Emotion Recognition on Scenes of films based on the speech and the image

by

Eleftherios Tzagkarakis

Submitted
in partial fulfilment of the requirements for the degree of

Master of Artificial Intelligence

at the

UNIVERSITY OF PIRAEUS

December 2023

Author . . . . Eleftherios Tzagkarakis. . . . .

II-MSc "Artificial Intelligence"

December, 2023

Certified by. . . . . Ilias Maglogiannis. . . . .

Professor
Thesis Supervisor

Certified by. . . . . Michael Filippakis . . . .

Professor
/Academic Title
Member of Examination Committee

Certified by. . . . . Maria Halkidi . . . .

Associate Professor
Member of Examination Committee

Emotion Recognition on Scenes of films based on the speech and the image.

By

Eleftherios Tzagkarakis

Submitted to the II-MSc “Artificial Intelligence” in December 2023,
in partial fulfilment of the
requirements for the MSc degree

Abstract

This thesis delves into the fascinating realm of experimentation and evaluation, exploring a diverse array of machine learning models applied to both the auditory and visual domains. Specifically, the focus is on emotion recognition within public datasets comprising photographs and speech excerpts. The research progresses to the discernment of optimal models, which are subsequently deployed on cinematic scenes featuring monologues. This allows for a comprehensive comparison of the outcomes produced by these two models, scrutinizing the consistency and correlation of their predictions.

The ultimate objective of this endeavour is to fashion an intelligent director, empowered by the capabilities of machine learning. This directorial intelligence extends beyond conventional boundaries, making decisions on whether a scene warrants a reiteration, particularly when the results of the two models exhibit disparities. The implementation of this groundbreaking approach integrates the training of open-source neural networks alongside the utilization of classical machine learning algorithms.

This multifaceted exploration underscores the fusion of innovative technologies and traditional methodologies, establishing a robust framework for the advancement of intelligent cinematic direction. The synergy between open-source neural networks and classical machine learning algorithms not only contributes to the evolution of film production methodologies but also charts new territories in the intersection of artificial intelligence and artistic expression.

Thesis Supervisor: Ilias Maglogiannis
Title: Professor

Acknowledgments

I extend my heartfelt gratitude to my esteemed professors, whose guidance and expertise have been invaluable throughout the journey of this thesis. Their mentorship has not only enriched my academic experience but has also shaped the depth and quality of this research.

A warm appreciation goes out to my co-students, whose collaborative spirit and shared enthusiasm have created an environment conducive to learning and innovation. The exchange of ideas and collective effort has undoubtedly played a crucial role in the development of this work.

I am profoundly grateful to my family for their unwavering support and encouragement. Their belief in my abilities has been a constant source of motivation, providing the foundation upon which this thesis stands.

To my friends, who have been a pillar of strength and a source of inspiration, I extend my deepest thanks. Their encouragement and camaraderie have been instrumental in navigating the challenges and celebrating the milestones of this academic endeavour.

This thesis stands as a testament to the collective support and encouragement of my professors, co-students, family, and friends. I am truly fortunate to have such a remarkable network of individuals who have contributed to the realization of this scholarly pursuit.

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the «funding body» or the view of University of Piraeus and Inst. of Informatics and Telecom. of NCSR “Demokritos”.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Definition

The fundamental challenge addressed in this thesis lies in gauging the fidelity of cinematic scenes by scrutinizing the congruence between facial expressions and vocal emotions. The primary aim is to assess the precision with which scenes are captured and to ascertain the alignment between the emotional portrayal in facial expressions and the corresponding emotional tonality in the voice. To achieve this, the research endeavours to develop an intelligent director employing two distinct classifiers—one dedicated to image classification and the other to speech classification. Through this approach, the study seeks to determine the effectiveness of these classifiers in collectively appraising the authenticity of film scenes, thereby providing filmmakers with valuable insights into the harmony between visual and auditory emotional cues.

1.2 Scope

The scope of this research extends to the exploration of machine learning models for emotion recognition, specifically applied to both visual and auditory data. The project delves into the intricacies of image and speech processing to accurately capture and classify emotions exhibited by actors in film scenes. The focus is not only on the individual performance of the classifiers but also on examining the correlation between their outputs. The goal is to determine whether the integration of these classifiers contributes to the effective evaluation of cinematic scenes, thereby empowering filmmakers with the assistance of artificial intelligence.

