," 2023, pp. 147–170. doi: 10.1007/978-981-19-6553-1_7.

[83] L. L. Iglesias *et al.*, "A primer on deep learning and convolutional neural networks for clinicians," *Insights Imaging*, vol. 12, p. 117, Aug. 2021, doi: 10.1186/s13244-021-01052-z.

[84] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[85] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, Apr. 2021, doi: 10.1016/j.ymssp.2020.107398.

[86] S. Huang, J. Tang, J. Dai, and Y. Wang, "Signal Status Recognition Based on 1DCNN and Its Feature Extraction Mechanism Analysis," *Sensors (Basel)*, vol. 19, no. 9, p. 2018, Apr. 2019, doi: 10.3390/s19092018.

[87] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang, "Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study," *Neural Comput & Applic*, vol. 34, no. 7, pp. 5321–5347, Apr. 2022, doi: 10.1007/s00521-022-06953-8.

[88] Y. Kim and Y.-K. Kim, "Time-Frequency Multi-Domain 1D Convolutional Neural Network with Channel-Spatial Attention for Noise-Robust Bearing Fault Diagnosis," *Sensors*, vol. 23, no. 23, p. 9311, Jan. 2023, doi: 10.3390/s23239311.

[89] S. H. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, Feb. 2020, doi: 10.1016/j.neucom.2019.10.008.

[90] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, Nov. 2019, doi: 10.1109/TASLP.2019.2925934.

[91] J. de Lope and M. Graña, "An ongoing review of speech emotion recognition," *Neurocomputing*, vol. 528, pp. 1–11, Apr. 2023, doi: 10.1016/j.neucom.2023.01.002.

[92] D. Bitouk, R. Verma, and A. Nenkova, "Class-Level Spectral Features for Emotion Recognition," *Speech Commun*, vol. 52, no. 7–8, pp. 613–625, 2010, doi: 10.1016/j.specom.2010.02.010.

[93] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *Journal of Big Data*, vol. 7, no. 1, p. 20, Mar. 2020, doi: 10.1186/s40537-020-00299-5.

[94] A. Khalil, M. Saad, K. Chaar, R. Tafreshi, S. Abdulla, and M. F. Wahid, "Enhanced Binary Classification of Gait Disorders Using a Machine Learning Majority Voting Approach," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2024, pp. 1–4. doi: 10.1109/EMBC53108.2024.10781639.

[95] A. Batool and Y.-C. Byun, "Toward Improving Breast Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm," *IEEE Access*, vol. 12, pp. 12869–12882, 2024, doi: 10.1109/ACCESS.2024.3356602.

[96] H. Roubhi, A. H. Gharbi, K. Rouabah, and P. Ravier, "Mutual Information-based Feature Selection Strategy for Speech Emotion Recognition using Machine Learning Algorithms Combined with the Voting Rules Method," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, Art. no. 1, Feb. 2025, doi: 10.48084/etasr.9066.

[97] F. Ghazali, A. Hacine-Gharbi, and P. Ravier, "Statistical features extraction based on the discrete wavelet transform for electrical appliances identification," in *Proceedings of the 1st International Conference on Intelligent Systems and Pattern Recognition*, in ISPR '20. New York, NY, USA: Association for Computing Machinery, décembre 2020, pp. 22–26. doi: 10.1145/3432867.3432900.

[98] A. Abdulboriy and J. S. Shin, "An Incremental Majority Voting Approach for Intrusion Detection System Based on Machine Learning," *IEEE Access*, vol. 12, pp. 18972–18986, 2024, doi: 10.1109/ACCESS.2024.3361041.

[99] F. GHAZALI, A. HACINE-GHARBI, P. RAVIER, and T. MOHAMADI, "Extraction and selection of statistical harmonics features for electrical appliancesidentification using k-NN classifier combined with voting rules method," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, no. 4, pp. 2980–2997, Jan. 2019, doi: 10.3906/elk-1812-80.

[100] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S. Yang Bang, "Constructing support vector machine ensemble," *Pattern Recognition*, vol. 36, no. 12, pp. 2757–2767, Dec. 2003, doi: 10.1016/S0031-3203(03)00175-4.

[101] J. M. Montgomery, F. M. Hollenbach, and M. D. Ward, "Improving Predictions using Ensemble Bayesian Model Averaging," *Political Analysis*, vol. 20, no. 3, pp. 271–291, Jul. 2012, doi: 10.1093/pan/mps002.

