LB/TH/30/2026
TH6279

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR SPEECH EMOTION RECOGNITION

S.S.J. Patabendige

239168X

Master of Science in Data Science and Artificial Intelligence

Department of Computer Science & Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

August 2025

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR SPEECH EMOTION RECOGNITION

S.S.J. Patabendige

239168X

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Data Science and Artificial Intelligence

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

August 2025

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 13 Oct 2025

The above candidate has carried out research for the Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Uthayasanker Thayasivam

15 Oct 2025

Signature of the Supervisor: Date:

DEDICATION

This work is dedicated to the pioneers of scientific inquiry and to all those who contribute to the advancement of knowledge. May the insights shared here inspire continued exploration, fostering progress that brings us closer to a brighter, more equitable future.

ACKNOWLEDGEMENT

The successful completion of this project owes much to the invaluable guidance and assistance provided by numerous individuals, and I am deeply grateful for their unwavering support throughout this journey.

Foremost, I extend my heartfelt appreciation to my project supervisors, Dr. Uthayasanker Thayasivam, for their indispensable contribution to this work. Their constant support, patience, motivation, and vast knowledge were instrumental in shaping the outcome of this project.

I would also like to express my gratitude to all the lecturers of the Department of Computer Science and Engineering. Their insightful knowledge, shared during the Master's program, has been a source of inspiration and intellectual growth.

ABSTRACT

The rapid advancement of Artificial Intelligence (AI) has led to its widespread adoption across multiple domains; however, the inherent opacity of conventional AI models has raised concerns regarding their interpretability and trustworthiness. In the context of Speech Emotion Recognition (SER), this lack of transparency limits understanding of how specific acoustic features influence emotional predictions. This research addresses this challenge by developing an explainable and human-centered framework for SER that prioritizes interpretability over classification accuracy.

A comprehensive feature extraction pipeline was designed to derive eighteen interpretable audio features from raw speech signals, grouped into five functional categories: Acoustic, Temporal, Spectral, Harmonic/Tonal, and Spectral Energy Distribution. Traditional machine learning models Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest, XGBoost, and k-Nearest Neighbors (k-NN) were employed to classify seven emotional states: Calm, Happy, Sad, Angry, Fearful, Surprised, and Disgust. Two model-agnostic interpretability techniques, SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), were used to quantify the contribution of each feature to model predictions.

A novel Aggregated Feature Importance (AFI) metric was introduced to combine interpretability outcomes by weighting each model's contribution according to its predictive performance. The analysis revealed noticeable divergence in feature importance rankings when aggregating across individual models, reflecting their inherent learning differences. However, as additional models were incorporated into the ensemble, these variations gradually diminished, showing a clear convergence and increased stability in AFI-based rankings. Key features such as mfcc_mean, mfcc_std, loudness, zero-crossing rate, and etc. consistently emerged as dominant predictors across emotional classes. Statistical validation using independent t-tests confirmed significant differences ($p < 0.05$) between emotional states, reinforcing the discriminative reliability of these features.

The resulting explainable feature–emotion map provides a transparent interpretation of how specific speech characteristics relate to distinct emotional states, supporting greater trust and usability in emotion-aware AI systems. This work establishes a reproducible framework for feature-level explainability in SER and offers a foundation for future research integrating reverse-engineered and temporally adaptive explainable models.

Keywords: Explainable Artificial Intelligence, Speech Emotion Recognition, Machine Learning, SHAP, LIME, Aggregated Feature Important

Table of Contents

1 Introduction 1
1.1 Background of Explainable AI for Speech Emotion Recognition 1
1.2 Research problem 2
1.3 Research Objectives 3
2 Literature review 4
2.1 Introduction to Artificial Intelligence (AI) and Explainable Artificial Intelligence (XAI) 4
2.1.1 Explainable Artificial Intelligence (XAI) 4
2.2 Value of XAI in Emotion Recognition 9
2.3 Methods of Speech Recognition 10
2.3.1 Feature Importance Analysis in Speech Emotion Recognition (SER) 10
2.3.2 Attention Mechanisms and Intrinsic Interpretation 11
2.3.3 Rule-Based Systems and Machine Learning Approaches 12
2.3.4 Model Visualization and Sonification 13
2.4 XAI in Speech-Emotional Recognition 14
2.4.1 Architectural Explainability in Speech Emotion Recognition 14
2.4.2 Embedding Explainability in Speech Emotion Recognition 15
2.4.3 Research Gaps of XAI for SER 16
3 Methodology 18
3.1 Methodology Overview 18
18
3.2 Step 1 Rational based feature extraction from dataset 18
Objective of step 1 is finding mp3 data set with labeled emotions and feature extraction from mp3 files 18
3.2.1 Mp3 dataset with labeled emotions 18
3.2.2 Dataset Overview 19
3.2.3 Emotional Expressions 19
3.2.4 Filename Structure 19
3.2.5 Licensing and Attribution 19
3.2.6 Feature identification and extraction from mp3 files 20
3.2.7 Rationale for Feature Selection 20

3.2.8 Why These Features Are Relevant for XAI in Voice Recognition..... 23
3.3 Step 2. Model training voice emotional recognize..... 24
3.3.1 Model Selection and Rationale..... 24
3.3.2 Data Preprocessing and Separation..... 24
3.3.3 Model Descriptions and Justification..... 24
3.3.4 Hyper parameter Tuning..... 27
3.4 Step 3. XAI for voice emotional recognize..... 27
3.4.1 Method for Selecting Feature Importance and Analyzing Value Ranges 27
3.4.2 Feature Importance Analysis..... 27
3.4.3 Instance Sampling and Explainability..... 28
3.5 Step 4. Aggregated Feature Importance (AFI) Explanation..... 28
3.6 Step 5. Value Range Analysis..... 30
3.6.1 Descriptive Statistical Analysis..... 30
3.6.2 Statistical Significance Testing with One-Sample t-Test..... 30
4 Results and Analysis..... 33
4.1 Each feature's distribution..... 33
4.2 Selected models with best parameters based on selected parameter grid ... 35
4.3 Each mode accuracy compression..... 36
4.4 Feature Importance Results for ANN, SVM, XGB, RF, KNN with LIME and SHAP..... 40
4.6 Aggregated Feature Importance (AFI) Explanation..... 45
4.7 Top five feature for each emotions..... 49
4.8 Value ranges for best features..... 50
5 Discussion..... 52
6 Conclusion and Future Work..... 55
8 References..... 57

List of tables

TABLE 2.1 Architectural explainability in speech emotion recognition..... 14

TABLE 2.2 Embedding explainability in speech emotion recognition 15

TABLE 4.4.1: Best hyperparamter and validation results for each models..... 35

TABLE 4.4.2: Precision, recall, f1, support value each class for random forest model 36

TABLE 4.4.3: Precision, recall, f1, support value each class for svm model..... 37

TABLE 4.4.4: Precision, recall, f1, support value each class for xgboost model..... 37

TABLE 4.4.5: Precision, recall, f1, support value each class for knn model 38

TABLE 4.4.6: Precision, recall, f1, support value each class for ann model..... 38

TABLE 4.4.7:Feature importance value for each emotions for m = 1..... 45

TABLE 4.4.8:Feature importance value for each emotions for m = 2..... 46

TABLE 4.4.9:Feature importance value for each emotions for m = 3..... 46

TABLE 4.4.10:Feature importance value for each emotions for m = 4..... 47

TABLE 4.4.11:Feature importance value for each emotions for m = 5..... 47

TABLE 4.4.12: Top five features for each emotion after applying afi with m=5 49

TABLE 4.4.13: Value ranges(min, max, mean, std) and t-test value for top five importance features 50

List of figures

Figure 2.2.1: Process of XAI 5
Figure 2.2.2 : XAI on different application domains with percentage.....6
Figure 2.2.3 : XAI adds new dimensions to AI by answering "wh" questions 7
Figure 3.1 :Overall workflow of Methodology.....18
Figure 4.4.1: Distribution of loudness (left) and speaking_rate (write) by Emotions 33
Figure 4.4.2: Distribution of chroma_mean (left) and mfcc_mean (write) by Emotions 33
Figure 4.4.3: Distribution of mfcc_std (left) and spectral_contrast (write) by Emotions 33
Figure 4.4.4: Distribution of zcr (left) and average_pitch (write) by Emotions 34
Figure 4.4.5: Distribution of chroma_std (left) and harmonic_mean (write) by Emotions ...34
Figure 4.4.6: Distribution of percussive_mean (left) and average_pitch (write) by Emotions 34
Figure 4.4.7: Distribution of tempo (left) and mfcc_mean (write) by Emotions 34
Figure 4.4.8: Distribution of spectral_rolloff_std (left) and spectral_rolloff_mean (write) .35
Figure 4.4.9: Distribution of tempo (left) and log_mel_spec_mean (write) by Emotions.....35
Figure 4.4.10:ANN Training & Validation Loss 36
Figure 4.4.11: ANN Training & Validation Accuracy 36
Figure 4.4.12:Precision, Recall and F1 score comparisons 39
Figure 4.4.13: Feature Importance Results for ANN with LIME 40
Figure 4.4.14: Feature Importance Results for ANN with SHAP.....40
Figure 4.4.15: Feature Importance Results for SVM with LIME 41
Figure 4.4.16: Feature Importance Results for SVM with SHAP.....41
Figure 4.4.17: Feature Importance Results for XGBoost with LIME.....42
Figure 4.4.18: Feature Importance Results for XGBoost with SHAP 42
Figure 4.4.19: Feature Importance Results for Random Forest with LIME 43
Figure 4.4.20: Feature Importance Results for Random Forest with SHAP.....43
Figure 4.4.21: Feature Importance Results for KNN with LIME 44
Figure 4.4.22: Feature Importance Results for KNN with SHAP.....44
Figure 4.4.23: Feature impotence value for each emotions after M = 5 with AFI equation...48
Figure 4.4.24: This visualization illustrates the convergence of feature importance values computed using the Aggregated Feature Importance (AFI) equation, as the number of models increases from M = 1 to M = 5. 48
Figure 4.4.25: Global feature importance for emotional recognize 49

List of abbreviations

Abbreviation	Description
XAI	Explainable Artificial Intelligence
SER	Speech Emotion Recognition
SVM	Support Vector Machine
ANN	Artificial Neural Network
RF	Random Forest
XGB	XGBoost
KNN	k-Nearest Neighbors
AFI	Aggregated Feature Importance
SHAP	SHapley Additive explanations
LIME	Local Interpretable Model-agnostic Explanations
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
ERNN	Enhanced Recurrent Neural Network

CHAPTER 1

INTRODUCTION

1.1 Background of Explainable AI for Speech Emotion Recognition

In recent years, the pervasive influence of Artificial Intelligence (AI) has extended into nearly every domain of human activity [5], demonstrating its capabilities across sectors such as healthcare, defense, and governance [1], [6]. Despite this widespread integration, the intrinsic opaqueness of AI models particularly in tasks involving Speech Emotion Recognition (SER) has raised critical concerns regarding transparency, accountability, and user trust [1]. Traditional AI approaches, often characterized as black-box models, fail to provide comprehensible explanations for their predictions and decisions, especially when deployed in emotionally sensitive or decision-critical contexts [7].

To mitigate these concerns, the framework of Explainable Artificial Intelligence (XAI) has been introduced, emphasizing the need for interpretable and transparent decision-making processes within AI systems [1], [4]. This study is situated within the field of SER, a domain where the ability to understand the reasoning behind AI-generated predictions is essential to user acceptance and system reliability [8]. As AI-powered systems such as virtual assistants like Alexa and Siri become increasingly embedded in daily life, largely depending on audio-based classification and interaction, the demand for explainable and trustworthy emotion recognition mechanisms continues to grow [1], [5].

The primary aim of this research is not to maximize classification performance, but rather to develop a structured and explainable feature matrix that maps acoustic characteristics of speech to recognized emotional states. Accordingly, deep learning models have been intentionally excluded in favor of traditional machine learning algorithms namely Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest, XGBoost, and k-Nearest Neighbors (k-NN) which are better suited for generating interpretable insights from tabular data formats [2], [3]. These models allow for a detailed exploration of how specific acoustic features contribute to emotion classification, supporting the development of a transparent, human-centered SER framework.

The methodology adopted in this work follows a structured progression: it begins with a comprehensive Literature Review to establish a clear understanding of the current landscape of XAI in SER, particularly focusing on voice embeddings and interpretable models [9]. This is followed by the development of an explainable matrix based on carefully selected audio features, the application of model-agnostic XAI techniques for interpretability enhancement, and the definition of evaluation metrics such as the proposed

Value of Feature Range's Metric to assess both model behavior and explanation quality [2], [3].

In doing so, this research seeks to bridge the gap between AI decision-making and human comprehension, offering valuable contributions at the intersection of AI, speech processing, and human-computer interaction [8]. The introduction of a feature-based value matrix, grounded in human-centered interpretability, aims to foster greater transparency in voice-based systems and pave the way for future explainable models that support reverse-engineering and sequential analysis for time-dependent SER tasks.

1.2 Research problem

The research problem we aim to address is the lack of human-centeredness and transparency in existing explainable AI (XAI) techniques for speech emotion recognition. Even while deep learning models in particular, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) and others have shown to be successful at identifying speech emotions, they frequently function as "black-box" models, making it difficult to comprehend how and why they generate particular predictions. Their practical utility is hindered by their opacity, particularly in applications that require interpretability and trustworthiness, such mental health monitoring and medical diagnosis. In the domain of speech emotion recognition, the existing XAI systems are unable to give concise, understandable explanations for their conclusions and have not fully taken into account human-centered explanations.

1.3 Research Objectives

- To critically review existing literature on Explainable Artificial Intelligence (XAI) with a specific focus on voice embedding techniques and their applicability to Speech Emotion Recognition (SER).

- To design and develop a human-centered feature extraction and model training framework that captures and explains key acoustic and spectral characteristics contributing to emotion recognition in voice data.

- To apply and evaluate XAI methods such as SHAP and LIME to identify and interpret the most influential features for each emotion category, and to define corresponding value ranges that enhance the transparency and interpretability of SER models.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to Artificial Intelligence (AI) and Explainable Artificial Intelligence (XAI)

Artificial Intelligence (AI) is an advanced paradigm in which systems or machines seek to emulate human intelligence, performing diverse tasks in the physical world. This cutting-edge technology enables systems to undergo training with data, facilitating thinking, learning, and deriving insights from experiences to adeptly tackle specific challenges. Through heuristic refinement based on the utilized data, the system can iteratively enhance its capabilities [1].

The applications of AI are manifold, encompassing advanced web search engines, automated driving cars, gaming, human speech recognition, recommendation systems, healthcare, and more. Originating around 1950 in computer science, AI aimed to replicate the complexities of the human mind, resulting in the development of machines capable of processing, organizing, and executing tasks based on provided datasets [2].

AI machinery has found extensive use in industrial domains, sparking comprehensive research in engineering fields such as DL, ML, Natural Language Processing (NLP), disease diagnosis and medicine, and scientific applications. The capacity of AI machines to assimilate knowledge from past experiences has proven crucial in problem-solving, contributing to enhanced performance across diverse application domains [3].

The application of Artificial Intelligence (AI) algorithms has empowered users to make business decisions. However, the opacity of AI-generated outputs and decision-making processes poses a challenge for users lacking insight into these aspects. To address this issue, Explainable Artificial Intelligence (XAI) has been introduced [2].

2.1.1 Explainable Artificial Intelligence (XAI)

Explainable AI (XAI) elucidates the internal workings of a model by providing clear explanations of methods, procedures, and output processes in a manner comprehensible to users. Coined by the Defense Advanced Research Project Agency (DARPA), XAI is often referred to as a "White box" due to its emphasis on transparently explaining the model's processes [2].

The input for training data is provided, and depending on specific requirements or application domains, users can select the methodology for prediction, incorporating XAI

techniques to clarify the model's inner workings. This explanation interface enhances user understanding of the AI model's output, fostering trust. User awareness of the output allows for improvements in accuracy and identification of model flaws, aiding informed decision-making for model enhancement. Process of XAI is give this **Figure 1**

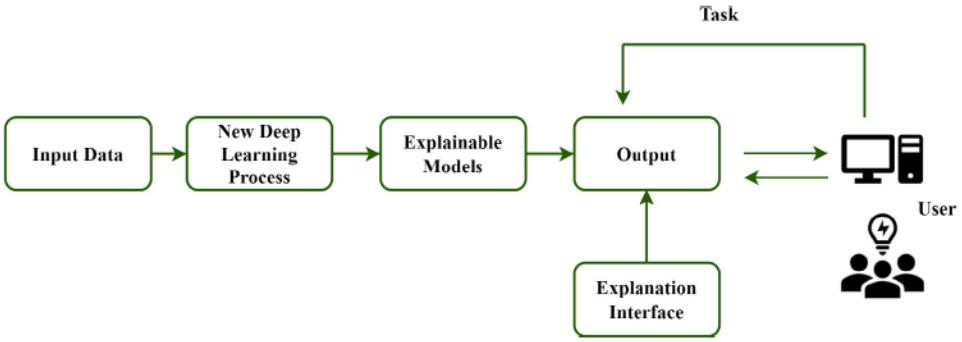


Figure 2.2.1: Process of XAI

Explainable Artificial Intelligence (XAI) has become increasingly important in sectors where decision transparency is critical, such as healthcare and signal processing. Generally, XAI techniques are divided into two major categories: knowledge-driven and data-driven approaches [13]. Knowledge-driven methods depend on established domain knowledge to interpret outcomes, whereas data-driven techniques focus on generating explanations through local, global, or instance-specific analysis of the model’s behavior [13].

Researchers have further organized XAI strategies according to the breadth and scope of the explanations they provide [14]. Some models, like linear regression, are naturally interpretable and fall under the category of "intrinsic models" because of their inherent simplicity (e.g., Logic Analysis of Data (LAD)). On the other hand, more complex models that are not self-explanatory require "post-hoc" interpretability methods. These include techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive explanations (SHAP), as well as approaches categorized as model-agnostic, model-specific, class activation maps (CAM), layer-wise relevance propagation (LRP), gradient-weighted CAM, cluster-based methods, filter-based approaches, attention mechanisms, rule-based systems, knowledge-based frameworks, interpretable models, auto encoder-based models, and tree-based models [14].

Over the past few years, AI technologies have deeply influenced nearly every area of human activity, from scientific research and business operations to finance and social media interactions [15]. AI algorithms have demonstrated success across various data types such as text, images, audio, and video enabling innovative applications in healthcare,

defense, law enforcement, governance, and autonomous systems. These algorithms are capable of solving a wide range of tasks, including classification, regression, clustering, transfer learning, and optimization problems [15], [11], [1]. Figure 2 provides a pie chart representation illustrating the distribution of XAI applications across different domains and their respective percentages.

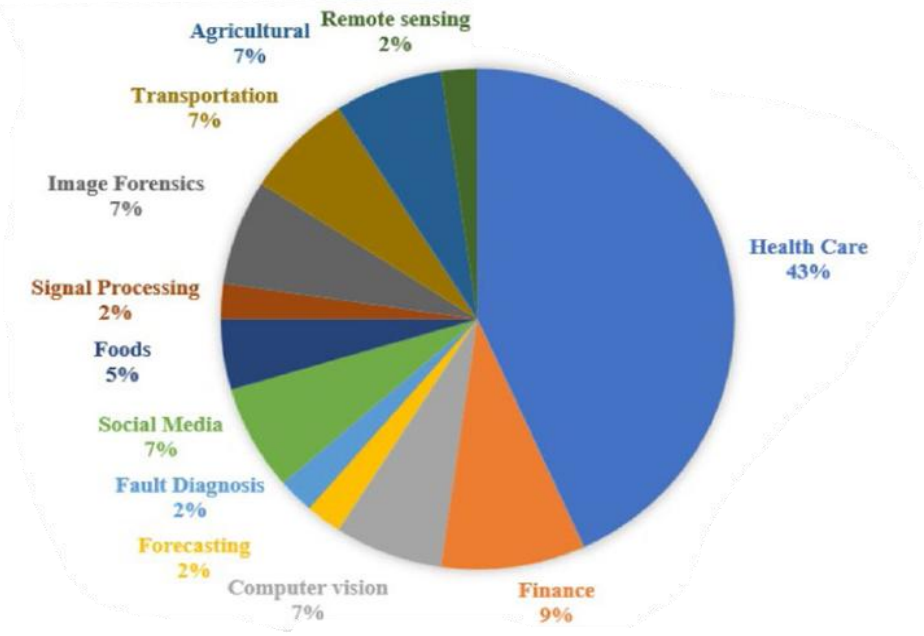


Figure 2.2.2 : XAI on different application domains with percentage

Modern developments in artificial intelligence (AI) are largely centered around a subfield known as machine learning (ML). In ML, a computer is exposed to a collection of examples—referred to as a training dataset—from which it learns patterns and relationships. After sufficient training, the system can respond to questions based on the knowledge it has acquired. Nevertheless, traditional AI models often operate like "black boxes," offering simple "yes" or "no" responses without providing any insight into how those decisions were made [1].