2 Theoretical Framework

In this chapter, an in-depth analysis of the theoretical underpinnings that form the basis for the experiments conducted in this thesis is presented. The exploration begins with a comprehensive examination of Artificial Intelligence (AI), including its historical evolution and various types that have shaped its trajectory. Subsequent sections delve into essential concepts in Machine Learning, Classification, Neural Networks, Image Classification. Each section provides a foundational understanding crucial for comprehending the methodologies employed in the subsequent chapters.

2.1 Artificial Intelligence (AI)

2.1.1 Definition and Evolution

Artificial Intelligence (AI) stands at the intersection of computer science, mathematics, and cognitive science. Its primary goal is to create machines that can mimic human intelligence and perform tasks that typically require human cognitive functions such as learning, problem-solving, and decision-making [1].

The term "Artificial Intelligence" was first coined by John McCarthy during the Dartmouth Conference in 1956 [2]. This seminal event marked the birth of AI as an academic field, attracting researchers from various disciplines who sought to explore the possibilities of creating intelligent machines.

The evolution of AI can be traced through several distinct phases. Early AI research focused on rule-based systems and symbolic reasoning. However, limitations in handling real-world complexity led to the emergence of machine learning paradigms, where systems could learn from data.

2.1.2 Types of AI

AI can be categorized into two main types: Narrow or Weak AI [3], [4] and General or Strong AI. Narrow AI is designed to perform a specific task or solve a particular problem, exhibiting intelligence only in that specific domain. Examples include virtual personal assistants like Siri and Alexa.

In contrast, General AI aspires to possess human-like intelligence across a broad range of tasks. Achieving General AI remains a long-term goal and involves creating systems that can adapt and learn across various domains [5].

The contemporary landscape of AI is dominated by narrow AI applications, which have demonstrated remarkable success in tasks such as image and speech recognition, natural language processing, and game playing [6].

2.2 Machine Learning (ML)

Machine Learning (ML) is a foundational concept in artificial intelligence, enabling computers to learn from data and make decisions or predictions without explicit programming. In ML, various tasks include classification, regression, and clustering.

2.2.1 Classification

Overview:

Classification is a supervised learning task where the algorithm assigns predefined labels to input data based on patterns learned during training. This task is fundamental in various applications, such as spam detection, image recognition, and sentiment analysis.

Algorithms:

- Decision Trees: Create a tree-like model of decisions based on features [7], [8].
- Random Forest: Ensemble of decision trees, offering improved accuracy and robustness [9].
- Support Vector Machines (SVM): Classify data points by finding the optimal hyperplane [10].

2.2.2 Regression

Overview:

Regression is a supervised learning task where the algorithm predicts a continuous output variable based on input features. Commonly used in finance, economics, and weather forecasting, regression helps model relationships between variables.

Algorithms:

- Linear Regression: Establishes a linear relationship between input and output variables [11].
- Lasso and Ridge Regression: Introduces regularization to prevent overfitting [12], [13].
- Decision Trees for Regression: Extends decision trees to predict continuous values [8].

2.2.3 Clustering

Overview:

Clustering is an unsupervised learning task that groups similar data points together based on certain features. Applications include customer segmentation and anomaly detection.

Algorithms:

- K-Means: Divides data into k clusters based on centroids [14].
- Hierarchical Clustering: Forms a tree of clusters, useful for visualizing relationships [15], [16].
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Groups dense areas and identifies outliers [17].

2.2.4 Other Aspects

While classification, regression, and clustering are fundamental, ML encompasses various other aspects. These include dimensionality reduction techniques, ensemble methods, and reinforcement learning, each serving specific purposes in diverse applications.

Algorithms:

- Principal Component Analysis (PCA): Reduces dimensionality while retaining data variance [18], [19].
- Random Forests: Ensemble method combining multiple decision trees [20].
- Q-Learning: A reinforcement learning algorithm for making sequential decisions [21].

2.3 Deep Learning (DL)

Deep Learning represents a subset of machine learning where artificial neural networks, inspired by the human brain's structure, are employed to model, and solve complex problems. The architecture involves interconnected nodes organized into layers, allowing for the learning of intricate hierarchical representations [22].