[102] G. R. Latif-Shabgahi, "A novel algorithm for weighted average voting used in fault tolerant computing systems," *Microprocessors and Microsystems*, vol. 28, no. 7, pp. 357–361, Sep. 2004, doi: 10.1016/j.micpro.2004.02.006.

[103] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection”.

[104] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, “Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection,” *Journal of Machine Learning Research*, vol. 13, no. 2, pp. 27–66, 2012.

[105] M. Verleysen and D. François, *The Curse of Dimensionality in Data Mining and Time Series Prediction*, vol. 3512. 2005, p. 770. doi: 10.1007/11494669_93.

[106] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural Comput & Applic*, vol. 24, no. 1, pp. 175–186, Jan. 2014, doi: 10.1007/s00521-013-1368-0.

[107] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.

[108] M. BÜYÜKKEÇECİ and M. Okur, “A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning,” *GAZI UNIVERSITY JOURNAL OF SCIENCE*, vol. 36, Sep. 2022, doi: 10.35378/gujs.993763.

[109] J. AK and D. Zongker, “Feature Selection: Evaluation, Application, and Small Sample Performance,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 153–158, Mar. 1997, doi: 10.1109/34.574797.

[110] K. Kira and L. A. Rendell, “The feature selection problem: traditional methods and a new algorithm,” in *Proceedings of the tenth national conference on Artificial intelligence*, in AAAI’92. San Jose, California: AAAI Press, juillet 1992, pp. 129–134.

[111] P. M. Narendra and K. Fukunaga, “A Branch and Bound Algorithm for Feature Subset Selection,” *IEEE Trans. Comput.*, vol. 26, no. 9, pp. 917–922, Sep. 1977, doi: 10.1109/TC.1977.1674939.

[112] D. Koller and M. Sahami, “Toward optimal feature selection,” in *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, in ICML’96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., juillet 1996, pp. 284–292.

[113] R. Kohavi and D. Sommerfield, “Feature subset selection using the wrapper method: overfltting and dynamic search space topology,” in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, in KDD’95. Montréal, Québec, Canada: AAAI Press, août 1995, pp. 192–197.

[114] M. Dash and H. Liu, “Feature selection for classification,” *Intelligent Data Analysis*, vol. 1, no. 1, pp. 131–156, Jan. 1997, doi: 10.1016/S1088-467X(97)00008-5.

[115] D. W. Aha and R. L. Bankert, “A Comparative Evaluation of Sequential Feature Selection Algorithms,” in *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher and H.-J. Lenz, Eds., New York, NY: Springer, 1996, pp. 199–206. doi: 10.1007/978-1-4612-2404-4_19.

[116] “FUNDAMENTALS of ALGORITHMS - Gilles Brassard and Paul Bratley - PDFCOFFEE.COM.” Accessed: Sep. 26, 2025. [Online]. Available: https://pdfcoffee.com/fundamentals-of-algorithms-gilles-brassard-and-paul-bratley-pdf-free.html

[117] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[118] J. Bins and B. A. Draper, “Feature selection from huge feature sets,” in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, Jul. 2001, pp. 159–165 vol.2. doi: 10.1109/ICCV.2001.937619.

[119] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, Jul. 1994, doi: 10.1109/72.298224.

[120] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy”.

[121] M. Dash and H. Liu, “Consistency-based search in feature selection,” *Artificial Intelligence*, vol. 151, no. 1, pp. 155–176, Dec. 2003, doi: 10.1016/S0004-3702(03)00079-1.

[122] V. Kumar, “Feature Selection: A literature Review,” *SmartCR*, vol. 4, no. 3, Jun. 2014, doi: 10.6029/smartcr.2014.03.007.

[123] P. Chauhan, N. Sharma, and H. Sharma, “GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES FEATURE SELECTION TECHNIQUES IN MACHINE LEARNING : A SURVEY”.

[124] R. Touahria, A. Hacine-Gharbi, and P. Ravier, “Feature selection algorithms highlight the importance of the systolic segment for normal/murmur PCG beat classification,” *Biomedical Signal Processing and Control*, vol. 86, p. 105288, 2023.