In many real-world scenarios, however, understanding the reasoning behind an AI-generated answer is just as important as the answer itself. For instance, in healthcare, it is not enough for an AI system to diagnose a disease from a CT scan; it is critical for practitioners to understand how that conclusion was reached. Such transparency builds trust and can help avoid potentially life-threatening mistakes. Similarly, in fields like law enforcement, the ability to answer questions such as "why," "what" and "how" is often necessary, something that traditional AI systems typically cannot accommodate [1].

This growing demand for interpretable and transparent AI systems has led to the emergence of a dedicated research area known as Explainable AI (XAI). As depicted in Figure 3, XAI extends the capabilities of traditional AI by addressing questions that were previously overlooked. The importance of XAI has been increasingly recognized across critical sectors such as healthcare, defense, and law enforcement, where understanding the basis of a decision is essential. Consequently, XAI has become a major focus of ongoing research and development efforts within both academic and industrial communities, with much work still needed to fully unlock its potential [1].

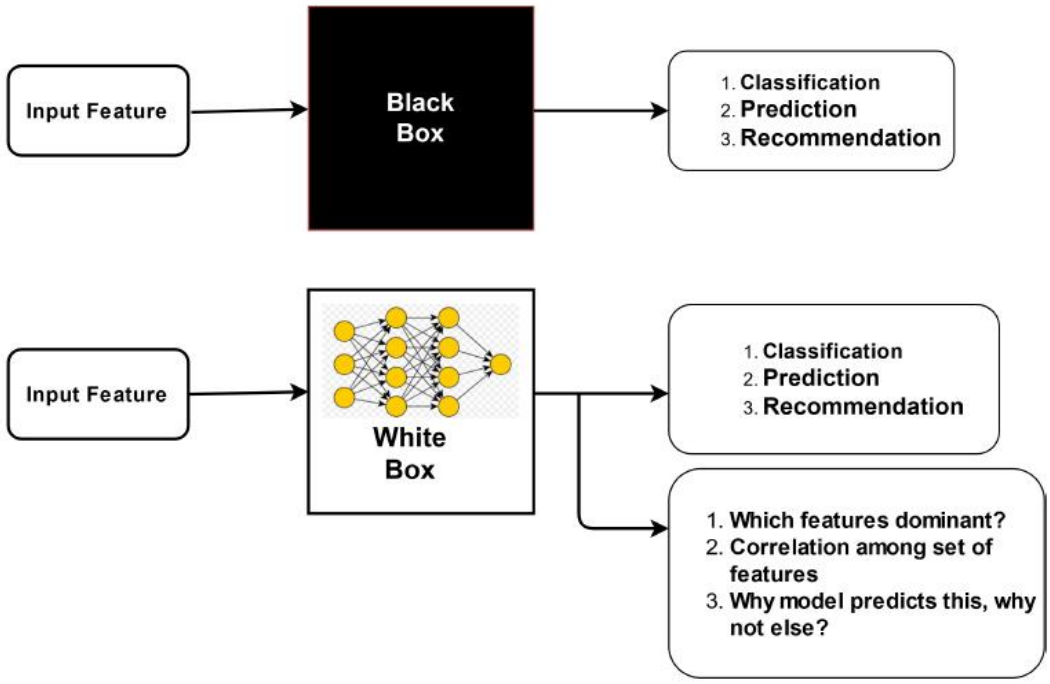


Figure 2.2.3 : XAI adds new dimensions to AI by answering questions

In the intricate landscape of Explainable AI (XAI), a multitude of imperatives underscores its profound significance and versatile applications. One crucial facet is the imperative to 'Explain to Justify,' where XAI becomes indispensable in elucidating results, especially in the context of real-world critical situations [4]. This explanatory function takes on heightened importance under the umbrella of the 'Right to Explanation' guidelines within the General Data Protection Regulation (GDPR), playing a pivotal role in delivering pertinent information and justifying outcomes. Beyond justification, the imperative to 'Explain to Control' emerges as paramount, involving not only the articulation of explanations but also the meticulous oversight to prevent the narrative from going astray. This entails identifying vulnerabilities and flaws, ensuring that the explanation remains

accurate, and does not deviate or mislead. Another essential imperative is 'Explain to Improve,' emphasizing the perpetual enhancement of models by understanding their inherent flaws. This continuous improvement is achieved by leveraging the insights provided by explainable models, shedding light on areas that require refinement and optimization [2].

Furthermore, the imperative to 'Explain to Discover' unveils a dimension of XAI focused on knowledge discovery. Beyond justification and improvement, XAI is instrumental in uncovering new insights and facts relevant to complex problems. By gathering and processing information, Explainable AI contributes to a deeper understanding of the intricacies surrounding various issues, fostering a proactive and informed approach to problem-solving. Moreover, the systematic categorization of explanations into three dimensions source, scope, and depth adds a structured layer to the understanding of the explanatory process [4].

Delving into the foundational principles of Explainable AI, the National Institute of Standards and Technology (NIST) has identified four fundamental tenets that govern its implementation [5]. The principle of 'Explanation' mandates that the system provides articulate insights into its processes and outcomes, elucidating the actions, design, and workflow involved [2]. The nature of the explanation is deemed task-dependent, adapting to the specific requirements of the undertaken task. 'Meaningful' explanations, as per another principle, should be interpretable to stakeholders, ensuring that the conveyed information is both relevant and comprehensible. 'Explanation Accuracy' introduces a stringent standard, emphasizing the need for accurate and high-quality explanations that align with the actual processes of the system [5]. Lastly, the principle of 'Knowledge Limits' mandates that systems recognize situations beyond their intended functionality or approval, thereby safeguarding against inaccurate, unsafe, or biased outcomes. By adhering to these principles, Explainable AI not only enhances transparency and accountability but also elevates the overall trust in AI systems by preventing unintended consequences and ensuring a thorough understanding of their functioning [2].

2.2 Value of XAI in Emotion Recognition

Explainable Artificial Intelligence (XAI) has become a crucial component in the domain of Emotion Recognition, which is vital for understanding human behavior and enhancing human-computer interactions. With advancements in Artificial Intelligence (AI) techniques, the accurate detection and classification of emotions from audio signals has gained increasing attention. Researchers and developers are utilizing AI algorithms to design sophisticated models capable of interpreting emotional cues embedded within audio data [18].

Recognizing emotions from audio information has a far-reaching impact across multiple fields. In healthcare, emotion recognition can assist in detecting emotional distress or diagnosing mental health issues. Within the entertainment sector, it enables personalized experiences by adapting content based on a user's emotional responses. In customer service, it helps gauge customer satisfaction levels, leading to improvements in service delivery. These applications illustrate the significant role that emotion recognition and classification play in developing human-centric technologies [18],[19],[20].

Traditionally, emotion recognition relied heavily on manual interpretation or subjective self-reports, both of which were prone to biases and inefficiencies. However, advancements in machine learning and deep learning have dramatically transformed the landscape. Techniques such as Support Vector Machines (SVM), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN) have demonstrated substantial success in reliably identifying and categorizing emotions from audio signals [21].

The emotion recognition process from audio data typically involves several key stages. Initially, important features such as pitch, intensity, and spectral attributes are extracted from audio recordings to capture emotional characteristics. These features are then used to train AI models, enabling the system to learn patterns linking audio features with specific emotions. Model evaluation is performed using labeled datasets where human annotators have pre-assigned emotional labels to each audio sample [21].

Despite considerable progress, certain challenges remain in the audio-based emotion recognition field. Subjectivity in emotion labeling and variations among human annotators can cause inconsistencies in training datasets. To address these limitations, emerging research is exploring multimodal approaches that combine audio and visual cues to enhance the accuracy and reliability of emotion recognition systems [22].

This article presents a comprehensive overview of AI-driven innovations in emotion recognition and classification based on audio data. It discusses the underlying AI

algorithms, techniques for feature extraction, methods for model training and evaluation, and real-world applications spanning healthcare, entertainment, and customer service. By understanding the current advancements and future directions in this field, researchers and practitioners can effectively leverage AI to build advanced emotion recognition systems, strengthening human-computer interactions and benefiting a wide range of industries [18],[19],[20],[21],[22].

2.3 Methods of Speech Recognition

This section explores various methods used to enhance the transparency and interpretability of Speech Emotion Recognition (SER) systems. It highlights techniques such as feature importance analysis, attention mechanisms, rule-based systems, and model visualization approaches that contribute to explainable SER models.

2.3.1 Feature Importance Analysis in Speech Emotion Recognition (SER)

Feature importance analysis plays a pivotal role in explainable Artificial Intelligence (XAI) applied to Speech Emotion Recognition (SER). This process involves identifying and emphasizing the most crucial features that significantly influence the predictions made by SER models. By unraveling the contribution of specific acoustic features, techniques like SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations) have emerged as powerful tools for enhancing interpretability in SER models [6],[7].

One prominent technique for feature importance analysis is SHAP, which provides a unified and theoretically grounded approach to interpreting model predictions [6] Developed by Lundberg and Lee in 2017, SHAP values assign a numerical measure to the contribution of each feature in the prediction, allowing for a detailed understanding of the impact of individual features on the final output. This not only aids in making predictions more interpretable but also helps in gaining insights into the underlying decision-making process of the neural network [6].

Another noteworthy technique is LIME, which focuses on providing local, interpretable explanations for model predictions [6],[8]. LIME generates perturbations in the input data and observes the corresponding changes in predictions, facilitating the identification of key features affecting specific instances. This approach is particularly beneficial in scenarios where a fine-grained understanding of model behavior is required at the instance level, contributing to the overall transparency of the SER model [6].

The reference to Lundberg and Lee's work [6] underscores the academic foundation of these techniques. In their paper, "A Unified Approach to Interpreting Model Predictions" (published in Advances in Neural Information Processing Systems, 2017), the authors

present a comprehensive overview of SHAP and its applications, emphasizing its potential in fostering interpretability across various machine learning domains, including SER.

Additionally, Caruana et al. [7] have contributed to the field of interpretable machine learning with their paper "Intelligible Models for Classification and Regression" by proposing an approach that enhances model transparency, making it more accessible for users.

Ribeiro et al. [8] introduced LIME with their paper "Why Should I Trust You? Explaining the Predictions of Any Classifier," providing valuable insights into local interpretability.

The adoption of feature importance analysis techniques such as SHAP and LIME in SER not only addresses the black-box nature of neural networks but also enhances model transparency. These approaches contribute to a more intuitive comprehension of SER model outputs, promoting trust and facilitating informed decision-making in applications related to emotion recognition in speech.

2.3.2 Attention Mechanisms and Intrinsic Interpretation

In the intricate landscape of Speech Emotion Recognition (SER), attention mechanisms in neural networks emerge as powerful tools contributing to model interpretability [9]. These mechanisms allow the model to focus on specific segments of the input sequence, unveiling the acoustic features crucial for accurate emotion recognition. This dual effect not only enhances the transparency of the model but also equips users with actionable insights into the elements shaping the decision-making process [9].

Recent research has emphasized the importance of attention mechanisms in enhancing interpretability in various domains. For instance, Xu et al. (2015) demonstrated the effectiveness of attention mechanisms in image captioning, where the model learned to focus on relevant regions of an image when generating textual descriptions [10]. This broader application showcases the versatility and generalizability of attention mechanisms across different modalities.

Furthermore, the work of Vaswani et al. (2017) on the Transformer architecture, which heavily relies on attention mechanisms, has become a cornerstone in natural language processing tasks [11]. The Transformer's success has sparked interest in understanding and applying attention mechanisms in various fields, including speech-related tasks such as SER.

Choi, Bahadori, and Schuetz's Retain model (2016) stands out as an example of the application of attention mechanisms in healthcare, using a reverse time attention mechanism for interpretable predictive modeling [9]. This work underscores the transferability of attention mechanisms and their capacity to provide interpretable insights across different domains.

The integration of attention mechanisms in SER not only aligns with broader trends in machine learning and neural network architectures but also draws from a rich body of research spanning multiple domains, further validating their utility in enhancing model interpretability and providing actionable insights into decision-making processes.

2.3.3 Rule-Based Systems and Machine Learning Approaches

Rule-based systems provide a structured and interpretable approach to Speech Emotion Recognition (SER), relying on explicit rules connecting input features to emotion labels. These rules not only yield human-readable outcomes but also offer transparency into the decision logic of SER models, enhancing user understanding and confidence in decision-making processes [12].

Dey and Ashour (2018) demonstrated the effectiveness of rule-based systems in SER by developing a rule-based expert system using lexical features. Their work exemplifies the interpretability and transparency achieved through rule-based approaches, contributing valuable insights to the domain of emotion recognition [12].

In a similar vein, Lopes et al. (2017) explored rule-based approaches in SER by introducing a system based on fuzzy logic and rule induction. This study provided further evidence of the interpretability and transparency benefits of rule-based systems in the context of speech emotion analysis [13].

Moreover, a study by Eyben et al. (2010) emphasized the utility of rule-based systems in SER by introducing a rule-based feature selection method that improved the efficiency of emotion recognition systems [14].

This emphasis on rule-based systems underscores the ongoing exploration and refinement of interpretable approaches in SER, offering valuable tools for understanding the decision-making processes involved in emotion recognition from speech.

While rule-based systems provide interpretability, machine learning approaches, such as ensemble methods, have also gained prominence in SER. Bagging and boosting techniques, as demonstrated by Kim et al. (2017), can combine the strengths of multiple models to enhance overall performance in emotion recognition tasks, while maintaining a degree of transparency in the model's decision process [15].

Kim et al. (2017) explored the application of ensemble methods in SER, specifically focusing on bagging and boosting techniques. Their findings highlighted the potential of these approaches to improve emotion recognition accuracy while preserving interpretability, offering an alternative perspective to the rule-based systems [15].

The integration of machine learning approaches with interpretability measures adds depth to the discourse surrounding SER, showcasing the ongoing exploration of diverse

methodologies to achieve a balance between accuracy and transparency in emotion recognition models.

2.3.4 Model Visualization and Sonification

In the pursuit of enhancing the explainability of Speech Emotion Recognition (SER) models, novel techniques such as model visualization and sonification have emerged as promising approaches. These strategies aim to make the decision-making process of SER models more understandable by providing users with intuitive representations of the internal workings of the model.

Zhang and Wang (2020) investigated the role of model visualization in SER, specifically focusing on Convolutional Neural Network (CNN)-based models. By visualizing the intermediate layers and filters of the model, their work demonstrated how these visualizations could offer insights into the features and patterns that the model deems crucial for emotion recognition [16].

The concept of sonification, as an auditory display of data, has also gained traction in the context of SER interpretability. Transforming complex model information into sound enables users to perceive patterns and nuances that may be challenging to discern through visual means alone. This auditory representation serves as an additional channel for users to gain insights into how the model interprets emotional cues within speech signals.

While Zhang and Wang (2020) provided a visual-focused approach, a study by Vogt et al. (2017) explored the use of sonification in SER. Their research demonstrated the effectiveness of using auditory feedback to convey emotion-related features, providing an alternative and complementary method to traditional visualizations [17].

Additionally, the work of Xu et al. (2019) delved into the visualization of attention mechanisms in SER models. By visualizing the attention weights, their study elucidated how the model allocates focus to different segments of the input, contributing to the interpretability of the decision-making process [18].

These studies collectively highlight the significance of model visualization and sonification in making SER models more interpretable, providing users with valuable insights into the features and patterns that underlie emotion recognition decisions.

2.4 XAI in Speech-Emotional Recognition

Explainable Artificial Intelligence (XAI) has become essential in Speech Emotion Recognition (SER) systems due to the intrinsic complexity and black-box nature of many deep learning architectures used in this domain. SER involves decoding subtle emotional cues from speech signals, which relies heavily on advanced neural network models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers. While these architectures deliver state-of-the-art accuracy, their layered, non-linear computations are often opaque, limiting trust and interpretability in critical sectors such as healthcare and law enforcement [19], [20], [21].

2.4.1 Architectural Explainability in Speech Emotion Recognition

Architectural explainability involves unpacking the internal workings of neural network components and layers to provide insights into how emotional information is processed. Key methods include

TABLE 2.1 ARCHITECTURAL EXPLAINABILITY IN SPEECH EMOTION RECOGNITION

Methods	Description	References
Attention Mechanisms	Transformer and RNN-based SER systems use multi-head self-attention layers, which compute context-dependent weights for each frame in the speech signal. The resulting attention matrices ($A \in R^{T \times T}$) quantify how much each segment of the input influences others. Visualization of these matrices (e.g., as temporal heat maps) reveals which phonemes, syllables, or acoustic events the network focuses on during emotion classification, allowing researchers and practitioners to confirm whether the model aligns with known emotional cues.	[22]
Iterative Feature Boosting with SHAP	Employs an iterative framework integrating SHAP values to explain and refine feature boosting in SER, iteratively selecting relevant feature combinations to enhance model accuracy and transparency using a feedback loop approach.	[23]

Class Activation Maps (CAM/Grad-CAM)	Grad-CAM leverages gradient information with respect to target emotion classes, projecting them onto intermediate feature maps (e.g., spectrograms). The resulting spatial maps highlight which time-frequency bins strongly activate network responses for each emotional category, making it possible to audit model attention in scenarios such as "anger" being strongly associated with high-energy regions or "sadness" with low frequency.	[24]

2.4.2 Embedding Explainability in Speech Emotion Recognition

Deep architectures for SER learn high-dimensional representations (embedding) capturing speech prosody, phonetics, and emotion. Explaining these embedding is an emerging technical priority.

TABLE 2.2 EMBEDDING EXPLAINABILITY IN SPEECH EMOTION RECOGNITION

Methods	Description	References
Embedding Visualization and Probing	High-dimensional representation vectors ($E \in R^{T\times d}$) are visualized using techniques such as t-SNE or UMAP, which facilitate the understanding of how different emotions cluster within the latent space. Probing involves training lightweight regresses to map embedding dimensions to interpretable features like pitch, intensity, or MFCCs, directly associating model internals with acoustic properties.	[25]
Layer Attribution in Multi-layer Embedding	Transformer encoders generate embedding at each network layer, and analyzing these through attribution metrics reveals which layers capture the most relevant acoustic or high-level information for emotion discrimination. This helps optimize model interpretability while maintaining predictive power.	[20]
Interpretable Embedding	Recent state-of-the-art speech emotion recognition models not only focus on predictive accuracy but also emphasize embedding	

Architectures and Regularization	interpretability by introducing auxiliary losses and sparsity-inducing regularization techniques. These approaches encourage learned embedding to correlate with annotated, interpretable acoustic cues, such as pitch, loudness, and prosodic variation, thereby enhancing model transparency and end-user trust. For example, models impose sparsity penaltie ($L_{reg} = \lambda \sum_i \lvert E_i \rvert$) on embedding dimensions, promoting feature disentanglement and interpretability. This aligns well with human-like perceptual processing, facilitating explanations that are more relatable to domain experts and lay users alike. The RexNet architecture embodies this principle, employing modular explanation networks that explicitly relate embedding vectors to contrastive vocal cues, improving both accuracy and explainability. User studies confirm that such interpretable embedding and semantically meaningful cue explanations significantly aid human understanding and decision-making in vocal emotion recognition tasks	[26],[24]

The collective application of architectural and embedding explainability tools including attention visualization, LRP, Grad-CAM, LIME, and SHAP enables researchers and practitioners to trace the SER model's reasoning from initial waveform processing through deep feature extraction, culminating in final emotion predictions. Such pipelines are critical for performance validation, bias detection, and ensuring system transparency requirements that are increasingly important in medical, social, and interactive applications of SER

2.4.3 Research Gaps of XAI for SER

Despite growing interest in Explainable AI for Speech Emotion Recognition, significant challenges remain in bridging the gap between technical model explanations and human-understandable insights. Addressing these gaps is crucial for advancing trustworthy, transparent, and practical SER systems.