Deep Learning has gained prominence for its ability to automatically learn and extract features from raw data, eliminating the need for manual feature engineering. The learning process involves the adjustment of weights in neural network connections through backpropagation, optimizing the model's performance.

Algorithms and Architectures:

- Convolutional Neural Networks (CNNs): Effective for image and video analysis [23].
- Recurrent Neural Networks (RNNs): Suited for sequential data, such as time series or natural language [24].
- Long Short-Term Memory Networks (LSTMs): A type of RNN, ideal for capturing long-range dependencies [25], [26].

2.3.1 Image Classification

2.3.1.1 Overview

Image classification is a fundamental task in computer vision, involving the assignment of predefined labels or categories to images. This process is essential for automated systems to recognize and interpret visual information in various applications.

2.3.1.2 Methods and Techniques

Various methods and techniques have been employed in image classification. Traditional machine learning approaches, including Support Vector Machines (SVM) and Random Forests, have been complemented and, in many cases, surpassed by deep learning methodologies, particularly Convolutional Neural Networks (CNNs) [27], [28].

2.3.1.3 Challenges in Image Classification

Despite its widespread use, image classification faces several challenges:

- Variability in Visual Appearance: Images can vary due to factors like lighting conditions, viewpoint changes, and object orientation, making it challenging for models to generalize effectively [29].
- Large-Scale Datasets and Training Complexity: Acquiring and managing large-scale labelled datasets for training image classification models can be resource intensive. Additionally, the computational complexity of training deep neural networks presents challenges [30].
- Fine-Grained Classification: Distinguishing between visually similar categories requires models to capture subtle differences in features, adding complexity to the classification task [31].
- Adversarial Attacks: Image classification models are vulnerable to adversarial attacks, where small, imperceptible perturbations to input images can lead to misclassifications [32].

- Interpretability and Explainability: In critical applications, such as healthcare and autonomous systems, the interpretability and explainability of image classification models are crucial for trust and reliability [33], [34].

2.3.1.4 Importance in Various Fields

Image classification holds significant importance across diverse domains due to its ability to interpret visual information. Specific applications include:

- Healthcare: Image classification contributes to medical diagnostics through tasks like identifying tumours in medical images [35].
- Autonomous Vehicles: Image classification is crucial for object detection and recognition in the development of autonomous vehicles, enhancing their ability to navigate and make informed decisions [36], [37].
- Agriculture: Image classification aids in crop monitoring and disease detection, contributing to precision agriculture for improved yield and resource management [38].
- Security and Surveillance: Image classification enhances security and surveillance systems by identifying and tracking objects or individuals in real-time [39].
- Retail and E-commerce: Image classification is employed for product recognition, recommendation systems, and inventory management, improving the efficiency of retail operations [40].

2.3.1.5 ImageNet Dataset

Overview:

ImageNet is a massive dataset comprising millions of labelled images across thousands of classes. Its significance lies in its role as a benchmark for evaluating and advancing image classification algorithms. ImageNet has played a pivotal role in driving breakthroughs in computer vision.

Algorithms Developed:

1. AlexNet (2012): Developed by Alex Krizhevsky and his team, AlexNet was the pioneering deep neural network architecture that demonstrated the effectiveness of deep learning for image classification. It significantly outperformed traditional methods, winning the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [41].
2. GoogLeNet (2014): Also known as Inception v1, GoogLeNet introduced the concept of inception modules, enabling the network to efficiently capture features at multiple scales. It won the ILSVRC 2014 competition [42].
3. ResNet (2015): ResNet, or Residual Network, addressed the vanishing gradient problem by introducing skip connections. This architecture enabled the training of extremely deep networks, leading to unprecedented accuracy in image classification [43].

2.3.1.6 Impact of ImageNet on Computer Vision

The success of ImageNet and the corresponding algorithms has had a profound impact on the field of computer vision. These models, and their subsequent iterations, have become foundational in various applications, from image and video analysis to object detection and segmentation [29], [30].