[125] R. Touahria, A. Hacine-Gharbi, P. Ravier, and M. Mostefai, “Relevant Multi Domain Features Selection Based on Mutual Information for Heart Sound Classification:,” in *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2024, pp. 918–923. doi: 10.5220/0012565300003654.

[126] A. Hacine-Gharbi and P. Ravier, “A binning formula of bi-histogram for joint entropy estimation using mean square error minimization,” *Pattern Recognition Letters*, vol. 101, pp. 21–28, Jan. 2018, doi: 10.1016/j.patrec.2017.11.007.

[127] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: a review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, Jan. 2000, doi: 10.1109/34.824819.

[128] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, Dec. 1997, doi: 10.1016/S0004-3702(97)00043-X.

[129] Y. Saeys, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics (Oxford, England)*, vol. 23, pp. 2507–17, Nov. 2007, doi: 10.1093/bioinformatics/btm344.

[130] A. Jovic, K. Brkić, and N. Bogunovic, *A review of feature selection methods with applications*. 2015, p. 1205. doi: 10.1109/MIPRO.2015.7160458.

[131] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, “MIFS-ND: A mutual information-based feature selection method,” *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, Oct. 2014, doi: 10.1016/j.eswa.2014.04.019.

[132] F. Azuaje, I. Witten, and F. E, “Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques,” *Biomedical Engineering Online - BIOMED ENG ONLINE*, vol. 5, pp. 1–2, Jan. 2006.

[133] L. Yu and H. Liu, *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*, vol. 2. 2003, p. 863.

[134] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[135] H. Yang and J. Moody, "Feature Selection Based on Joint Mutual Information," 1999. Accessed: Feb. 08, 2026. [Online]. Available: https://www.semanticscholar.org/paper/Feature-Selection-Based-on-Joint-Mutual-Information-Yang-Moody/dd691540c3f28decb477a7738f16aa92709b0f59

[136] D. Lin and X. Tang, "Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 68–82. doi: 10.1007/11744023_6.

[137] A. Jakulin, "Machine Learning Based on Attribute Interactions," phd, Univerza v Ljubljani, 2005. Accessed: Dec. 20, 2023. [Online]. Available: http://eprints.fri.uni-lj.si/205/

[138] R. Duda, P. Hart, and D. G.Stork, "Pattern Classification," in *Wiley Interscience*, vol. xx, 2001.

[139] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1, pp. 23–69, Oct. 2003, doi: 10.1023/A:1025667309714.

[140] S. Alelyani, J. Tang, and H. Liu, "Feature Selection for Clustering: A Review," in *Data Clustering*, 1st ed., C. C. Aggarwal and C. K. Reddy, Eds., Chapman and Hall/CRC, 2018, pp. 29–60. doi: 10.1201/9781315373515-2.

[141] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *J Am Stat Assoc*, vol. 105, no. 490, pp. 713–726, Jun. 2010, doi: 10.1198/jasa.2010.tm09415.

[142] Y. Li, M. Dong, and J. Hua, "Localized feature selection for clustering," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 10–18, Jan. 2008, doi: 10.1016/j.patrec.2007.08.012.

[143] D. S. Modha and W. S. Spangler, "Feature Weighting in k-Means Clustering," *Machine Learning*, vol. 52, no. 3, pp. 217–237, Sep. 2003, doi: 10.1023/A:1024016609528.

[144] M. Dash and Y.-S. Ong, "RELIEF-C: Efficient Feature Selection for Clustering over Noisy Data," in *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, Nov. 2011, pp. 869–872. doi: 10.1109/ICTAI.2011.135.

[145] A. H. Gharbi, P. Ravier, and M. N. Meziane, "Relevant harmonics selection based on mutual information for electrical appliances identification," *IJCAT*, vol. 62, no. 2, p. 102, 2020, doi: 10.1504/IJCAT.2020.104691.

[146] A. Jiménez-Cordero, J. M. Morales, and S. Pineda, "A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification," *European Journal of Operational Research*, vol. 293, no. 1, pp. 24–35, Aug. 2021, doi: 10.1016/j.ejor.2020.12.009.

[147] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*. Hoboken, N.J: Wiley-Interscience, 2006.

[148] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[149] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, Jun. 1954, doi: 10.1007/BF02289159.