2.4.3.1 Limited Human-Centered Explainability

Many existing SER systems emphasize classification accuracy, often employing deep learning models like CNNs and RNNs that act as black boxes, offering little insight into why specific emotional predictions are made. While some studies propose post-hoc explanation methods such as SHAP or saliency maps, these explanations are often technical and not sufficiently aligned with human cognitive understanding. There is a lack of frameworks that provide intuitive, relatable, and human-centered explanations linking acoustic features directly to emotional states.

2.4.3.2 Lack of Numerical Feature Range Mapping

While some recent studies have introduced ordinal feature matrices that categorize emotions based on general trends in acoustic features, they lack precise numerical feature range mappings**.** Most existing works do not define the specific value intervals of acoustic features associated with different emotional states. This absence of structured, quantitative mapping limits the development of transparent, rule-based systems that offer clear and interpretable explanations. Without such mappings, it becomes difficult for users and domain experts to understand or verify model decisions, thereby reducing trust, usability, and applicability in sensitive domains where interpretability is critical.

2.4.3.3 Underutilization of Traditional Machine Learning for Interpretability

While deep learning models dominate SER research due to their high accuracy, traditional machine learning algorithms (e.g., Random Forest, SVM, XGBoost) have shown greater potential for explainability through explicit feature importance and decision rules. However, existing research rarely leverages these interpretable models to develop structured, explainable SER frameworks that balance performance and transparency.

2.4.3.4 Insufficient Integration of Model-Agnostic XAI with Feature Engineering

Current XAI methods applied in SER largely focus on post-hoc explanations without deeply integrating feature engineering and explanation techniques. There is a gap in research that combines carefully extracted, rational acoustic features with model-agnostic XAI techniques to enhance interpretability from the ground up, rather than as an afterthought

CHAPTER 3

METHODOLOGY

3.1 Methodology Overview

This research follows a structured methodology starting with an emotional speech dataset selection and rational-based feature extraction. Interpretable machine learning models are then trained for speech emotion recognition. Explainable AI techniques (SHAP and LIME) are applied to evaluate feature importance, which is used to develop a human-sensitive value matrix. Statistical significance testing using the t-test is conducted to validate the results. The figure below illustrates the overall workflow.

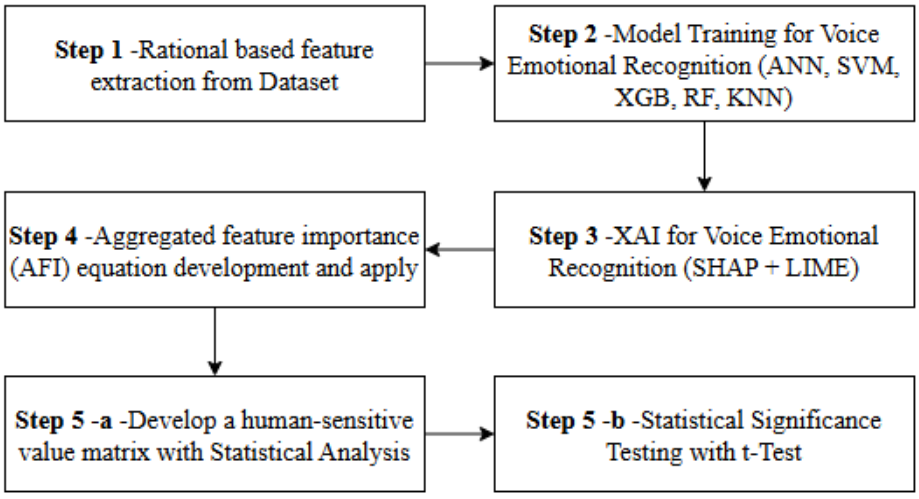


Figure 3.1 :Overall workflow of Methodology

3.2 Step 1 Rational based feature extraction from dataset

Objective of step 1 is finding mp3 data set with labeled emotions and feature extraction from mp3 files

3.2.1 Mp3 dataset with labeled emotions

RAVDESS dataset (Livingstone & Russo, 2018) was used in my research with 1440 mp3 files with different voice labels. Dataset Description -Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [27].

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a comprehensive dataset designed to support research on emotional speech and song recognition. It is widely used in the fields of speech processing, affective computing, and human-computer interaction. This dataset provides high-quality audio files that capture a range of emotional expressions delivered by professional actors.

3.2.2 Dataset Overview

The portion of the RAVDESS used in this research comprises 1440 audio-only files (.wav format, 16-bit, 48kHz). These files are organized as follows

- 60 trials per actor x 24 actors = 1440 recordings
- The dataset features 24 professional actors (12 males, 12 female) delivering two lexically-matched statements in a neutral North American accent.

3.2.3 Emotional Expressions

The RAVDESS includes a variety of emotional states to facilitate robust emotional recognition analysis. The emotional expressions covered are

- Calm, Happy, Sad, Angry, Fearful, Surprise, and Disgust
- Each emotion is produced at two levels of intensity (Normal and Strong), with the exception of the Neutral emotion, which is presented only at normal intensity.

3.2.4 Filename Structure

Each file in the dataset is uniquely named using a 7-part numerical identifier that encodes key stimulus characteristics. For example, the filename 03-01-06-01-02-01-12.wav can be decoded as follows

- Modality: 03 (Audio-only)
- Vocal Channel: 01 (Speech)
- Emotion: 06 (Fearful)
- Emotional Intensity: 01 (Normal)
- Statement: 02 ("Dogs are sitting by the door")
- Repetition: 01 (1st Repetition)
- Actor: 12 (Female, as even numbers represent female actors)

3.2.5 Licensing and Attribution

The RAVDESS is licensed under CC BY-NA-SC 4.0, allowing its use in academic research and non-commercial projects with appropriate attribution.

This dataset's structured emotional representation and high-quality audio recordings make it an essential resource for advancing research in emotion detection and affective computing.

3.2.6 Feature identification and extraction from mp3 files

Selected features - 'loudness', 'zcr','tempo', 'speaking_rate', 'average_pitch', 'spectral_contrast', 'mfcc_mean', 'mfcc_std', 'chroma_mean', 'chroma_std', 'harmonic_mean', 'percussive_mean', 'spectral_rolloff_mean', 'spectral_rolloff_std', 'log_mel_spec_mean', 'log_mel_spec_std', 'log_mel_spec_min', 'log_mel_spec_max'

3.2.7 Rationale for Feature Selection

The selected features are carefully chosen to capture a wide range of acoustic, temporal, and spectral characteristics of the audio signal. These features are widely used in speech and audio processing tasks and are particularly relevant for voice recognition and emotion analysis. Below is the rationale for selecting each feature

- **Loudness**

Loudness is a perceptual measure of the intensity of the audio signal. It is essential for distinguishing between soft and loud speech segments, which can indicate emotional states (e.g., anger vs. calmness) or emphasis in speech. Loudness is also useful for detecting background noise and normalizing audio signals. Loudness is a key feature for distinguishing between soft and loud speech segments. It is also useful for detecting emphasis or stress in spoken language [28] .

It is closely related to the Root Mean Square (RMS) energy of the signal, which quantifies the power of the audio waveform. Expressed in Eq 3.1

$$\text{RMS} = \sqrt{\frac{\sum_{n=1}^{N} y(n)^2}{N}} \tag{3.1}$$

$y(n)$ = Amplitude of the audio signal at sample n.

N = Total number of samples in the signal.

- **Zero-Crossing Rate (ZCR)**

ZCR is a simple yet effective feature for distinguishing between voiced and unvoiced sounds. Voiced sounds (e.g., vowels) have lower ZCR, while unvoiced sounds (e.g., consonants like "s" or "f") have higher ZCR. This feature is particularly useful for segmenting speech and identifying phonetic boundaries [29]. Expressed in Eq 3.2

$$ZCR = \frac{1}{2(N-1)} \sum_{n=1}^{N-1} |sgn\big(y(n+1)\big) - sgn\big(y(n)\big)| \quad (3.2)$$

$sgn(y(n))$ Sign function, returns +1 if $y(n) > 0$, -1 if $y(n) < 0$, and 0 if $y(n) = 0$.

- **Tempo**

Tempo represents the estimated beats per minute (BPM) of the audio signal. It is useful for identifying patterns in speech delivery, such as fast-paced speech (e.g., excitement or urgency) or slow-paced speech (e.g., sadness or hesitation). Tempo is also relevant for applications like speaker identification and emotion recognition [30].

- **Speaking Rate**

Speaking rate measures the number of speech segments (e.g., words or syllables) per unit time. It is a key feature for understanding the fluency and rhythm of speech. Variations in speaking rate can indicate emotional states (e.g., nervousness vs. confidence) or cognitive load [31].

- **Average Pitch**

Pitch is the perceptual frequency of the sound. The average pitch is computed as the mean of the spectral centroid, which represents the "center of mass" of the spectrum. Expressed in Eq 3.3

$$Spectral\ centroid = \frac{\sum_{k=1}^{k} f(k).S(k)}{\sum_{k=1}^{k} S(k)} \quad (3.3)$$

$f(k)$ Frequency of bin k.

$S(k)$ Magnitude of the spectrum at bin k. Average pitch is useful for distinguishing between male and female voices, as well as identifying emotional states (e.g., high pitch for excitement, low pitch for sadness). It is also relevant for prosody analysis [32].

- **Spectral Contrast**

Spectral contrast measures the difference in amplitude between peaks and valleys in the frequency spectrum. It is useful for distinguishing between harmonic (e.g., vowels) and non-harmonic (e.g., noise) components of speech. This feature is particularly relevant for identifying timbral characteristics and emotional cues [33].

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

MFCCs are one of the most widely used features in speech and audio processing. They capture the short-term power spectrum of the audio signal, which is essential for modeling the timbral and spectral characteristics of speech. MFCCs are particularly effective for tasks like speaker recognition, emotion detection, and speech-to-text conversion [34]. Expressed in Eq 3.4

$$Mel(f) = 2595.log_{10}\left(1 + \frac{f}{700}\right) \quad (3.4)$$

Apply Discrete Cosine Transform (DCT) to the log Mel spectrum to obtain MFCCs.

- **Chroma Features**

Chroma features represent the 12 distinct pitch classes in the audio signal. They are useful for analyzing harmonic content and musical structure, making them relevant for applications like music information retrieval and speech emotion recognition. Chroma features can also help identify tonal variations in speech [14].

- **Harmonic and Percussive Components**

Harmonic components represent the pitched elements of the audio, while percussive components represent the non-pitched, transient elements. These are separated using the Harmonic-Percussive Source Separation (HPSS) algorithm. Separating the audio signal into harmonic and percussive components allows for a more detailed analysis of the signal's structure. Harmonic components capture pitched elements (e.g., vowels), while percussive components capture transient elements (e.g., consonants). This separation is useful for tasks like speech enhancement and emotion recognition [35].

- **Spectral Rolloff**

Spectral rolloff measures the frequency below which a specified percentage of the total spectral energy lies. It is useful for distinguishing between tonal and noise-like sounds. This feature is particularly relevant for identifying the brightness or sharpness of speech, which can vary with emotional states [36].

- **Log Mel Spectrogram**

Compute the Mel spectrogram Expressed in Eq 3.5

$$Mel(f) = 2595.log_{10}\left(1 + \frac{f}{700}\right) \quad (3.5)$$

Convert to decibel scale

$$Log\ Mel = 10.log_{10}(Mel\ Spectrogram) \quad (3.6)$$

The log Mel spectrogram is a time-frequency representation of the audio signal that maps the frequency axis to the Mel scale. It is widely used in speech and audio processing because it closely mimics the human auditory system's perception of sound. The log Mel spectrogram is particularly effective for tasks like speech recognition and emotion detection [37]. Expressed in Eq 3.6

3.2.8 Why These Features Are Relevant for XAI in Voice Recognition

- **Comprehensive Representation** - The speech signal is fully represented by the chosen features, which record a variety of the audio signal's acoustic, temporal, and spectral properties.

- **Explainability** - It is simpler to describe the model's choices in terms of concepts that are easy for humans to understand because each attribute has a distinct physical or perceptual meaning.

- **Robustness** - These properties are appropriate for real-world applications because they can withstand changes in recording settings, such as background noise and microphone quality.

- **Emotion and Speaker identification** - A lot of these parameters (including loudness, pitch, and MFCCs) are closely related to speaker traits and emotional states, which makes them very important for emotion analysis and voice identification.

- **Wide Adoption** - These features are widely used in the speech and audio processing community, ensuring that the results of your study are comparable to existing research.

3.3 Step 2. Model training voice emotional recognize

Objective of step 2 is selecting proper ML model for voice emotional reorganization with extracted features

3.3.1 Model Selection and Rationale

For this research, a diverse set of machine learning models was chosen to ensure robust performance and to explore different learning paradigms. The selected models include an Artificial Neural Network (ANN), Random Forest, Support Vector Machine (SVM), XGBoost, and k-Nearest Neighbors (k-NN). Each model was chosen based on its ability to handle tabular and numerical data effectively, as well as its unique strengths in learning complex patterns from data.

3.3.2 Data Preprocessing and Separation

Since all feature data are in tabular and numerical format, a MinMaxScaler was applied to normalize the features, ensuring that each feature contributes equally to the model's learning process. This normalization was crucial for models sensitive to feature scaling, such as ANN and SVM. To maintain the integrity of the model evaluation process, the dataset was shuffled and split into training and testing sets using an 80:20 ratios with stratified sampling This approach preserves the proportion of each class label, ensuring that the model learns from a balanced representation of classes. Class imbalance was mitigated using Synthetic Minority Oversampling Technique (SMOTE). This method was employed on the training set to synthesize new instances of minority classes, preventing the model from becoming biased toward the majority class

3.3.3 Model Descriptions and Justification

A range of traditional machine learning models ANN, SVM, Random Forest, XGBoost, and k-NN were selected for their balance of interpretability and classification ability. Each model brings unique strengths, allowing for comparison and analysis of how acoustic features contribute to emotion recognition.

- **Artificial Neural Network (ANN**)

The Artificial Neural Network (ANN) model used in this research is designed to capture complex, non-linear relationships in high-dimensional tabular data. The architecture begins with an input layer corresponding to the number of scaled features. It includes two hidden layers: the first with 128 units using ReLU activation, batch normalization, and a dropout rate of 0.3, and the second with 64 units similarly configured. The output layer

uses a softmax activation function and has a number of neurons equal to the number of emotion classes encoded.

The model is compiled using the **Adam** optimizer, although **RMSprop** and **SGD** were also tested. The loss function is **categorical cross entropy**, suitable for multi-class classification tasks, and the evaluation metric is **accuracy**.

Hyper parameters were tuned across various values: batch size (16, 32, 64), optimizer (Adam, RMSprop, SGD), dropout rate (0.2, 0.3, 0.4), and activation functions (ReLU, Tanh). The training configuration involved 200 epochs with a validation split of 0.2 and verbose set to 1 for monitoring progress.

The rationale for using ANN lies in its strength to model non-linear interactions between features, making it a robust choice for emotion classification tasks based on acoustic features.

- **Random Forest**

The Random Forest algorithm is an ensemble learning method that constructs multiple decision trees using a bagging approach to improve classification accuracy and reduce overfitting. In this research, several hyper parameters were tuned to optimize model performance. These include the number of trees in the forest (**n_estimators**) with values [50, 100, 200], the maximum depth of each tree (**max_depth**) set to [None, 10, 20, 30], the minimum number of samples required to split an internal node (**min_samples_split**) tested at [2, 5, 10], and the minimum number of samples required to be at a leaf node (**min_samples_leaf**) set to [1, 2, 4].

Additional configurations included enabling **bootstrap** sampling (True) and setting the **random_state** to 42 to ensure reproducibility.

The rationale for using Random Forest lies in its robustness against overfitting and its ability to handle high-dimensional numerical data effectively. Moreover, it provides useful insights into feature importance, which supports explainability and aids in interpreting the role of individual acoustic features in emotion recognition.

- **Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a powerful classification algorithm that works by finding the optimal hyperplane that best separates classes in the feature space. For this study, several hyperparameters were tuned, including the regularization parameter **C** with values [0.1, 1, 10, 100], the **gamma** parameter (which controls the kernel's influence) set

to ['scale', 'auto', 0.01, 0.1, 1], and the **kernel** type, using both 'rbf' and 'linear' to test performance under linear and non-linear mappings.

Additional configurations involved enabling **probability=True** to allow probability estimates and setting **random_state=42** for consistent results.

The rationale for selecting SVM lies in its effectiveness in handling high-dimensional datasets, especially when there is a clear margin of separation between classes. Its mathematical formulation ensures strong generalization and stability, making it suitable for structured speech emotion recognition tasks.

- **XGBoost**

XGBoost (Extreme Gradient Boosting) is a high-performance ensemble learning algorithm based on gradient boosting. It constructs an additive model by sequentially fitting new trees to correct the residuals of prior predictions. In this research, key hyperparameters tuned include **n_estimators** [50, 100, 200], **max_depth** [3, 6, 9], **learning_rate** [0.01, 0.1, 0.2], **subsample** [0.8, 1.0], and **colsample_bytree** [0.8, 1.0]. Additional settings included the use of **objective='multi:softmax'** for multi-class classification, **eval_metric='mlogloss'** to evaluate model performance, and **random_state=42** to maintain consistency.

The rationale for choosing XGBoost is its capability to model complex patterns with high accuracy, along with built-in regularization that helps prevent overfitting. It is particularly efficient on tabular data and manages class imbalance effectively, making it well-suited for SER tasks involving diverse emotional states.

- **k-Nearest Neighbors (k-NN)**

The k-Nearest Neighbors (k-NN) algorithm is a simple, non-parametric method that classifies instances based on the majority class of their **k** closest neighbors in the feature space. In this study, the key hyperparameters tuned were **n_neighbors** [3, 5, 7, 9], **weights** ['uniform', 'distance'], and the **p** parameter for distance metric selection, where *p=1* corresponds to Manhattan distance and *p=2* to Euclidean distance.

The algorithm configuration was set to **algorithm='auto'**, which allows the model to choose the most appropriate internal method for computing nearest neighbors.

The rationale behind using k-NN is its simplicity and interpretability. Despite its basic nature, it performs competitively on certain tabular datasets and serves as an effective baseline model, especially when the class boundaries are complex or irregular.

3.3.4 Hyper parameter Tuning

GridSearchCV was employed for hyper parameter optimization. A 5-fold cross-validation approach was used to evaluate each combination of hyper parameters. This method ensures robust model selection by testing performance across multiple data splits, preventing overfitting and enhancing generalization. The selection of diverse models allows a comprehensive evaluation of different learning paradigms, ensuring the best model is chosen based on performance metrics such as accuracy, precision, recall, and F1-score. By addressing class imbalance, optimizing hyper parameters, and implementing robust data preprocessing, this approach enhances the model's ability to generalize and make accurate predictions on unseen data.