Figure 1: ImageNet

2.3.2 Audio Classification

2.3.2.1 Overview

Audio classification is a fundamental task in the field of signal processing and machine learning, aimed at automatically categorizing audio signals into predefined classes or categories [44]. This process enables the extraction of meaningful information from vast amounts of audio data, with applications spanning various domains such as speech recognition, music genre classification, emotion recognition, environmental sound analysis, and more.

2.3.2.2 Types of Audio Signals

Audio signals encompass a diverse range, including but not limited to:

- Speech: Human speech is a crucial component of communication systems and voice-activated technologies.
- Music: Classification of music genres aids in content recommendation and organization in music streaming services [45].
- Environmental Sounds: Recognizing sounds from the environment, such as sirens, footsteps, or bird calls, has applications in surveillance and environmental monitoring [46].

2.3.2.3 Challenges in Audio Classification

Despite its significance, audio classification poses several challenges, including:

- Variability: Audio signals can exhibit substantial variability due to factors like pitch, tempo, and background noise [44].
- Feature Extraction: Selecting relevant features from audio signals is a critical step, requiring careful consideration of the characteristics of different signal types [45].
- Data Imbalance: Imbalanced datasets, where certain classes have fewer examples, can affect the model's ability to generalize [46].

2.3.2.4 Importance in Various Fields

The application of audio classification extends across numerous domains, and specific examples are outlined below:

- Healthcare: Identifying medical conditions through speech analysis or monitoring patient well-being through audio cues [44].
- Entertainment: Improving content recommendation systems and enhancing user experience in gaming and virtual reality [45].
- Security: Detecting abnormal sounds for security and surveillance purposes [46].

3 Related work

3.1 Introduction

Emotion recognition, a captivating interdisciplinary field, intersects with the realms of psychology, artificial intelligence, and multimedia processing. As human-computer interaction becomes increasingly nuanced, the ability to comprehend and respond to human emotions is a pivotal aspect of technological advancement. This chapter embarks on a comprehensive exploration of existing literature, seeking to unravel the intricate tapestry of emotion recognition, particularly within the captivating context of films.

Understanding emotions has been a fundamental quest for scholars across disciplines, and the emergence of advanced technologies has propelled emotion recognition into the forefront of research endeavours. This chapter delves into the rich landscape of emotion recognition, dissecting the evolution of methodologies and paradigms that have shaped our understanding of emotional cues.

Through an examination of seminal works, landmark studies, and contemporary research efforts, this chapter illuminates the critical milestones in emotion recognition. As we embark on this journey, the convergence of speech and image modalities stands as a central theme, reflecting the increasing recognition of the multi-dimensional nature of human expression. This exploration extends into the specific domain of films, where emotions manifest uniquely, challenging researchers to develop sophisticated models capable of capturing the essence of cinematic emotional landscapes.

The synthesis of past achievements, ongoing endeavours, and the distinct challenges posed by the film domain will serve as a foundation for the subsequent chapters. As we navigate through the intricacies of related work, a comprehensive understanding of the existing landscape will pave the way for the novel contributions and methodologies presented in this thesis.

3.2 Emotion Recognition in Images

Visual cues play a pivotal role in conveying emotions, and the application of emotion recognition in images has emerged as a transformative field, influencing various domains, including computer vision, psychology, and filmmaking.

3.2.1 Early Foundations

Pioneering research, exemplified by Ekman and Friesen's (1971) work on "Constants across Cultures in the Face and Emotion" [47], focused on the universality of facial expressions. This study and others in the early phase established the role of facial expressions as a key element in image emotion recognition.

The work of Russell (1980) in "A Circumplex Model of Affect" [48] introduced a circumplex model of affect, emphasizing the diverse range of emotional signals within visual stimuli. Understanding this diversity is fundamental for developing comprehensive image emotion recognition models.

3.2.2 Machine Learning Approaches in Image Emotion Recognition

The analysis of facial expressions remains a cornerstone in image emotion recognition. Studies like Lucey et al.'s (2010) "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression" [49], featuring the Extended Cohn-Kanade dataset, provide a comprehensive resource for facial expression analysis and emotion-specific expression recognition.

3.2.3 Facial Expression Analysis with Deep Learning

Deep learning has revolutionized facial expression analysis in image emotion recognition. Studies like Liu et al.'s (2017) "Deep Learning for Generic Object Detection: A Survey" [50] showcase the effectiveness of deep neural networks in capturing intricate facial features, contributing to improved accuracy in recognizing emotions.