[150] T. Tsujishita, "On Triple Mutual Information," *Advances in Applied Mathematics*, vol. 16, no. 3, pp. 269–274, Sep. 1995, doi: 10.1006/aama.1995.1013.

[151] A. Hacine-Gharbi, M. Deriche, P. Ravier, R. Harba, and T. Mohamadi, "A new histogram-based estimation technique of entropy and mutual information using mean squared error minimization," *Computers & Electrical Engineering*, vol. 39, no. 3, pp. 918–933, Apr. 2013, doi: 10.1016/j.compeleceng.2013.02.010.

[152] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002, doi: 10.1109/TPAMI.2002.1114861.

[153] H. Joe, "Estimation of Entropy and Other Functionals of a Multivariate Density," *Annals of the Institute of Statistical Mathematics*, vol. 41, pp. 683–697, Feb. 1989, doi: 10.1007/BF00057735.

[154] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1168–1174, Jul. 2010, doi: 10.1016/j.patrec.2009.11.010.

[155] H. A. Sturges, "The Choice of a Class Interval," *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, Mar. 1926, doi: 10.1080/01621459.1926.10502161.

[156] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979, doi: 10.1093/biomet/66.3.605.

[157] D. Freedman and P. Diaconis, "On the histogram as a density estimator:L2 theory," *Z. Wahrscheinlichkeitstheorie verw Gebiete*, vol. 57, no. 4, pp. 453–476, Dec. 1981, doi: 10.1007/BF01025868.

[158] A. Hacine-Gharbi, P. Ravier, R. Harba, and T. Mohamadi, "Low bias histogram-based estimation of mutual information for feature selection," *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1302–1308, Jul. 2012, doi: 10.1016/j.patrec.2012.02.022.

[159] P. A. Legg, P. L. Rosin, D. Marshall, and J. E. Morgan, "Improving accuracy and efficiency of registration by mutual information using Sturges' histogram rule," *Proc. Med. Image Understand. Anal*, pp. 26–30, 2007.

[160] M. Verleysen, F. Rossi, and D. François, "Advances in Feature Selection with Mutual Information," in *Similarity-Based Clustering: Recent Developments and Biomedical Applications*, Berlin, Heidelberg: Springer-Verlag, 2009, pp. 52–69. Accessed: Oct. 05, 2025. [Online]. Available: https://doi.org/10.1007/978-3-642-01805-3_4

[161] S. Liu and M. Motani, "Improving Mutual Information Based Feature Selection by Boosting Unique Relevance," *J. Artif. Int. Res.*, vol. 82, avril 2025, doi: 10.1613/jair.1.17219.

[162] D. François, F. Rossi, V. Wertz, and M. Verleysen, "Resampling methods for parameter-free and robust feature selection with mutual information," *Neurocomputing*, vol. 70, no. 7, pp. 1276–1288, Mar. 2007, doi: 10.1016/j.neucom.2006.11.019.

[163] H. Roubhi, A. Hacine-Gharbi, T. Dhieb, K. Rouabah, and P. Ravier, "A Novel Approach to Enhancing Performance in 1D-CNN-Based Speech Emotion Recognition Using Mutual Information-Based Feature Selection," *Journal of Engineering Science and Technology Review*, vol. 18, pp. 104–112, Aug. 2025, doi: 10.25103/jestr.184.15.

[164] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, *A database of German emotional speech*, vol. 5. 2005, p. 1520. doi: 10.21437/Interspeech.2005-446.

[165] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, May 2011, doi: 10.1016/j.specom.2010.08.013.

[166] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, "The HTK book," Sep. 1995. Accessed: Nov. 22, 2025. [Online]. Available: https://www.semanticscholar.org/paper/The-HTK-book-Young-Jansen/a98e481ce418a437cdfae107d85f009a5da6a790

[167] N. Aifanti, C. Papachristou, and A. Delopoulos, “The MUG facial expression database,” May 2010, pp. 1–4.

[168] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, pp. 94–101. doi: 10.1109/CVPRW.2010.5543262.

[169] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Dec. 2001, p. I–I. doi: 10.1109/CVPR.2001.990517.

[170] A. Chouchane, M. Belahcene, A. Ouamane, and S. Bourennane, “3D face recognition based on histograms of local descriptors,” in *2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Oct. 2014, pp. 1–5. doi: 10.1109/IPTA.2014.7001925.