3.4 Step 3. XAI for voice emotional recognize

The Objective of identifying importance feature for each emotions & define value range for it

3.4.1 Method for Selecting Feature Importance and Analyzing Value Ranges

To identify the most influential features for each emotion category and analyze their value ranges, a comprehensive methodology is employed that ensures feature importance is calculated across a representative set of data instances, avoiding reliance on any single observation. This approach provides a robust and accurate understanding of feature behavior for each emotion, facilitating both local and global interpretability.

3.4.2 Feature Importance Analysis

To determine the most significant features contributing to each emotion's classification, both LIME and SHAP methods are used. These two techniques offer complementary insights into feature importance, providing a detailed understanding of how each feature influences model predictions.

- LIME is employed for local explanations by approximating complex model decisions with interpretable surrogate models. This method is well-suited for explaining individual predictions by perturbing features and observing how changes affect model outputs.
- SHAP, based on Shapley values, provides both local and global interpretability. It ensures consistent and accurate feature importance across both individual instances and the entire dataset, by fairly attributing contributions to each feature in a model's prediction.

3.4.3 Instance Sampling and Explainability

1. **Emotion-Specific Sampling** - For each emotion class, instances from the test set corresponding to that emotion are identified. A subset of these instances is sampled to ensure feature importance calculation is not biased by outliers.
2. **LIME Explanations** - LIME explains each sampled instance by perturbing its features and recording the change in model predictions. Each perturbation provides a set of feature weights, representing how important each feature is for the given prediction.
3. **SHAP Explanations** - SHAP values are computed for each instance, capturing the marginal contribution of each feature to the predicted outcome. SHAP ensures that the contribution of each feature is calculated in a way that satisfies fairness and consistency principles.

3.5 Step 4. Aggregated Feature Importance (AFI) Explanation

The Aggregated Feature Importance (AFI) is a custom metric developed to evaluate and combine feature importance scores from multiple machine learning models, with each model's contribution weighted by its predictive accuracy. The motivation for AFI is to derive a unified, interpretable, and performance-aware measure of feature relevance that accounts for both the strength of predictive models and the interpretability of feature attribution techniques. The construction of AFI follows a stepwise integration of models. Initially, feature importance is computed using only a single model. Then, additional models are progressively added in a fixed order, with AFI recalculated at each stage to observe how the importance values evolve. The model inclusion sequence is as follows:

1. Begin with **Model 1**.
2. Add **Model 2**, and aggregate their feature importance scores.
3. Continue until all five models are included.
4. At each stage, evaluate **AFI convergence** by comparing feature importance rankings and values, to determine whether the set of dominant features stabilizes

Let:

- M denote the total number of models included at a given stage, where M∈{1,2,3,4,5},
- a_i be the predictive accuracy of the i^{th} model,
- FI_i^{LIME} and FI_i^{SHAP} represent the LIME and SHAP feature importance vectors for the i^{th} model.

The AFI is calculated using the following equation in 3.7:

$$AFI = \sum_{i=1}^{M} \frac{a_i}{\sum_{j=1}^{M} a_j} \left(\frac{\left\| FI_i^{LIME} \right\|_1 + \left\| FI_i^{SHAP} \right\|_1}{2} \right) \quad (3.7)$$

This formulation is built on two core principles:

1. **Accuracy-weighted aggregation**: The contribution of each model to the AFI is scaled by its predictive accuracy. This ensures that models with higher performance have greater influence on the final aggregated importance scores. As such, AFI prioritizes interpretations from models that are more trustworthy in terms of prediction quality.
2. **Use of the L1 norm**: The L1 norm (i.e., the sum of the absolute values of feature importance scores) is used to quantify the magnitude of each feature's influence. This choice is intentional and justified for the following reasons:
    - It **preserves the interpretability** of additive explanation methods like SHAP and LIME by measuring the total impact without canceling out positive and negative values.
    - It is **robust to sparsity**, which is particularly beneficial in high-dimensional datasets where many features may have near-zero importance.
    - It provides a **simple and consistent scale** across models, making it easier to compare and aggregate importance values.

By monitoring the AFI values at each stage of model addition (e.g., from 1 model to 5 models), the analysis can detect whether the most influential features remain consistent. If the feature importance rankings and scores stabilize, this indicates **convergence**, suggesting that the identified features are robust across diverse model architectures and learning mechanisms. This lends greater credibility to feature selection and interpretation in the modeling pipeline.

In conclusion, the AFI framework provides a structured, interpretable, and performance-aware method for aggregating feature importance across multiple models, supporting comprehensive insight into the underlying data-generating process.

3.6 Step 5. Value Range Analysis

Once the most important features are identified through the AFI framework, a **value range analysis** is conducted to explore the distribution of these features within each emotion category. This step provides a clearer understanding of how the values of influential features vary between emotional states. For each emotion class, the value ranges (i.e., minimum and maximum) of the top-ranked features are extracted using instances from the test dataset, ensuring the analysis reflects the true variability present in the model evaluation phase. Once the most important features are identified for each emotion, the value ranges of these features are analyzed to gain deeper insight into their distribution and variability within each emotional category.

3.6.1 Descriptive Statistical Analysis

For each selected top feature, descriptive statistics are calculated, including:

- **Minimum (Min)** - The lowest observed value for the feature.
- **Maximum (Max)-** The highest observed value.
- **Mean** - The average value across all instances within the emotion category.
- **Standard Deviation (Std)** - The variability or spread of the feature values around the mean.

These statistics are computed using all instances from the test set that belong to a specific emotion, ensuring that the value ranges are representative of the full data distribution for that emotion.

3.6.2 Statistical Significance Testing with One-Sample t-Test

To validate the emotion-specific relevance of features identified through the Aggregated Feature Importance (AFI) process, a two-tailed one-sample t-test was performed on each top-ranked feature within each emotion category. The purpose of this test is to determine whether the average value of a given feature is statistically different from zero in a given emotional context. A significant deviation from zero suggests that the feature has a consistent, non-random influence in characterizing that particular emotion.

Why the One-Sample t-Test is Used?

The **one-sample t-test** is appropriate in this context for several key reasons:

- The test evaluates whether the mean of a single group of values (i.e., a feature under a specific emotion category) is significantly different from a reference value in this case, zero, which represents the absence of a systematic emotional signal.
- Unlike independent or paired t-tests that compare two groups, this test suits scenarios where a single sample distribution is evaluated against a hypothesized benchmark.
- The population mean and standard deviation are unknown, but the sample size for each emotion exceeds 30, which, under the Central Limit Theorem, allows for reliable approximation using sample statistics.
- The use of zero as the null mean reflects a neutral baseline if the mean of a feature is statistically different from zero, it implies that the feature exhibits a consistent directional pattern (positive or negative) related to the emotion being analyzed.

Thus, the one-sample t-test serves as a statistically sound method to evaluate the **consistency and significance** of features identified through AFI and descriptive analyses**.**

Hypotheses for each feature–emotion pair, the following hypotheses are tested:

- **Null Hypothesis (H_0):** The feature's true mean is equal to zero.

$$H_0: \mu = 0$$

- **Alternative Hypothesis (H_1):** The feature's true mean is not equal to zero.

$$H_1: \mu \neq 0$$

This is a two-tailed test, allowing detection of deviations in both positive and negative directions.

The **t-statistic** is calculated using the formula in 3.8:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\bar{x}}{s/\sqrt{n}} \quad (3.8)$$

Where:

- $\bar{x}$ = sample mean of the feature for the emotion
- $\mu_0 = 0$ (the null hypothesis mean)
- s = sample standard deviation
- n = number of instances in the emotion category

The resulting t-value is compared to the critical value from the t-distribution with n−1n - 1 degrees of freedom. A feature is considered statistically significant if either:

- The **p-value < 0.05**, or
- $|t| > t_{\text{critical}}$ (e.g., approximately 1.96 for large n, as the t-distribution approaches normality)

A statistically significant result indicates that the feature's average value is not due to random variation, but reflects a consistent pattern specific to the emotion. This adds further credibility to the feature's interpretability and complements the findings from both the AFI rankings and descriptive statistical analysis.

Together, this one-sample t-test provides an important layer of statistical validation, ensuring that selected features are not only ranked as important but also demonstrate meaningful and reliable differences in emotional expression.

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Each feature's distribution

To visually examine the distribution and variability of the extracted features, box plots were generated for each parameter. These plots help identify the range, median, and presence of outliers in the dataset. The visualization confirms that most features such as loudness, ZCR, MFCCs, chroma, and log Mel spectrogram values exhibit distinct distribution patterns, reflecting the acoustic and emotional variability present in the speech samples. This exploratory analysis validates the diversity of the dataset and supports the suitability of these features for modeling emotion recognition with explainable AI (XAI). Notably, features like tempo and spectral contrast show wider spread, suggesting they could be key discriminators in emotional classification.

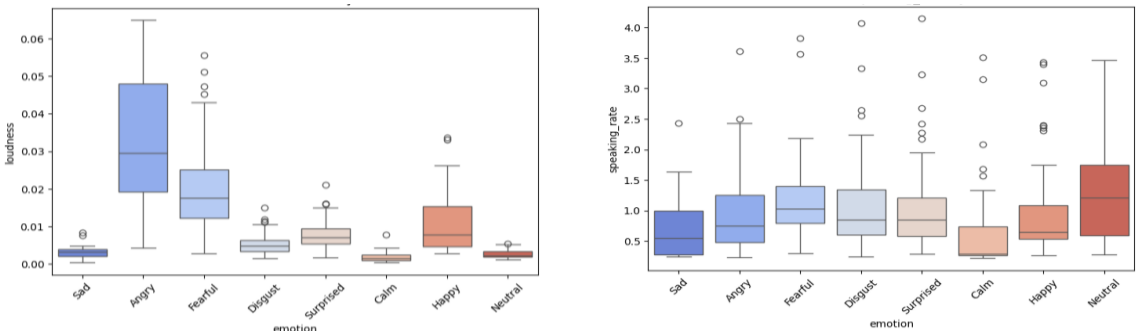


Figure 4.4.1: Distribution of loudness (left) and speaking_rate (write) by

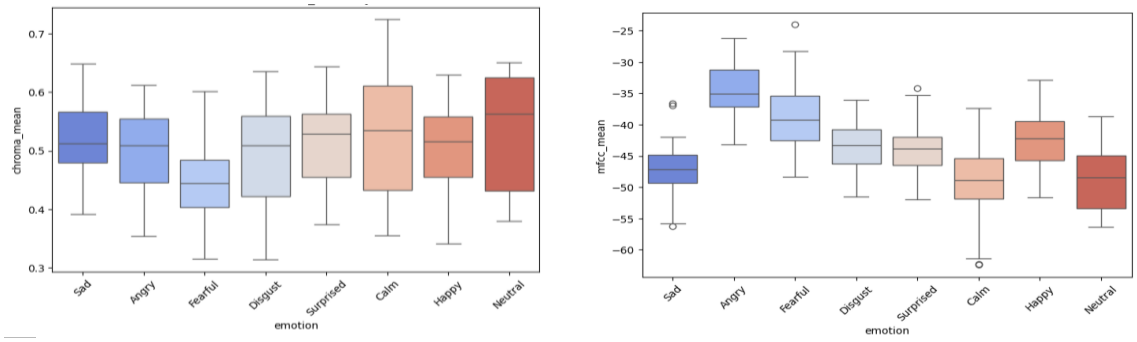


Figure 4.4.2: Distribution of chroma_mean (left) and mfcc_mean (write) by Emotions

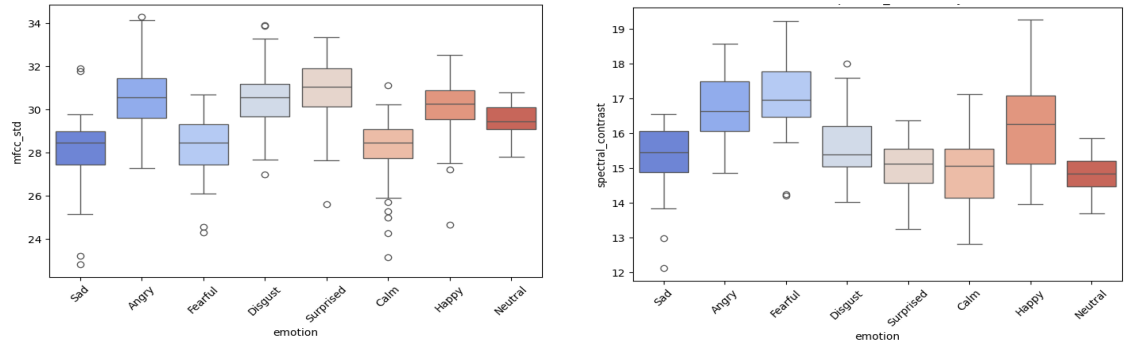


Figure 4.4.3: Distribution of mfcc_std (left) and spectral_contrast (write) by Emotions

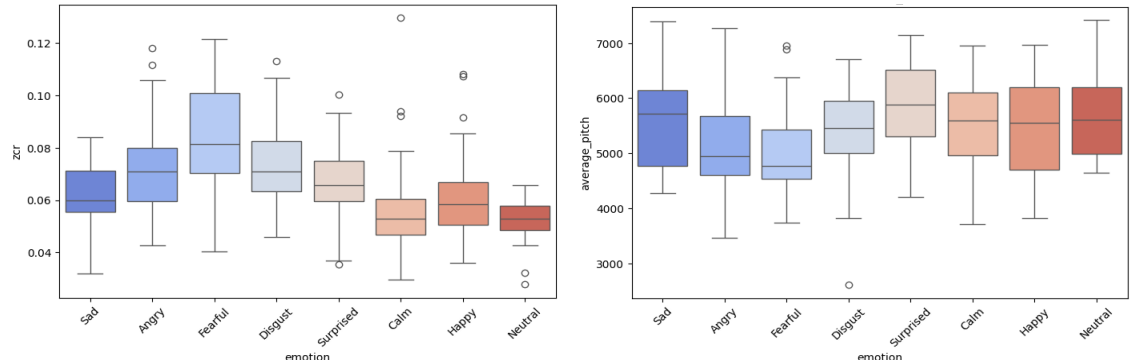


Figure 4.4.4: Distribution of zcr (left) and average_pitch (write) by Emotions

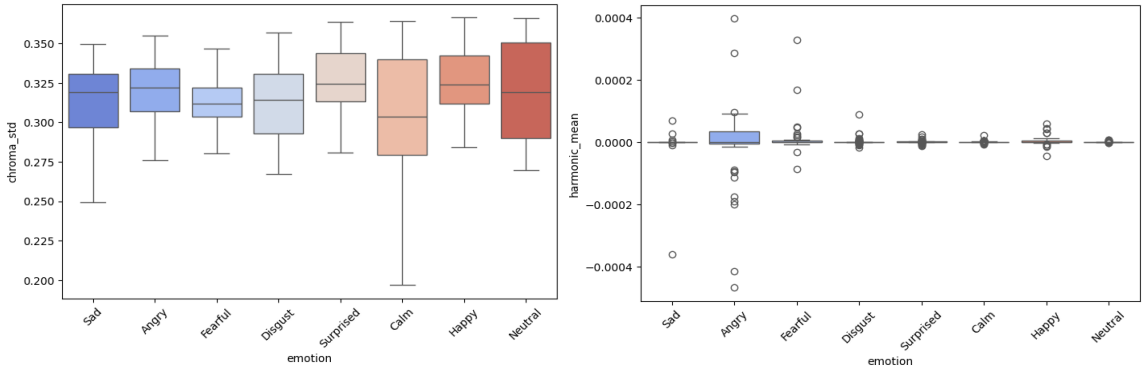


Figure 4.4.5: Distribution of chroma_std (left) and harmonic_mean (write) by Emotions

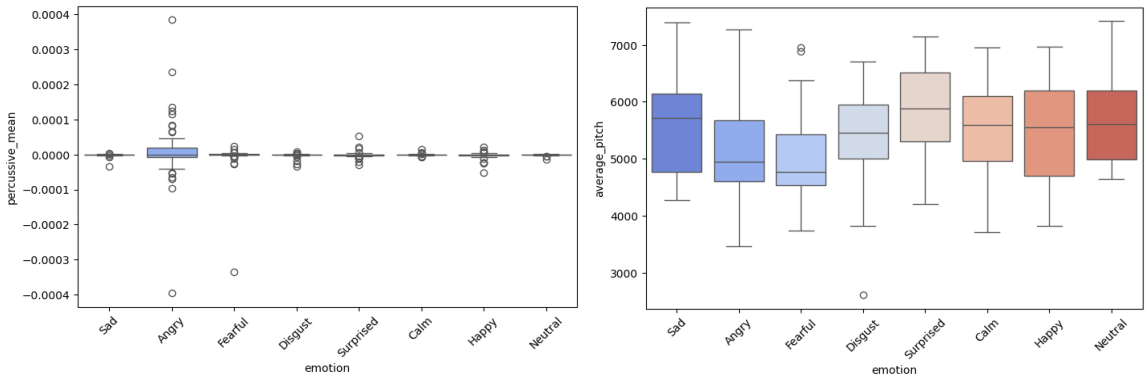


Figure 4.4.6: Distribution of percussive_mean (left) and average_pitch (write) by Emotions

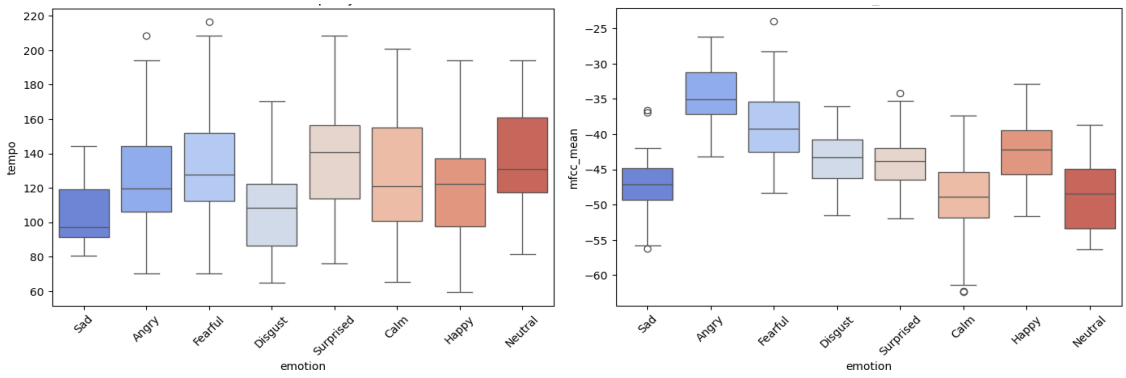


Figure 4.4.7: Distribution of tempo (left) and mfcc_mean (write) by Emotions

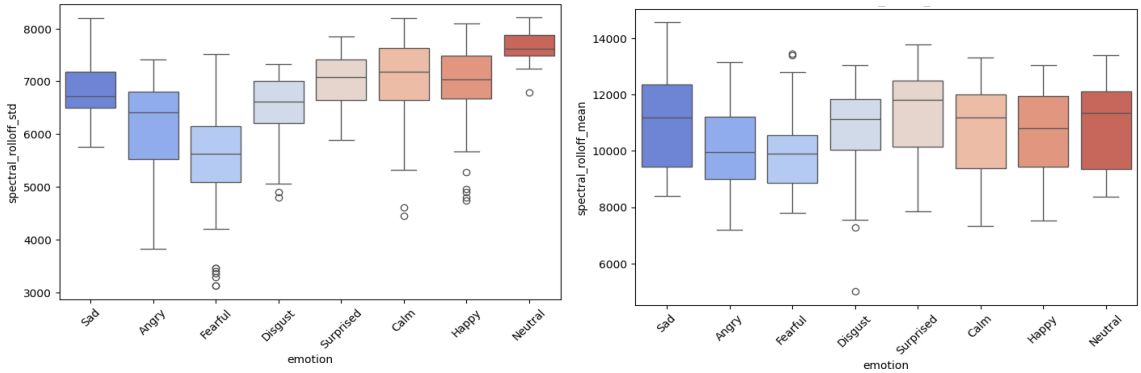


Figure 4.4.8: Distribution of spectral_rolloff_std (left) and spectral_rolloff_mean (write) by Emotions