Recognizing that emotions extend beyond facial expressions; multimodal deep learning approaches have evolved to consider other visual cues. The work of Zhang et al. (2020) in "Deep Learning for Emotion Recognition: A Comparative Review of Recent Advances"

[51] provides a comprehensive review of recent advances in deep learning for emotion recognition, including the integration of diverse visual cues.

3.3 Emotion Recognition in Speech

Speech, as a fundamental mode of human expression, harbours a wealth of emotional information encoded in acoustic features and linguistic nuances. The exploration of emotion recognition in speech has traversed a multifaceted journey, marked by key studies and technological advancements.

3.3.1 Early Foundations

In the early stages of speech emotion recognition, pioneering studies by Scherer (2003) laid the groundwork for understanding the vocal communication of emotion [52]. Scherer's work highlighted the significance of acoustic cues and prosody in conveying emotional states. This foundational research spurred subsequent investigations into the acoustic features indicative of specific emotions, providing a baseline for the development of automated recognition systems.

3.3.2 Machine Learning Approaches

The shift towards machine learning approaches marked a turning point in speech emotion recognition. Notably, alongside Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs) have played a pivotal role in capturing the statistical distribution of speech features related to emotions. A significant study by Schuller et al. (2011) titled "The INTERSPEECH 2011 Speaker State Challenge" showcased the effectiveness of both GMMs and SVMs in modelling dynamic changes in emotion states using speech data [53]. These machine learning works offer complementary approaches, with GMMs providing a probabilistic framework for capturing complex emotion patterns, and SVMs excelling in discerning intricate patterns in speech data. Additionally, the influential work by Vapnik and Cortes (1995) titled "Support-Vector Networks" significantly influenced the adoption of SVMs in various machine learning applications, including speech emotion recognition [54]. Together, these paradigms have contributed significantly to automating the intricate task of decoding emotions from speech signals.

3.3.3 Deep Learning Advancements

The advent of deep learning ushered in a new era for speech emotion recognition. Deng et al.'s (2013) work on "Recent Advances in Deep Learning for Speech Research at Microsoft" demonstrated the effectiveness of deep neural networks (DNNs) in capturing complex hierarchical representations of speech features [55]. DNNs, with their capacity to automatically learn intricate patterns, have become instrumental in enhancing the accuracy and robustness of speech emotion recognition systems. The current state of the art is characterized by the integration of advanced deep learning architectures, particularly recurrent neural networks (RNNs) and long short-term memory networks (LSTMs). Recent studies, such as Trigeorgis et al.'s (2016) exploration of "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network" [56], showcase the effectiveness of end-to-end architectures in capturing both temporal and spectral dependencies. This state-of-the-art integration underlines the significance of end-to-end solutions, minimizing the need for handcrafted features and maximizing the potential for nuanced emotion decoding.

The journey through emotion recognition in speech encompasses seminal works, ranging from foundational research on acoustic cues to the integration of sophisticated machine learning and deep learning approaches. This evolution sets the stage for the subsequent exploration of emotion recognition in images and the fusion of modalities in the context of films.

3.4 Multimodal Approaches

Emotion recognition is inherently multimodal, and the integration of various sensory modalities, such as speech, images, and potentially other modalities like physiological signals, has garnered significant attention. Multimodal approaches offer a holistic understanding of human emotions, capturing the nuances that arise from the combination of different expressive channels.

3.4.1 Speech-Image Fusion for Comprehensive Emotion Understanding

Combining speech and image modalities provides a synergistic approach to emotion recognition, enriching the depth and accuracy of emotional understanding.

Studies such as Huang et al.'s (2019) "AVLnet: Learning Audio-Visual Language Representations from Instructional Videos" [57] exemplify the fusion of audio and visual information for improved emotion representation. Integrating both speech and image features enhances the robustness of emotion recognition systems, particularly in dynamic and complex scenarios.

Ensuring temporal alignment between speech and image cues is crucial for coherent emotion interpretation. Research, such as the work by Han et al. (2020) on "Mutual Guidance for Cross-Modality Emotion Recognition" [58], explores mutual guidance mechanisms to align temporal dynamics in speech and image data, contributing to more accurate cross-modal emotion recognition.