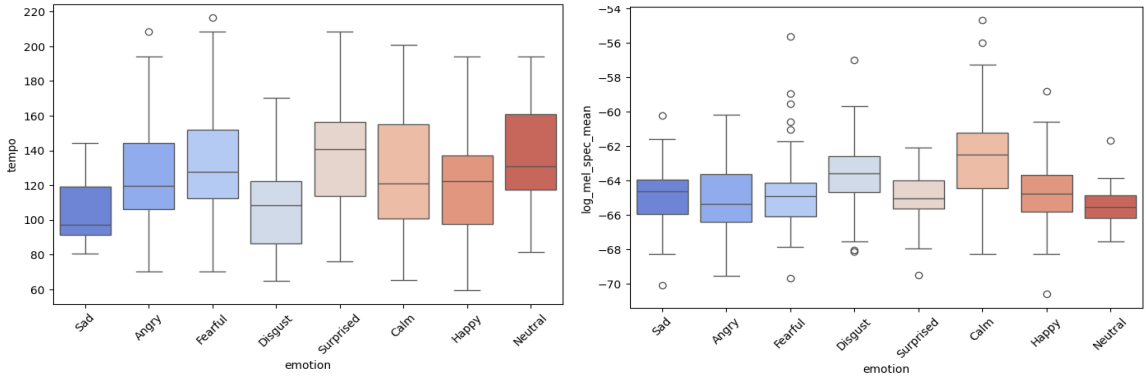


Figure 4.4.9: Distribution of tempo (left) and log_mel_spec_mean (write) by Emotions

4.2 Selected models with best parameters based on selected parameter grid

TABLE 4.4.1: BEST HYPERPARAMTER AND VALIDATION RESULTS FOR EACH MODELS

Model	Best Parameters	CV	Tuning Method
Random Forest	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}	0.7937	GridSearchCV
SVM	{'C': 100, 'gamma': 1, 'kernel': 'rbf'}	0.8318	GridSearchCV
XGBoost	{'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200, 'subsample': 0.8}	0.7963	GridSearchCV
k-NN	{'n_neighbors': 3, 'p': 1, 'weights': 'distance'}	0.6761	GridSearchCV
ANN	{'batch_size': 32, 'epochs': 200, 'optimizer': 'adam', 'dropout_rate': 0.3}	0.9620	GridSearchCV

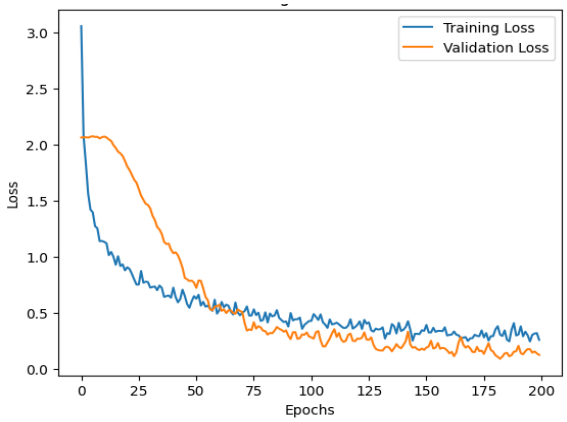


Figure 4.4.10:ANN Training & Validation Loss

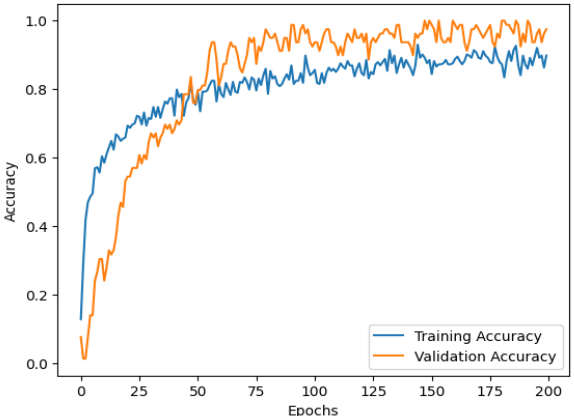


Figure 4.4.11: ANN Training & Validation Accuracy

4.3 Each mode accuracy compression

Random Forest Test Accuracy: 0.6533

TABLE 4.4.2: PRECISION, RECALL, F1, SUPPORT VALUE EACH CLASS FOR RANDOM FOREST MODEL

Class	Precision	Recall	F1-Score	Support
Angry	0.61	0.92	0.73	12
Calm	1	0.75	0.86	12
Disgust	0.5	0.36	0.42	11
Fearful	0.71	0.62	0.67	8
Happy	0.33	0.22	0.27	9
Neutral	0.8	0.67	0.73	6
Sad	0.5	1	0.67	5
Surprised	0.75	0.75	0.75	12
Macro Avg	0.65	0.66	0.64	75
Weighted Avg	0.66	0.65	0.64	75

SVM Test Accuracy: 0.7867

TABLE 4.4.3: PRECISION, RECALL, F1, SUPPORT VALUE EACH CLASS FOR SVM MODEL

Class	Precision	Recall	F1-Score	Support
Angry	1	0.92	0.96	12
Calm	0.89	0.67	0.76	12
Disgust	0.75	0.55	0.63	11
Fearful	0.71	0.62	0.67	8
Happy	0.64	0.78	0.7	9
Neutral	0.67	1	0.8	6
Sad	0.71	1	0.83	5
Surprised	0.85	0.92	0.88	12
Macro Avg	0.78	0.81	0.78	75
Weighted Avg	0.8	0.79	0.78	75

XGBoost Test Accuracy: 0.7200

TABLE 4.4.4: PRECISION, RECALL, F1, SUPPORT VALUE EACH CLASS FOR XGBOOST MODEL

Class	Precision	Recall	F1-Score	Support
Angry	0.69	0.92	0.79	12
Calm	1	0.67	0.8	12
Disgust	0.64	0.82	0.72	11
Fearful	0.71	0.62	0.67	8
Happy	0.5	0.22	0.31	9
Neutral	0.71	0.83	0.77	6
Sad	0.71	1	0.83	5
Surprised	0.75	0.75	0.75	12
Macro Avg	0.72	0.73	0.7	75
Weighted Avg	0.73	0.72	0.7	75

k-NN Test Accuracy: 0.5200

TABLE 4.4.5: PRECISION, RECALL, F1, SUPPORT VALUE EACH CLASS FOR KNN MODEL

Class	Precision	Recall	F1-Score	Support
Angry	0.92	0.92	0.92	12
Calm	0.86	0.5	0.63	12
Disgust	0.44	0.36	0.4	11
Fearful	0.67	0.5	0.57	8
Happy	0.15	0.22	0.18	9
Neutral	0.4	0.67	0.5	6
Sad	0.36	0.8	0.5	5
Surprised	0.57	0.33	0.42	12
Macro Avg	0.55	0.54	0.52	75
Weighted Avg	0.59	0.52	0.53	75

ANN Test Accuracy: 0.8133

TABLE 4.4.6: PRECISION, RECALL, F1, SUPPORT VALUE EACH CLASS FOR ANN MODEL

Class	Precision	Recall	F1-Score	Support
Angry	1	0.83	0.91	12
Calm	0.92	0.92	0.92	12
Disgust	0.71	0.45	0.56	11
Fearful	0.8	1	0.89	8
Happy	0.62	0.89	0.73	9
Neutral	1	0.83	0.91	6
Sad	0.8	0.8	0.8	5
Surprised	0.77	0.83	0.8	12
Accuracy			**0.81**	**75**
Macro avg	0.83	0.82	0.81	75
Weighted avg	0.83	0.81	0.81	75

Instead of selecting only the top-performing models, this study adopts a broader, more inclusive strategy by retaining all five evaluated classifiers ANN, Random Forest, SVM, XGBoost, and k-Nearest Neighbors (k-NN) for further analysis. This decision is supported

by class-wise performance visualizations, such as the precision, recall, and F1-score heatmaps, which reveal that different models excel in different emotional categories. For instance, ANN yields strong results for Angry, Calm, and Surprised, while SVM performs consistently well across Disgust, Sad, and Neutral. Similarly, Random Forest and XGBoost show stronger recall and F1-scores for Fearful and Happy, suggesting their strength in handling underrepresented or complex categories. Although k-NN demonstrates relatively weaker overall performance, its inclusion contributes to classifier diversity, which is a valuable asset in ensemble or hybrid modeling. By retaining all classifiers, the study acknowledges the complementary strengths of each, moving beyond global accuracy and embracing the nuance required for emotion recognition tasks. All models were optimized using GridSearchCV to ensure fair and consistent evaluation.

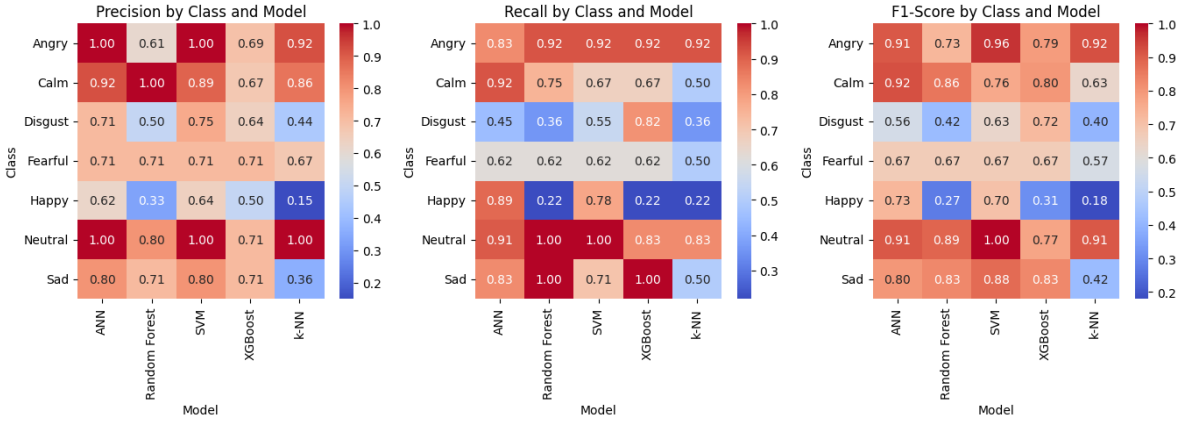


Figure 4.4.12:Precision, Recall and F1 score comparisons

Building on this rationale, the study further applies model interpretability techniquesspecifically LIME and SHAP to each model individually. Given the class-specific strengths observed across classifiers, it becomes essential to examine not only performance metrics but also the reasoning behind model predictions. By using LIME and SHAP separately for each classifier, the study enables a fine-grained analysis of how features influence decision-making, allowing for transparent interpretation of results. This is particularly important in affective computing, where understanding the underlying decision logic enhances both trust and applicability. Since no single model consistently outperforms the rest across all emotions, individualized interpretability analysis ensures that each model's unique behavioral patterns and biases are captured. This dual focus on performance and explainability creates a more robust and interpretable framework for emotion recognition,

4.4 Feature Importance Results for ANN, SVM, XGB, RF, KNN with LIME and SHAP

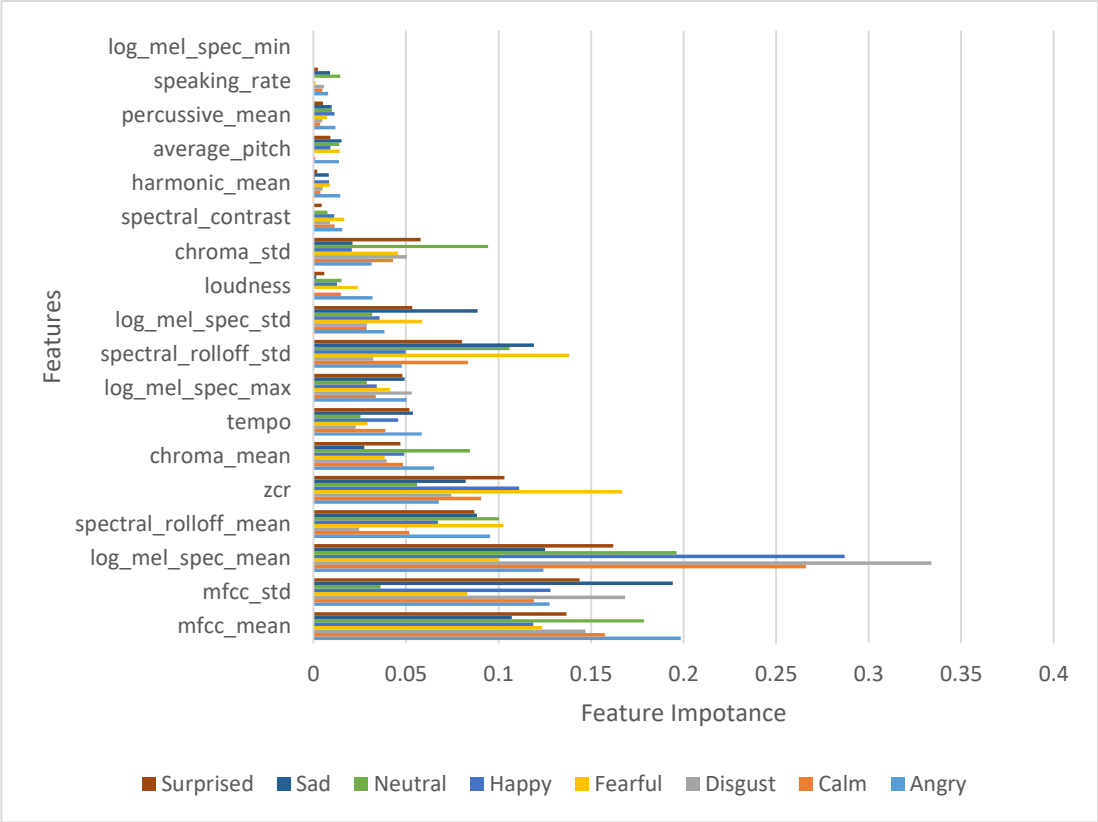


Figure 4.4.13: Feature Importance Results for ANN with LIME

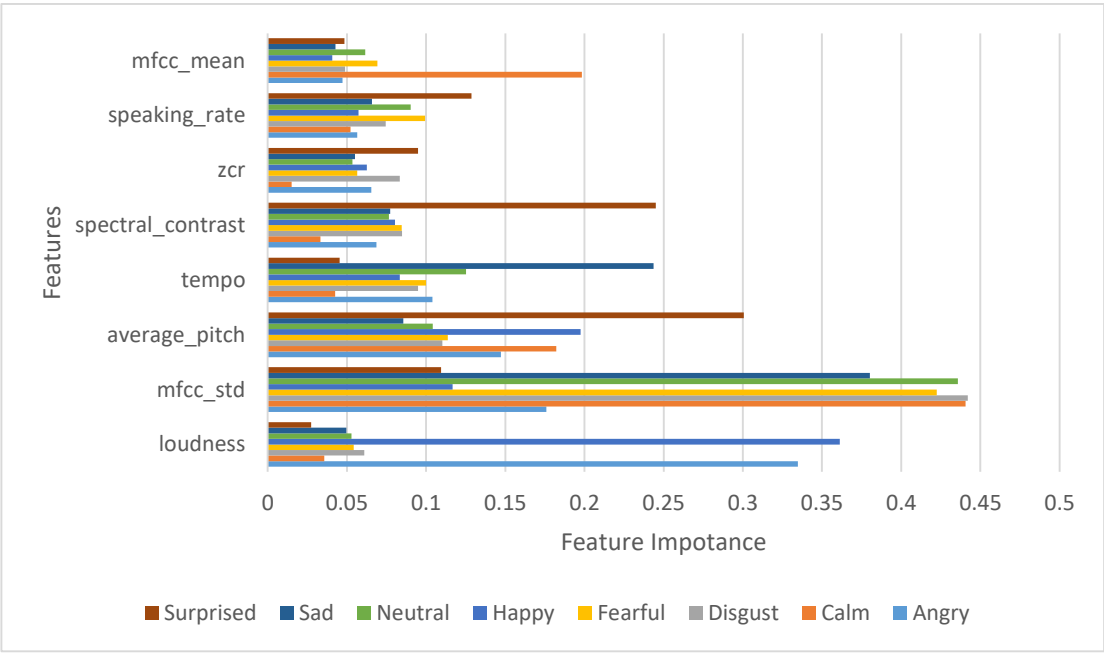


Figure 4.4.14: Feature Importance Results for ANN with SHAP

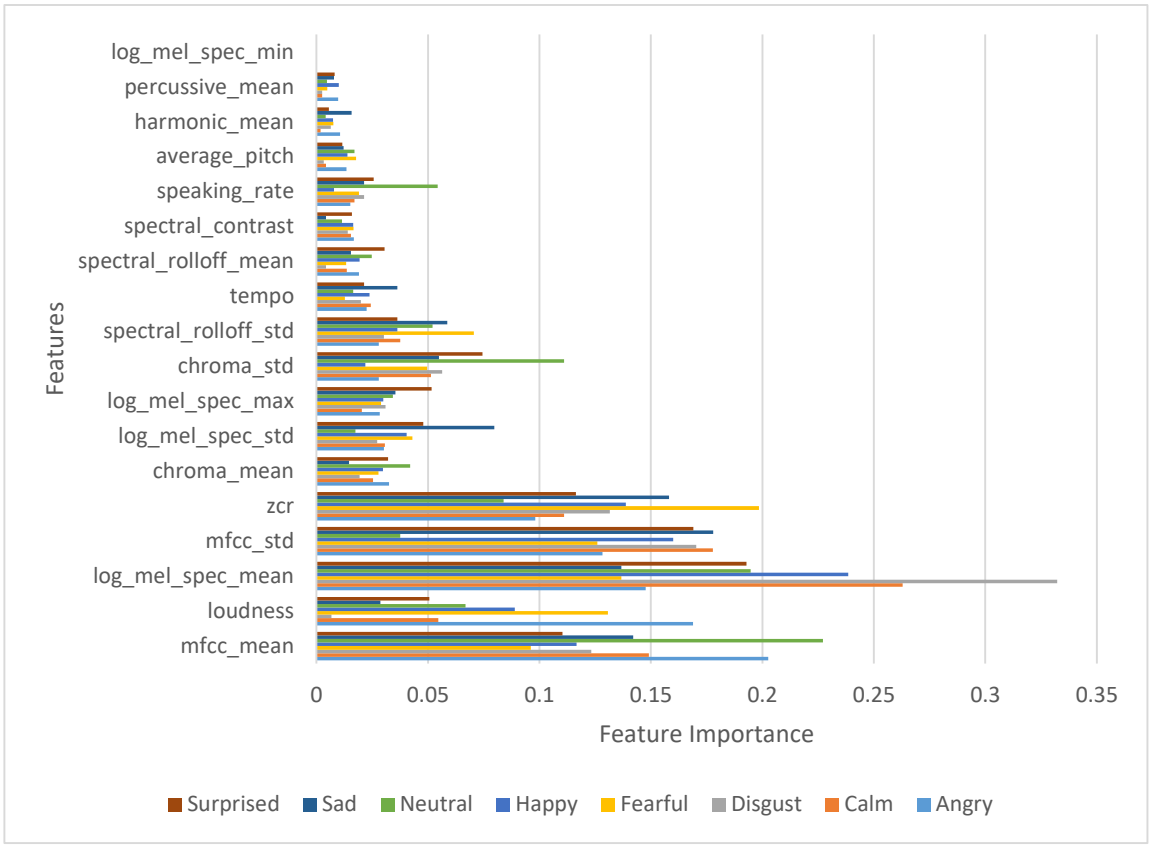


Figure 4.4.15: Feature Importance Results for SVM with LIME

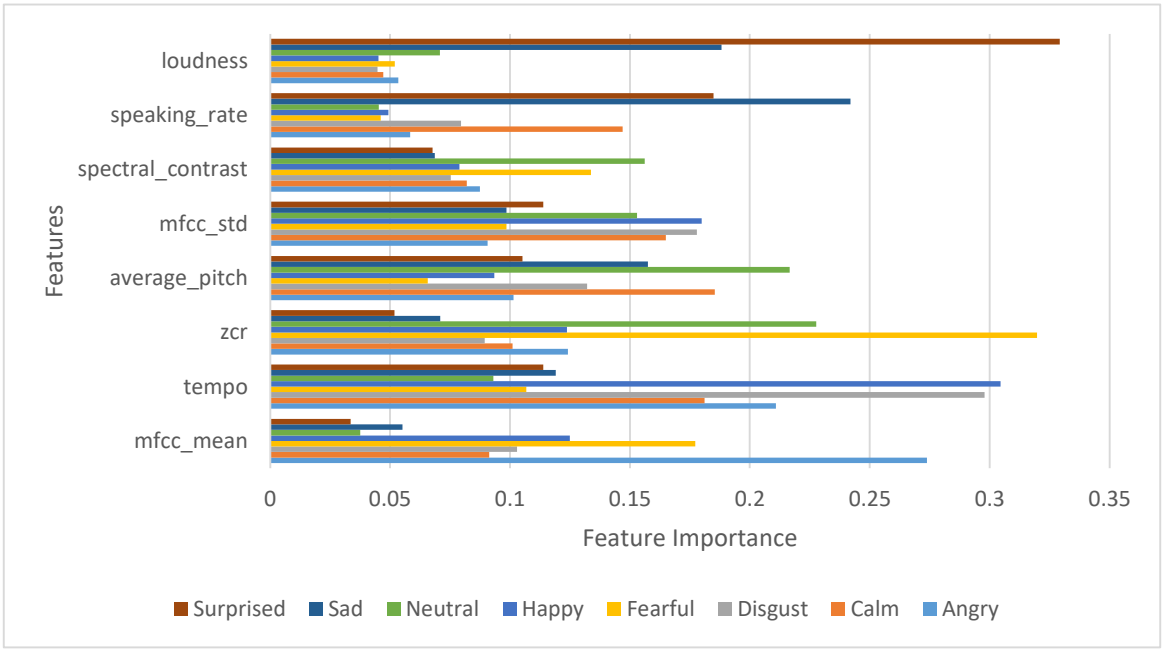


Figure 4.4.16: Feature Importance Results for SVM with SHAP

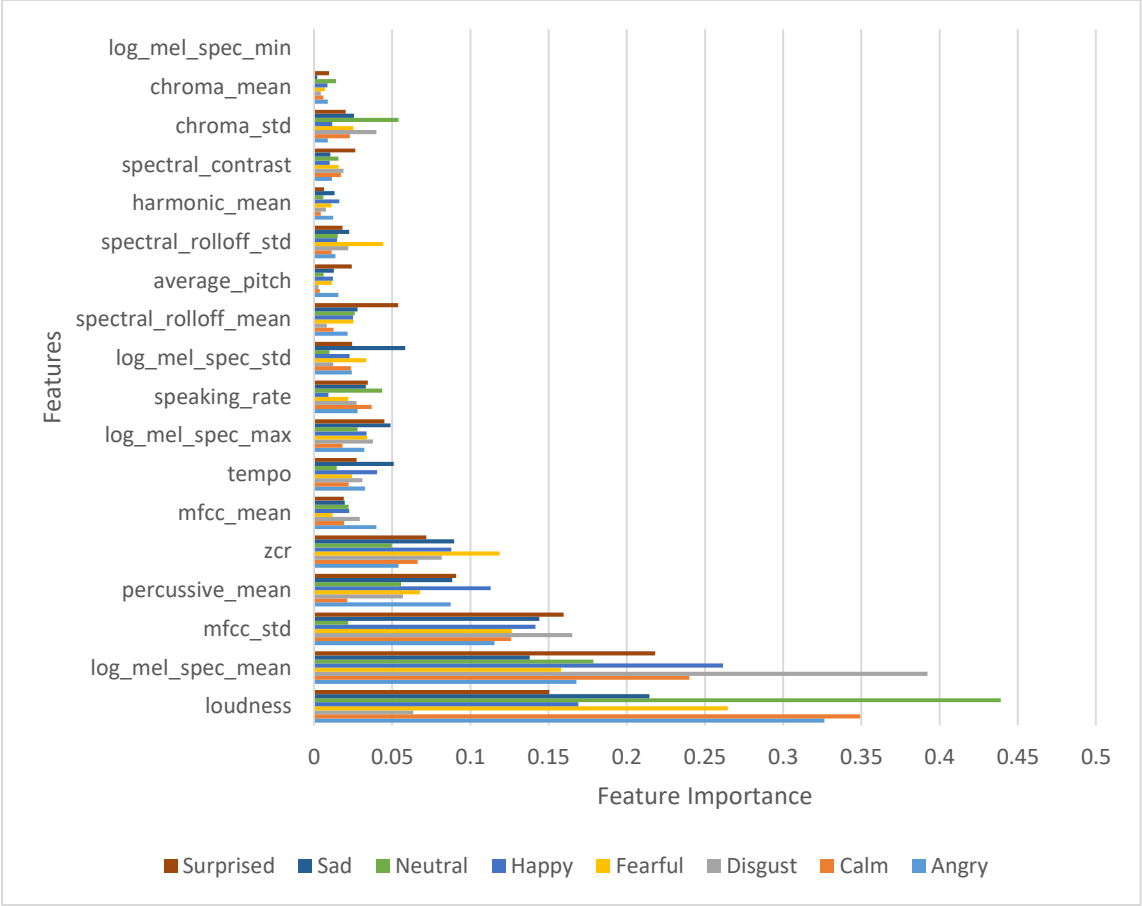


Figure 4.4.17: Feature Importance Results for XGBoost with LIME

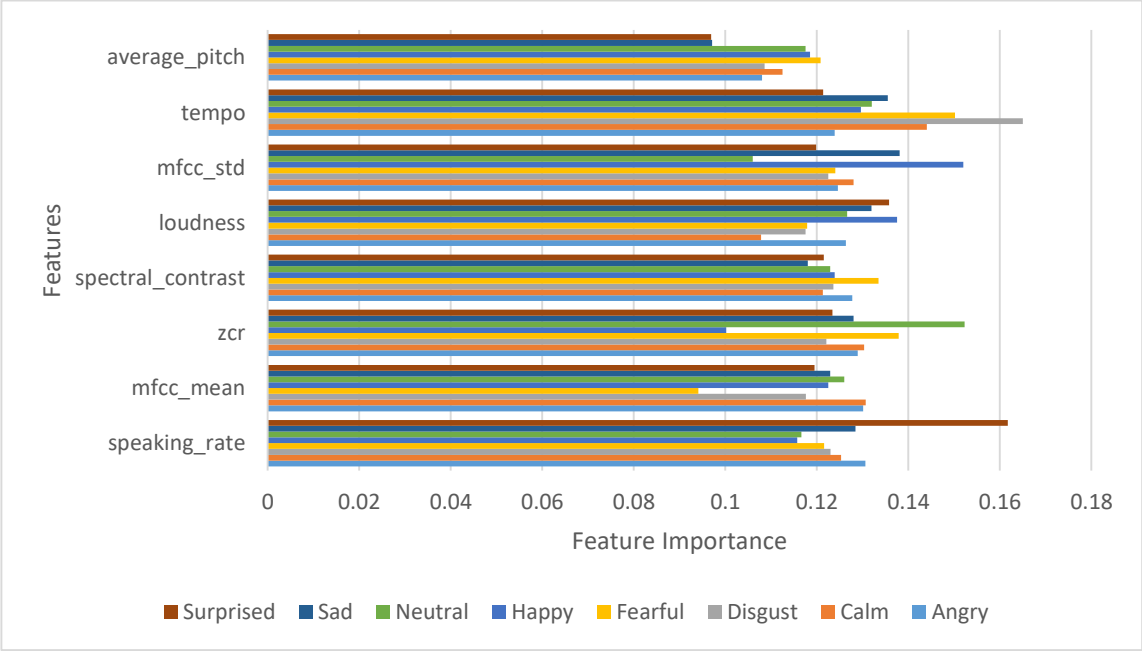


Figure 4.4.18: Feature Importance Results for XGBoost with SHAP

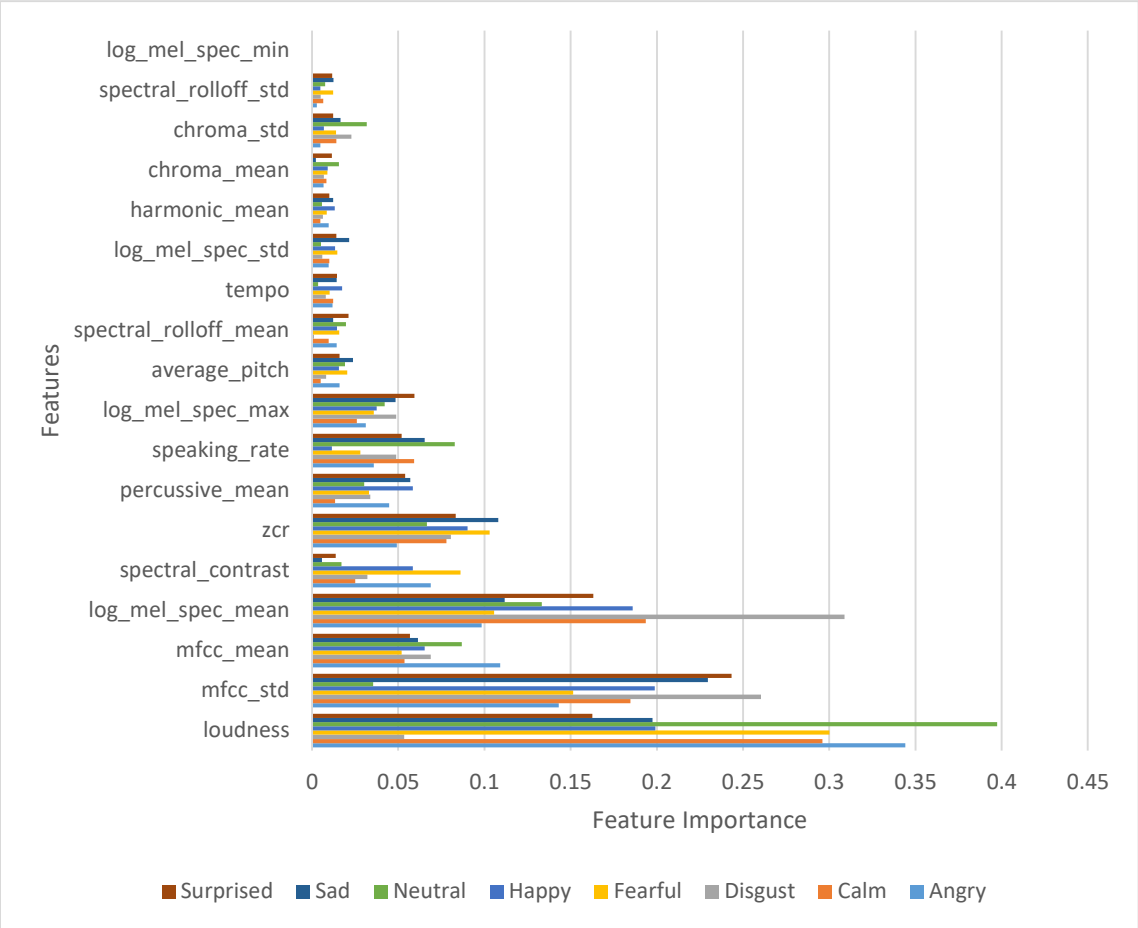


Figure 4.4.19: Feature Importance Results for Random Forest with LIME

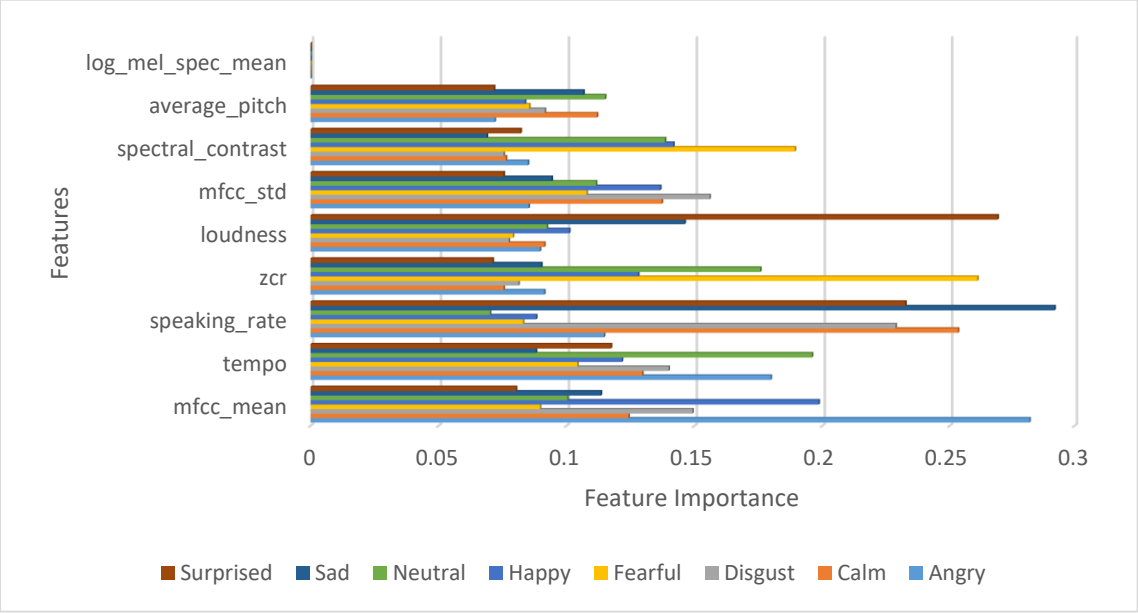


Figure 4.4.20: Feature Importance Results for Random Forest with SHAP

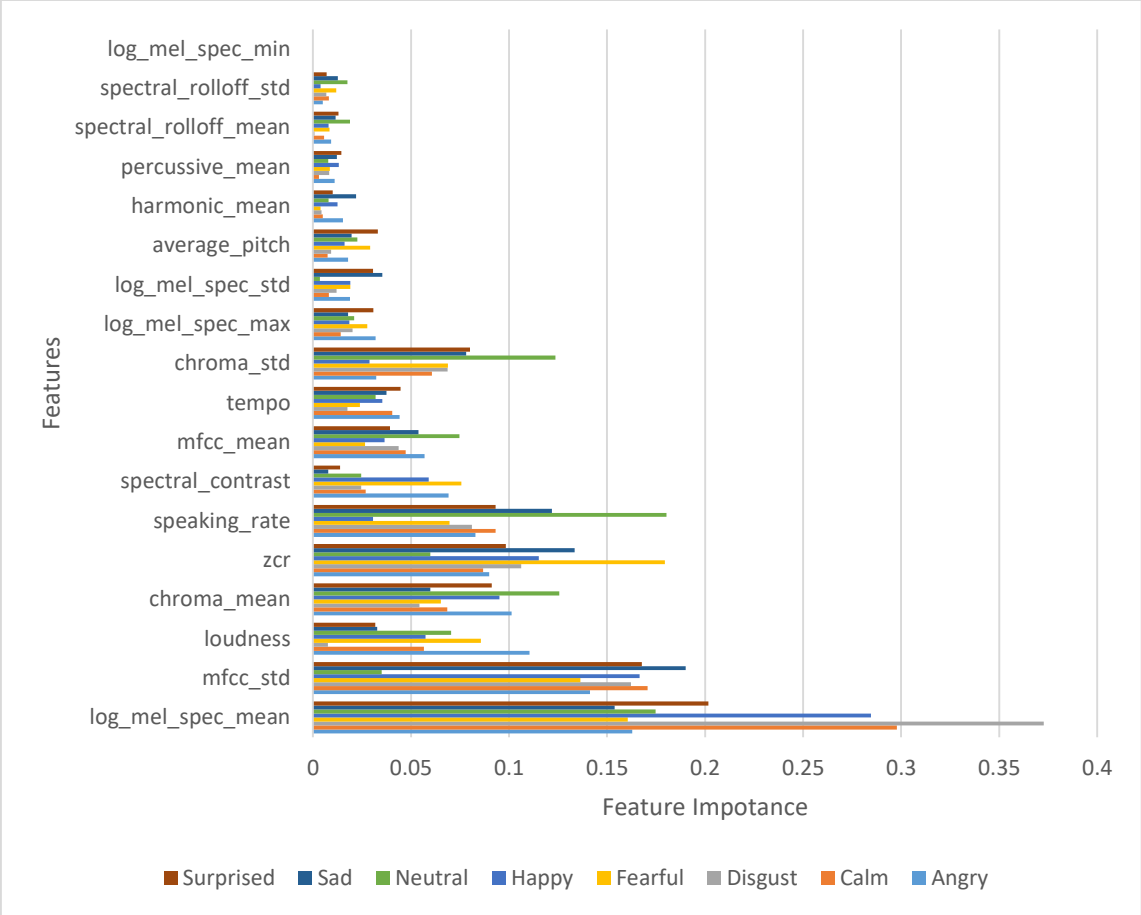


Figure 4.4.21: Feature Importance Results for KNN with LIME

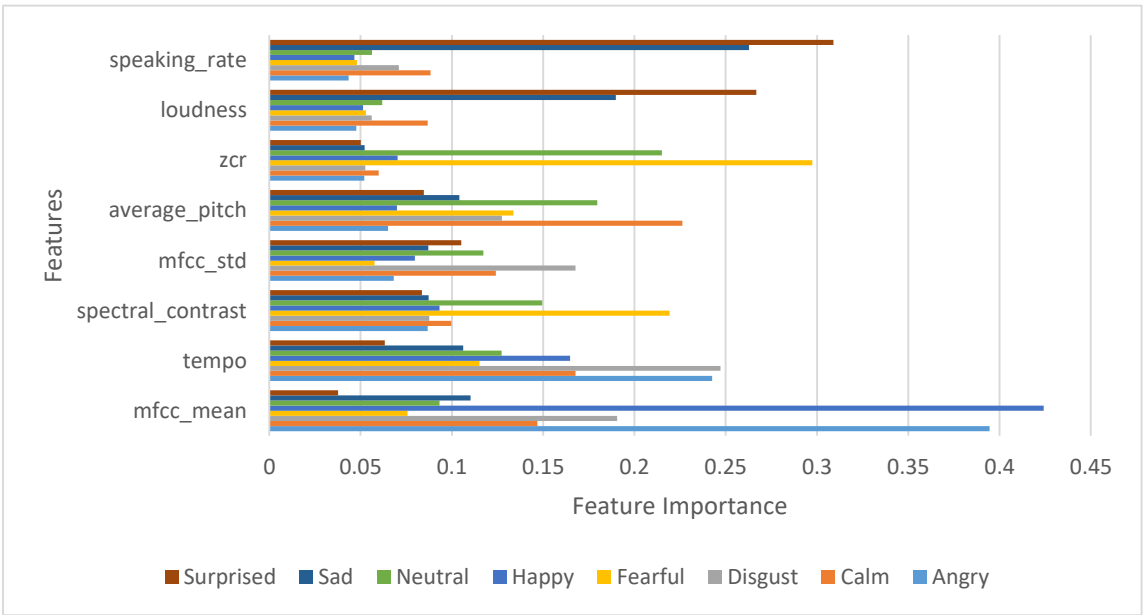


Figure 4.4.22: Feature Importance Results for KNN with SHAP

4.6 Aggregated Feature Importance (AFI) Explanation

Let:

- M denote the total number of models included at a given stage, where M∈{1-ANN, 2-SVM, 3-XGB,4-RF, 5-KNN},
- a_i be the predictive accuracy of the i^{th} model,
- FI_i^{LIME} and FI_i^{SHAP} represent the LIME and SHAP feature importance vectors for the i^{th} model.

The AFI is calculated using the following equation in (4.1):

$$AFI = \sum_{i=1}^{M} \frac{a_i}{\sum_{j=1}^{M} a_j} \left(\frac{\left\| FI_i^{LIME} \right\|_1 + \left\| FI_i^{SHAP} \right\|_1}{2} \right) \quad (4.1)$$

TABLE 4.4.7:FEATURE IMPORTANCE VALUE FOR EACH EMOTIONS FOR M = 1

Feature-shape	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
mfcc_mean	0.1228	0.17785	0.0979	0.09635	0.0798	0.1201	0.07495	0.0926
loudness	0.1833	0.0253	0.0305	0.0391	0.187	0.03395	0.0256	0.01665
log_mel_spec_mean	0.0621	0.13305	0.16695	0.05005	0.1435	0.09815	0.0626	0.08095
mfcc_std	0.1518	0.2799	0.30525	0.25275	0.1224	0.23605	0.2872	0.12665
zcr	0.06655	0.05275	0.0789	0.1117	0.08685	0.0547	0.06865	0.0991
chroma_mean	0.0326	0.0242	0.0198	0.01925	0.0245	0.04225	0.0137	0.02345
chroma_std	0.01575	0.0215	0.02525	0.0228	0.01035	0.0472	0.0105	0.0289
log_mel_spec_std	0.0192	0.0143	0.01445	0.02935	0.01785	0.01585	0.0443	0.0267
log_mel_spec_max	0.0251	0.01685	0.0265	0.02065	0.0171	0.01445	0.02465	0.024
spectral_rolloff_std	0.0238	0.0417	0.0161	0.0691	0.0249	0.05295	0.05955	0.0401
spectral_rolloff_mean	0.0477	0.0259	0.01235	0.05135	0.0336	0.05015	0.04415	0.0435
tempo	0.0813	0.0407	0.0589	0.06455	0.06455	0.0753	0.1487	0.04855
spectral_contrast	0.0421	0.02225	0.04695	0.05075	0.0457	0.0421	0.03865	0.1248
speaking_rate	0.03225	0.0285	0.04	0.05025	0.02875	0.0524	0.0374	0.06555
average_pitch	0.0805	0.0916	0.0553	0.06395	0.1033	0.0591	0.0504	0.15485
harmonic_mean	0.00725	0.00185	0.00245	0.00435	0.0042	0.0004	0.00415	0.00105
percussive_mean	0.0059	0.0018	0.00235	0.00365	0.0057	0.00495	0.0049	0.00255
log_mel_spec_min	0	0	0	0	0	0	0	0

TABLE 4.4.8:FEATURE IMPORTANCE VALUE FOR EACH EMOTIONS FOR M = 2

Feature-shape	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
mfcc_mean	0.17959135	0.149503	0.105349	0.11619	0.099984	0.126123	0.086603	0.082422
loudness	0.147823845	0.037912	0.028164	0.064791	0.128021	0.05111	0.066337	0.101862
log_mel_spec_mean	0.06785289	0.132288	0.166557	0.059073	0.131576	0.097781	0.065452	0.088571
mfcc_std	0.130976505	0.226551	0.240764	0.183666	0.14578	0.166819	0.213961	0.133927
zcr	0.088455235	0.078933	0.094462	0.184201	0.108681	0.104435	0.091194	0.0917
chroma_mean	0.024560705	0.018521	0.014834	0.016619	0.01978	0.031801	0.010553	0.019811
chroma_std	0.01486494	0.023565	0.026676	0.023783	0.010645	0.051306	0.018834	0.033006
log_mel_spec_std	0.017208615	0.014816	0.014032	0.02549	0.019005	0.012359	0.042137	0.025348
log_mel_spec_max	0.01974047	0.013556	0.021091	0.017601	0.016067	0.015753	0.021233	0.024885
spectral_rolloff_std	0.018956755	0.03044	0.015608	0.052481	0.021556	0.039723	0.044701	0.029283
spectral_rolloff_mean	0.02891706	0.016509	0.007335	0.029371	0.021824	0.031588	0.026252	0.029609
tempo	0.09870618	0.07121	0.10807	0.062214	0.113523	0.065171	0.113765	0.057917
spectral_contrast	0.047017	0.03528	0.045819	0.062772	0.046683	0.062604	0.037544	0.083989
speaking_rate	0.03446265	0.054806	0.045163	0.041571	0.028676	0.051122	0.083767	0.08507
average_pitch	0.069166315	0.093198	0.061372	0.053034	0.078912	0.087471	0.067339	0.107426
harmonic_mean	0.006291185	0.001383	0.002819	0.004055	0.003954	0.001211	0.005969	0.00191
percussive_mean	0.005383715	0.00153	0.001809	0.00306	0.005356	0.003672	0.004433	0.003288
log_mel_spec_min	0	0	0	0	0	0	0	0

TABLE 4.4.9:FEATURE IMPORTANCE VALUE FOR EACH EMOTIONS FOR M = 3

Feature-shape	**Angry**	**Calm**	**Disgust**	**Fearful**	**Happy**	**Neutral**	**Sad**	**Surprised**
mfcc_mean	0.150249935	0.126382	0.095449	0.096579	0.091454	0.109978	0.081838	0.07835
loudness	0.172194653	0.097075	0.047494	0.104021	0.135867	0.123029	0.099501	0.114658
log_mel_spec_mean	0.072817409	0.12849	0.175725	0.065241	0.131336	0.095165	0.066553	0.094927
mfcc_std	0.127585886	0.195656	0.210688	0.165523	0.146112	0.13488	0.191335	0.135734
zcr	0.089399778	0.084943	0.096755	0.166836	0.104125	0.103384	0.096673	0.093547
chroma_mean	0.018304075	0.013689	0.010882	0.012532	0.014976	0.02412	0.007604	0.015153
chroma_std	0.011632733	0.019805	0.024589	0.020297	0.009157	0.043762	0.016961	0.025912
log_mel_spec_std	0.015623218	0.013865	0.011586	0.022762	0.01663	0.010044	0.038106	0.021267
log_mel_spec_max	0.018595246	0.012173	0.02038	0.017415	0.016279	0.015193	0.022247	0.024145
spectral_rolloff_std	0.015215088	0.022731	0.014163	0.043037	0.017148	0.029739	0.03432	0.023019
spectral_rolloff_mean	0.023279336	0.013309	0.006331	0.024151	0.01893	0.025804	0.022435	0.028769
tempo	0.092341888	0.074884	0.104928	0.069984	0.10467	0.067694	0.107399	0.063001
spectral_contrast	0.054025431	0.045807	0.053696	0.066458	0.052989	0.064666	0.045832	0.08089
speaking_rate	0.048346616	0.06295	0.054469	0.050922	0.039173	0.060099	0.08283	0.089098
average_pitch	0.066864899	0.08229	0.059612	0.057089	0.074688	0.079504	0.063478	0.092879
harmonic_mean	0.006247386	0.001605	0.003108	0.004519	0.005225	0.001751	0.006134	0.002311
percussive_mean	0.017274985	0.004345	0.010077	0.012616	0.021228	0.011175	0.016774	0.016372
log_mel_spec_min	0	0	0	0	0	0	0	0

TABLE 4.4.10:FEATURE IMPORTANCE VALUE FOR EACH EMOTIONS FOR M = 4

Feature-shape	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
mfcc_mean	0.160203365	0.118169	0.098438	0.090926	0.100352	0.10639	0.083071	0.076196
loudness	0.18201741	0.118317	0.05144	0.122825	0.139005	0.149807	0.115387	0.136903
log_mel_spec_mean	0.06762814	0.121516	0.17104	0.062508	0.122901	0.088888	0.064212	0.091998
mfcc_std	0.12464471	0.188042	0.210152	0.157652	0.150855	0.121393	0.184878	0.140945
zcr	0.085203133	0.083132	0.093282	0.170179	0.105229	0.107299	0.097207	0.089977
chroma_mean	0.015018333	0.011593	0.009238	0.010757	0.012685	0.020534	0.006186	0.013087
chroma_std	0.009604096	0.017003	0.021702	0.017375	0.007892	0.037629	0.015058	0.021559
log_mel_spec_std	0.013245117	0.011928	0.009688	0.019365	0.014448	0.008409	0.032095	0.018144
log_mel_spec_max	0.017937124	0.012344	0.021275	0.017521	0.016822	0.016491	0.022687	0.025365
spectral_rolloff_std	0.012168621	0.018451	0.011611	0.034921	0.013918	0.024039	0.028152	0.019247
spectral_rolloff_mean	0.019724377	0.01144	0.005061	0.020569	0.016364	0.022299	0.018856	0.024777
tempo	0.093123684	0.07402	0.098122	0.067197	0.096965	0.074737	0.09504	0.063649
spectral_contrast	0.059051482	0.046882	0.05373	0.082123	0.06334	0.067563	0.043957	0.073641
speaking_rate	0.054257887	0.083461	0.072998	0.05195	0.041519	0.063692	0.103895	0.100777
average_pitch	0.061840973	0.077074	0.057467	0.05618	0.069197	0.076789	0.063857	0.082106
harmonic_mean	0.005929363	0.001791	0.003117	0.004471	0.005527	0.002003	0.006137	0.002913
percussive_mean	0.018390078	0.004851	0.011576	0.01348	0.02298	0.012049	0.019351	0.018719
log_mel_spec_min	0	0	0	0	0	0	0	0

TABLE 4.4.11:FEATURE IMPORTANCE VALUE FOR EACH EMOTIONS FOR M = 5

Feature-shape	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
mfcc_mean	0.171774953	0.115948	0.102069	0.086053	0.120623	0.104065	0.08367	0.071125
loudness	0.167543857	0.111758	0.048712	0.115545	0.126914	0.137871	0.115604	0.14018
log_mel_spec_mean	0.070240125	0.126594	0.174584	0.065677	0.126697	0.089401	0.066629	0.094043
mfcc_std	0.122501785	0.183297	0.204749	0.149485	0.148018	0.115384	0.179035	0.141346
zcr	0.08392	0.082483	0.092057	0.182315	0.104363	0.112966	0.097437	0.088276
chroma_mean	0.020446071	0.015062	0.011978	0.014114	0.017988	0.026986	0.009779	0.018041
chroma_std	0.010677277	0.019178	0.023799	0.020099	0.008944	0.041658	0.018839	0.024595
log_mel_spec_std	0.01279555	0.010872	0.009243	0.01806	0.013865	0.007484	0.030256	0.017903
log_mel_spec_max	0.017749455	0.011633	0.019729	0.017078	0.015809	0.015722	0.020777	0.024078
spectral_rolloff_std	0.010835608	0.01645	0.010504	0.030877	0.012274	0.021969	0.02513	0.017033
spectral_rolloff_mean	0.017552798	0.010213	0.004383	0.018183	0.01459	0.020481	0.016964	0.022172
tempo	0.101488813	0.07928	0.104434	0.068003	0.098666	0.075878	0.092179	0.062713
spectral_contrast	0.062261366	0.049697	0.054415	0.09242	0.065603	0.071102	0.044754	0.070247
speaking_rate	0.05586137	0.085178	0.07383	0.053225	0.041302	0.072192	0.118057	0.116505
average_pitch	0.059235661	0.083717	0.059602	0.060252	0.065699	0.081298	0.064211	0.079089
harmonic_mean	0.006218405	0.001896	0.003005	0.004117	0.005663	0.002309	0.006921	0.003245
percussive_mean	0.016515681	0.004369	0.010473	0.012125	0.020572	0.010854	0.017401	0.017042
log_mel_spec_min	0	0	0	0	0	0	0	0

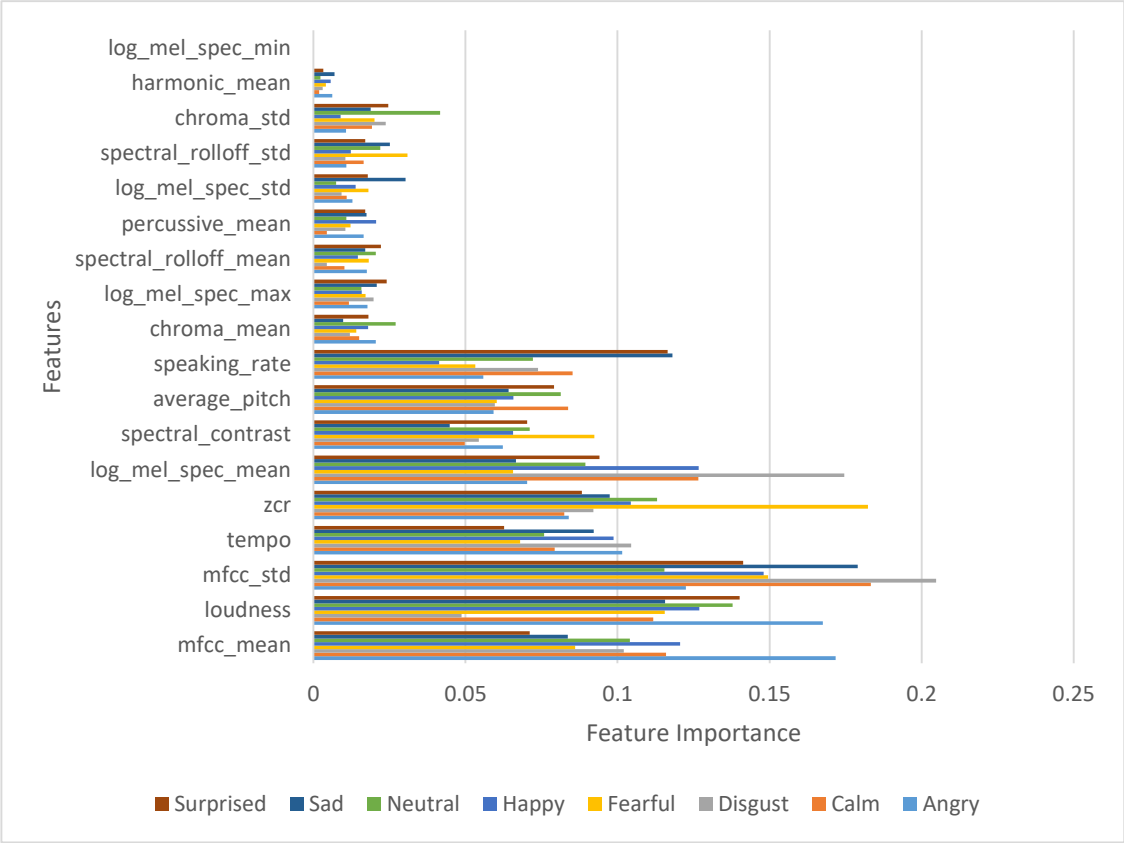


Figure 4.4.23: Feature impotence value for each emotions after M = 5 with AFI equation

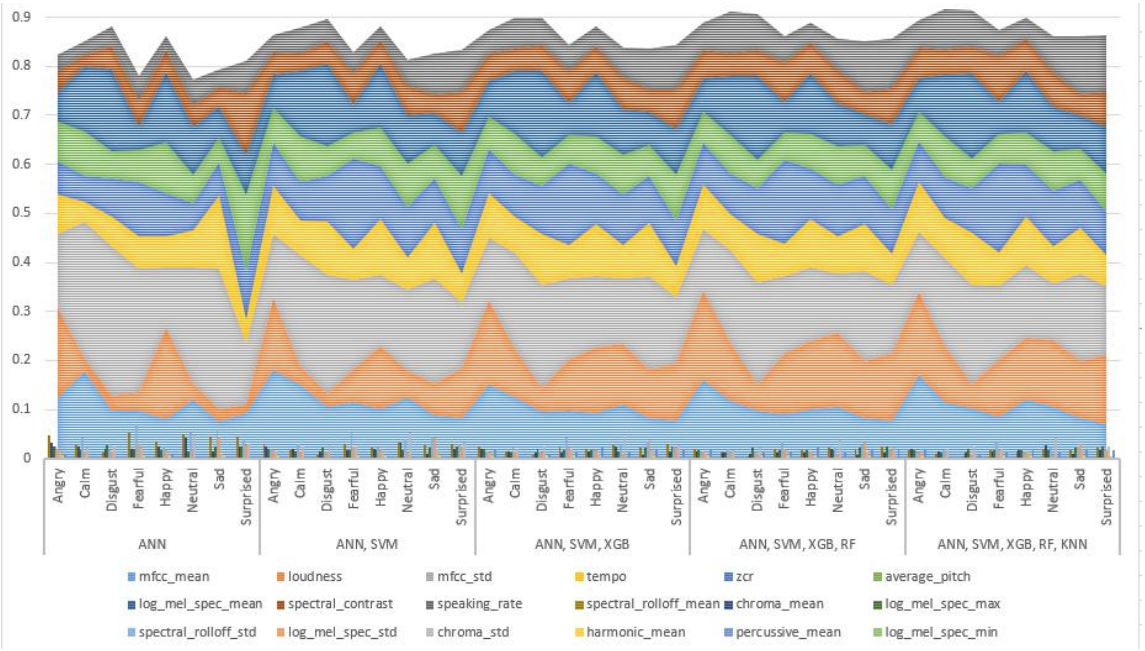


Figure 4.4.24: This visualization illustrates the convergence of feature importance values computed using the Aggregated Feature Importance (AFI) equation, as the number of models increases from M = 1 to M = 5.

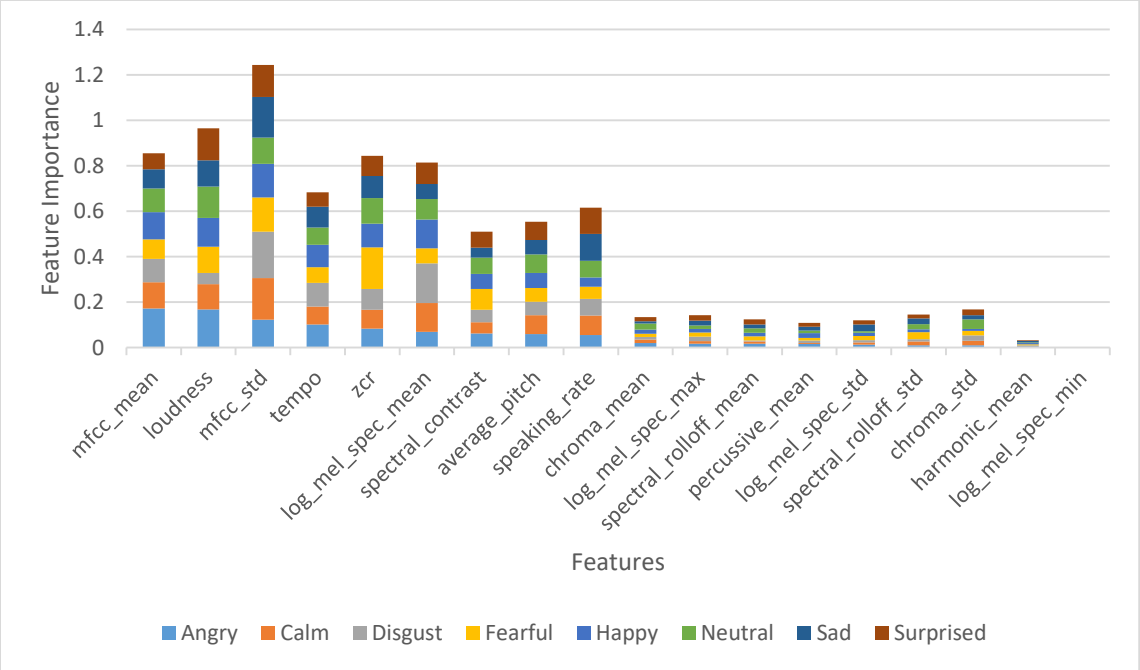


Figure 4.4.25: Global feature importance for emotional recognize

4.7 Top five feature for each emotions

TABLE 4.4.12: TOP FIVE FEATURES FOR EACH EMOTION AFTER APPLYING AFI WITH M=5

Emotion	Top five feature for each emotions
Angry	mfcc_mean, loudness, mfcc_std, tempo, zcr
Calm	mfcc_std, mfcc_mean, log_mel_spec_mean, loudness, speaking_rate
Disgust	mfcc_std, log_mel_spec_mean, tempo, mfcc_mean, zcr
Fearful	zcr, mfcc_std, mfcc_mean, loudness, spectral_contrast
Happy	mfcc_std, log_mel_spec_mean, loudness, mfcc_mean, zcr
Neutral	mfcc_std, mfcc_mean, zcr, log_mel_spec_mean, loudness
Sad	mfcc_std, tempo, zcr, loudness, speaking_rate
Surprised	mfcc_std, speaking_rate, loudness,zcr, log_mel_spec_mean

4.8 Value ranges for best features

TABLE 4.4.13: VALUE RANGES(MIN, MAX, MEAN, STD) AND T-TEST VALUE FOR TOP FIVE IMPORTANCE FEATURES

Feature	Min	Max	Mean	Std	t-value	Significant?
mfcc_mean	-43.2087	-26.2279	-34.2523	4.0339	-46.51	Yes
loudness	0.0042	0.0649	0.0323	0.0173	10.23	Yes
mfcc_std	27.2756	34.2794	30.6029	1.6145	103.82	Yes
tempo	70.3125	208.3333	126.3029	33.7496	20.5	Yes
zcr	0.0428	0.118	0.0722	0.0162	24.41	Yes
mfcc_std	23.1698	31.1249	28.179	1.5279	101.02	Yes
mfcc_mean	-62.4148	-37.423	-49.0575	6.0949	-44.09	Yes
log_mel_spec_mean	-68.255	-54.681	-62.4637	2.5394	-134.73	Yes
loudness	0.0005	0.0078	0.0022	0.0017	7.09	Yes
speaking_rate	0.222	1.6806	0.5527	0.4415	6.86	Yes
mfcc_std	26.9914	33.9059	30.4141	1.3751	121.14	Yes
log_mel_spec_mean	-68.1061	-56.9874	-63.5577	2.113	-164.75	Yes
tempo	64.6552	170.4545	106.3055	24.5836	23.68	Yes
mfcc_mean	-51.5448	-36.0332	-43.5389	4.0382	-59.05	Yes
zcr	0.0458	0.1132	0.0739	0.0135	29.98	Yes
zcr	0.0404	0.1217	0.0847	0.0213	21.78	Yes
mfcc_std	24.2947	30.6851	28.2879	1.3951	111.06	Yes
mfcc_mean	-48.3801	-24.0639	-38.8981	5.3465	-39.85	Yes
loudness	0.0028	0.0555	0.0207	0.0134	8.46	Yes
spectral_contrast	14.1919	19.22	17.0497	1.0608	88.03	Yes
mfcc_std	24.6763	32.5423	30.0642	1.4079	116.96	Yes
log_mel_spec_mean	-70.5923	-58.7817	-64.6686	2.2098	-160.29	Yes
loudness	0.0029	0.0337	0.0106	0.0077	7.54	Yes
mfcc_mean	-51.6278	-32.9075	-40.424	5.2889	-41.86	Yes

zcr	0.0361	0.1082	0.0609	0.0161	20.72	Yes
mfcc_std	27.8015	30.7789	29.529	0.7411	218.24	Yes
mfcc_mean	-56.3306	-38.7488	-48.2844	5.3516	-49.42	Yes
zcr	0.0277	0.0656	0.0521	0.0084	33.97	Yes
log_mel_spec_mean	-67.5405	-61.6654	-65.4067	1.1973	-299.21	Yes
loudness	0.0014	0.0051	0.0028	0.0011	13.94	Yes
mfcc_std	22.8156	31.906	28.0197	2.0765	73.91	Yes
tempo	80.3571	144.2308	106.3517	19.6244	29.68	Yes
zcr	0.0319	0.0842	0.0621	0.0131	25.96	Yes
loudness	0.0011	0.0076	0.0034	0.002	9.31	Yes
speaking_rate	0.2379	2.43	0.7551	0.5633	7.34	Yes
mfcc_std	25.6213	33.3562	30.8705	1.4533	116.35	Yes
speaking_rate	0.2827	2.4218	0.9277	0.5863	8.67	Yes
loudness	0.0017	0.0212	0.0078	0.0037	11.55	Yes
zcr	0.0355	0.1002	0.0667	0.0128	28.54	Yes
log_mel_spec_mean	-69.4568	-62.0508	-64.8746	1.602	-221.81	Yes

CHAPTER 5

DISCUSSION

In this research, a methodical approach was undertaken to construct an explainable speech emotion recognition (SER) system through the integration of a carefully curated set of audio features and high-performing classification models. The audio features were deliberately selected to encompass a broad range of speech characteristics, categorized into five principal groups: Acoustic, Temporal, Spectral, Harmonic/Tonal, and Spectral Energy Distribution. Each category was included based on its contribution to a comprehensive and interpretable understanding of the speech signal.

The Acoustic features, such as Loudness and RMS Energy, were incorporated due to their fundamental role in representing the intensity of the signal, where their perceptual association with emotional expressiveness and environmental dynamics has been previously acknowledged in speech analysis. These features, being intuitive in nature, facilitate a better understanding of variations in speech linked to emotional fluctuations.

Temporal features like Zero-Crossing Rate, Tempo, and Speaking Rate were instrumental in capturing the rhythmic and phonetic timing patterns embedded in human speech. Their utility in distinguishing between phonetic components and assessing fluency was found essential, particularly in applications related to emotion recognition and psychological assessment. The incorporation of these features enabled a deeper interpretation of the temporal dynamics underlying vocal expressions of emotion.

Spectral characteristics, which included Average Pitch, Spectral Contrast, and Mel-Frequency Cepstral Coefficients (MFCCs), served as key indicators of frequency domain attributes. Their role in identifying prosodic cues, emotional intonation, and speaker identity has been well-established, and this study reaffirmed their importance, particularly in differentiating nuanced emotional states such as anger, sadness, and surprise. MFCCs, in particular, emerged as pivotal for capturing spectral nuances tied to vocal tone and mood.

Harmonic and Tonal descriptors, including Chroma Features and the separation of Harmonic and Percussive Components, contributed to an enhanced understanding of speech tonality. These features proved valuable in segregating vowel-rich segments from transient consonants, thus aiding clarity in recognition tasks. Their inclusion was vital for improving speech intelligibility and interpretability in real-time scenarios.

Additionally, features representing the Spectral Energy Distribution, such as Spectral Rolloff and Log Mel Spectrogram statistics, were integrated to facilitate a perceptual

approximation of human auditory processing. These features allowed the system to distinguish between noise and tonal elements, reinforcing robustness against variations in speech quality and background noise.

In this study, the primary objective was not to maximize classification accuracy but to uncover consistent and interpretable patterns of feature importance for each emotional category. To achieve this, a diverse ensemble of classifiers including Artificial Neural Networks (ANN), Support Vector Machines (SVM), XGBoost, Random Forest, and k-Nearest Neighbors (k-NN) was employed. Feature attributions were derived from two model-agnostic explainability techniques, namely LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations). These attribution scores were aggregated using the Aggregated Feature Importance (AFI) method, which weighted each model's contribution based on its performance, enabling a balanced and robust measure of feature relevance.

The resulting feature importance scores, as visualized in **Figure 4.4.23**, revealed that features such as mfcc_mean, loudness, mfcc_std, and zcr played dominant roles across multiple emotional classes. These features are well-established in the literature for capturing both spectral and temporal aspects of speech, and their recurrence across emotions reinforces their general importance. **Figure 4.4.24** further supports this finding by showing how feature rankings stabilize as the number of models (M) increases from 1 to 5, confirming the consistency of AFI-derived rankings. The convergence trend illustrates that as more classifiers contribute to the ensemble, the resulting importance scores become increasingly reliable and less susceptible to model-specific fluctuations.

A global comparison of feature importance across all emotion classes is shown in **Figure 4.4.25**. While certain features such as mfcc_std, tempo, and log_mel_spec_mean demonstrate high global relevance, others are more emotion-specific. **Table 4.12** summarizes the top five features for each emotional class after applying AFI with M = 5. These emotion-wise results reflect how each emotion is characterized by a unique set of features. For instance, anger is primarily described by a combination of acoustic and temporal features including mfcc_mean, loudness, mfcc_std, tempo, and zcr. Calm is marked by smoother and softer characteristics such as low loudness, speaking_rate, and log_mel_spec_mean. Disgust and happy emotions prominently involve mfcc_std and log_mel_spec_mean, suggesting strong spectral energy dynamics. Emotions like fearful and sad emphasize abrupt signal changes and slower pacing, captured by high zcr, mfcc_std, and tempo. Surprised is characterized by rapid articulation and abrupt vocal energy, reflected in the high importance of speaking_rate, zcr, and loudness.

To strengthen the interpretability results, statistical validation was carried out using independent t-tests for the top-ranked features of each emotion. The results demonstrated

that all selected features showed statistically significant differences ($p < 0.05$) across emotional categories. This confirms their discriminative power and reinforces the validity of AFI rankings. Features such as mfcc_std, log_mel_spec_mean, and zcr not only scored high in AFI analysis but also achieved large t-values in statistical testing, underscoring their robustness in separating emotions such as Disgust, Happy, and Fearful.

The combination of ensemble-based interpretability and statistical significance analysis provides a strong foundation for building explainable speech emotion recognition systems. Rather than depending solely on classification metrics, this approach prioritizes transparency, stability, and domain relevance. The consistent identification of core features across both AFI and statistical analysis highlights their practical utility in real-world applications such as affective computing, mental health screening, and forensic speech analysis. Beyond mere classification accuracy, the integration of Explainable AI (XAI) introduces a structured understanding of how individual acoustic and spectral features contribute to emotional inference. By leveraging model-agnostic techniques such as SHAP and LIME, the system provides clear insights into the relationships between speech characteristics such as MFCCs, loudness, and zero-crossing rate and the corresponding emotional states. This transparency ensures that users and practitioners can trace and interpret model decisions, enhancing trust and accountability in applications where human comprehension of AI output is essential.

Furthermore, the proposed XAI-based feature-value matrix adds an additional layer of practical significance. By defining the value ranges for each critical feature, the framework quantifies how variations in acoustic and spectral properties correspond to specific emotions, bridging the gap between computational predictions and human-understandable cues. This human-centered interpretability allows clinicians, educators, and system designers to not only recognize which features influence emotion recognition but also to understand the magnitude and direction of their effect, making the AI system actionable and meaningful in practice. Overall, this approach demonstrates that the value of SER systems extends beyond predictive performance; it creates transparent, reliable, and interpretable models that can be responsibly applied in diverse domains requiring emotion-sensitive decision-making.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This research presents a significant advancement in the field of Explainable Artificial Intelligence (XAI) for Speech Emotion Recognition (SER), with a purposeful emphasis on interpretability rather than mere classification accuracy. By prioritizing feature-level transparency, the study moves away from conventional deep learning pipelines and instead utilizes traditional machine learning techniques applied to a carefully selected set of handcrafted acoustic features extracted from raw mp3 audio signals.

A total of 18 interpretable features including loudness, zero-crossing rate (ZCR), tempo, speaking rate, pitch-related statistics, spectral contrast, MFCCs, and other harmonic descriptors were systematically extracted and grouped into five major functional categories. These features served as the foundation for building an explainable feature-emotion relationship matrix spanning seven emotional classes: Calm, Happy, Sad, Angry, Fearful, Surprised, and Disgust.

To determine the relevance of each feature in distinguishing emotional states, two model-agnostic interpretability techniques SHAP and LIME were applied across a diverse ensemble of classifiers: Artificial Neural Networks (ANN), Support Vector Machines (SVM), XGBoost, Random Forests, and k-Nearest Neighbors (k-NN). A novel metric, Aggregated Feature Importance (AFI), was introduced to synthesize these interpretability outputs by weighting them according to the predictive accuracy of each model. This resulted in a unified and interpretable feature importance score for each emotion.

A major contribution of this work lies in the convergence analysis of feature importance scores. As the number of models (M) in the ensemble increased from 1 to 5, the rankings of critical features across emotions became progressively more stable. This stability indicates a strong consensus among different models about the most relevant features, such as mfcc_std, mfcc_mean, loudness, and zcr, which appeared consistently across most emotional classes. The robustness of this convergence reduces reliance on any one model's behavior and supports the reliability of the interpretability framework.

Furthermore, a complementary statistical analysis was performed on the top-ranked features for each emotion using descriptive metrics (mean, standard deviation, min, max) and independent t-tests. The results confirmed statistically significant differences ($p < 0.05$) between emotional states for most key features, reinforcing their discriminative power. For instance, log_mel_spec_mean and mfcc_std stood out as particularly strong indicators for emotions such as Disgust, Fearful, and Happy.

The final explainable feature-emotion map (Figure 4.4.25 and Table 4.12) is not only analytically sound but also practically valuable, offering human-understandable insights into which features matter most for each emotion. This structure supports transparency and trust, which is crucial for applying SER in domains like healthcare, psychology, education, and forensics.

Future research will explore two main directions. First, the development of a reverse-engineered framework where known emotional outputs can guide the reconstruction of approximate input feature profiles enabling emotion-guided audio synthesis and deeper insight into emotional composition. Second, the integration of temporal dynamics through sequence-based models such as LSTMs, not for improving accuracy, but for introducing time-aware explanations that preserve interpretability while reflecting the natural flow of speech. Together, these directions aim to enrich the proposed explainability paradigm with both generative and temporal dimensions, contributing further to trustworthy and human-centered AI systems in emotion recognition.

REFERENCES

[1] K. P. Exarchos *et al.*, "Review of Artificial Intelligence Techniques in Chronic Obstructive Lung Disease," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 5, pp. 2331–2338, 2022, doi: 10.1109/JBHI.2021.3135838.

[2] A. Saranya and R. Subhashini, "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends," *Decis. Anal. J.*, vol. 7, no. April, p. 100230, 2023, doi: 10.1016/j.dajour.2023.100230.

[3] F. Shi *et al.*, "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19.," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 4–15, 2021, doi: 10.1109/RBME.2020.2987975.

[4] R. Sheh and I. Monteath, "Defining Explainable AI for Requirements Analysis," *KI - Künstliche Intelligenz*, vol. 32, no. 4, pp. 261–266, 2018, doi: 10.1007/s13218-018-0559-3.

[5] R. Guidotti, A. Monreale, D. Pedreschi, and F. Giannotti, "Principles of explainable artificial intelligence," *Explain. AI Within Digit. Transform. Cyber Phys. Syst. XAI Methods Appl.*, pp. 9–31, 2021, doi: 10.1007/978-3-030-76409-8_2.

[6] K. Zhang, Y. Zhang, and M. Wang, "A Unified Approach to Interpreting Model Predictions Scott," *Nips*, vol. 16, no. 3, pp. 426–430, 2012.

[7] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 150–158, 2012, doi: 10.1145/2339530.2339556.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," *NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess.*, pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.

[9] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," *Adv. Neural Inf. Process. Syst.*, no. Nips, pp. 3512–3520, 2016.

[10] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 3, pp. 2048–2057, 2015.

[11] R. Ullah *et al.*, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer," *Sensors*, vol. 23, no. 13, pp. 1–20, 2023, doi: 10.3390/s23136212.

[12] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020, doi: 10.1016/j.specom.2019.12.001.

[13] A. Chaturvedi and A. Tripathi, "Emotion Recognition using Fuzzy Rule- based System," *Int. J. Comput. Appl.*, vol. 93, pp. 25–28, 2014, doi: 10.5120/16260-5920.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in

openSMILE, the munich open-source multimedia feature extractor," *MM 2013 - Proc. 2013 ACM Multimed. Conf.*, no. May, pp. 835–838, 2013, doi: 10.1145/2502081.2502224.
[15] C. Sun, H. Li, and L. Ma, "Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network," *Front. Psychol.*, vol. 13, no. January, pp. 1–14, 2023, doi: 10.3389/fpsyg.2022.1075624.
[16] J. Singh, L. B. Saheer, and O. Faust, "Speech Emotion Recognition Using Attention Model," *Int. J. Environ. Res. Public Health*, vol. 20, no. 6, 2023, doi: 10.3390/ijerph20065140.
[17] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, 2015, doi: 10.1109/TAFFC.2014.2334294.
[18] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Inf. Fusion*, vol. 99, no. January, p. 101805, 2023, doi: 10.1016/j.inffus.2023.101805.
[19] N. A. Malk and S. A. Diwan, "Artificial Intelligence in Speech Emotion Detection : Trends , Challenges , and Future Directions," pp. 19–29.
[20] I. Gurowiec and N. Nissim, "Improving speech emotion recognition capabilities in the short and long term using temporal bucketing and active learning," *Comput. Biol. Med.*, vol. 196, p. 110912, 2025, doi: https://doi.org/10.1016/j.compbiomed.2025.110912.
[21] M. Norval and Z. Wang, "Explainable Artificial Intelligence Techniques for Speech Emotion Recognition: A Focus on XAI Models," *Intel. Artif.*, vol. 28, no. 76, pp. 85–123, 2025, doi: 10.4114/intartif.vol28iss76pp85-123.
[22] T. Leygue, A. Sabourin, C. Bolzmacher, S. Bouchigny, M. Anastassova, and Q.-C. Pham, "Explainable speech emotion recognition through attentive pooling: insights from attention-based temporal localization," 2025, [Online]. Available: http://arxiv.org/abs/2506.15754.
[23] A. Nfissi, W. Bouachir, N. Bouguila, and B. Mishara, "Unveiling hidden factors: explainable AI for feature boosting in speech emotion recognition," *Appl. Intell.*, vol. 54, no. 11–12, pp. 7046–7069, 2024, doi: 10.1007/s10489-024-05536-5.
[24] N. Sripriya, S. Subramanian, and S. Jagathisan, "Multimodal Emotional Analysis using XAI for Psychotherapy," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 5, pp. 1058–1067, 2025, doi: 10.38124/ijisrt/25may474.
[25] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, pp. 551–555, 2021, doi: 10.21437/Interspeech.2021-703.
[26] W. Zhang and B. Y. Lim, "Towards Relatable Explainable AI with the Perceptual Process," *Conf. Hum. Factors Comput. Syst. - Proc.*, 2022, doi: 10.1145/3491102.3501826.
[27] S. R. Livingstone and F. A. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. 2018.
[28] S. W. Byun and S. P. Lee, "A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms," *Appl. Sci.*, vol. 11, no. 4, pp. 1–15, 2021, doi: 10.3390/app11041890.
[29] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal," *Adv.*

Tech. Comput. Sci. Softw. Eng., no. November, pp. 279–282, 2010, doi: 10.18260/1-2-1153-53698.
[30] M. Alonso, B. David, and G. Richard, "Tempo and Beat Estimation of Musical Signals," *ISMIR 2004 - 5th Int. Symp. Music Inf. Retr.*, no. May 2014, 2004.
[31] M. Baese-Berk and T. Morrill, "Speaking rate consistency in native and non-native speakers of English," *J. Acoust. Soc. Am.*, vol. 136, p. 2146, Oct. 2014, doi: 10.1121/1.4899749.
[32] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," *Odyssey 2010 Speak. Lang. Recognit. Work.*, no. June, pp. 34–39, 2010.
[33] J. M. Hillenbrand, R. A. Houde, T. M. Nearey, and M. J. Clark, "Vowel recognition from harmonic spectra," *J. Acoust. Soc. Am.*, vol. 105, no. 2_Supplement, pp. 1396–1396, 1999, doi: 10.1121/1.426597.
[34] S. Majeed, H. HUSAIN, S. Samad, and T. Idbeaa, "Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: A comparison study," *J. Theor. Appl. Inf. Technol.*, vol. 79, pp. 38–56, Sep. 2015.
[35] N. Ono *et al.*, "Harmonic and Percussive Sound Separation and Its Application to MIR-Related Tasks," in *Studies in Computational Intelligence*, vol. 274, 2010, pp. 213–236.
[36] J. Beauchamp, "Analysis and Synthesis of Musical Instrument Sounds," 2007, pp. 1–89.
[37] K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network," 2022, pp. 241–250.