3.4.2 Physiological Signals Integration

Incorporating physiological signals, such as heart rate or skin conductance, alongside speech and image modalities provides a more comprehensive view of emotional states.

Studies like Chanel et al.'s (2009) "Emotion Assessment: Arousal Evaluation Using EEG's and Peripheral Physiological Signals" [59] showcase the integration of electroencephalogram (EEG) and peripheral physiological signals to assess arousal levels. Combining physiological data with speech and image features enables a more nuanced understanding of emotional responses, particularly in contexts where physiological changes play a significant role.

3.5 Applications in Film

Emotion recognition technologies have found compelling applications in the realm of filmmaking, transforming the cinematic experience and storytelling. Understanding audience emotions opens new avenues for creating engaging narratives and enhancing overall viewer satisfaction.

3.5.1 Real-Time Scene Adaptation

In the work of Johnson et al. (2015) on "Emotionally Responsive Storytelling for Digital Games" [60], real-time emotion recognition is employed to adapt the progression of digital game narratives. Similar techniques can be applied to filmmaking, allowing scenes to dynamically respond to the emotional state of the audience. [61]

3.5.2 Character Animation in Animated Films

The study by Zhu et al. (2020) on "EmoGen: Deep Learning-Based Emotional Character Generation for Interactive Storytelling" [62] focuses on emotional character generation for interactive storytelling. Deep learning techniques enable the creation of animated characters that express emotions authentically, contributing to a more immersive viewing experience.

4 Methodology

4.1 Introduction

In the world of movies, where pictures and words come together, emotions come alive. This study aims to unravel how the emotions expressed on characters' faces connect with how they say things. Rather than focusing solely on pictures or words, the interest lies in understanding how these elements work together to evoke emotions.

What sparked curiosity is the impactful way movies make us feel emotions. The goal is to better grasp how the emotions displayed on characters' faces might be linked to how they express themselves. The objective is to create a complete picture of how emotions unfold in movies, with a specific emphasis on the connection between facial expressions and the way characters speak. In these initial sections, the groundwork is laid for this exploration by explaining what needs to be discovered, why it's important, and where this study fits into the broader context of computers understanding emotions, analysing media, and how people and technology interact. This endeavour is like a journey, aiming to uncover the magic of emotions in movies, with a particular focus on assisting film directors in achieving a balance between facial and spoken emotions within scenes.

4.2 Use Case

Within the expansive landscape of emotion recognition in movies, this study zeroes in on a specific use case: scenes featuring characters expressing themselves through speech. These instances, often termed as monologue scenes, provide a controlled setting for analysis. The focus on monologue scenes allows us to hone in on the nuanced interplay between facial expressions and the way characters articulate their feelings. By concentrating on these specific moments, the study aims to uncover patterns and correlations between facial emotions and speech, contributing to a more detailed understanding of how these elements synchronize to convey emotions within cinematic narratives.

This targeted use case selection is not only about dissecting technical intricacies but also holds practical implications. The insights gained from monologue scenes can potentially provide guidance to film directors seeking to strike a balance between facial and spoken emotions. As explored in the introduction, this aligns with the broader goal of the study: to assist filmmakers in achieving a harmonious blend of visual and auditory emotional cues within scenes. The following sections will delve into the methodology, featuring Machine Learning algorithms, to untangle the complexities of this use case and shed light on the correlation between facial and speech emotions in movie scenes.

4.3 Techniques

4.3.1 Transfer Learning

Central to the fine-tuning process for pre-trained ImageNet algorithms is the utilization of transfer learning. This technique capitalizes on the existing knowledge acquired by the models during training on the ImageNet dataset [6]. By leveraging this wealth of generalizable features, the models become adept at recognizing emotional nuances within cinematic scenes. The transfer of knowledge from a diverse range of images to the specific context of emotion recognition in films enhances the models' adaptability and performance in the dynamic cinematic environment.

4.3.2 Data Augmentation (Rotation, Random Crop for Images)

To augment the models' ability to generalize to various emotional expressions, data augmentation techniques were deployed. Rotation and random cropping emerged as key augmentation strategies, diversifying the training dataset. The introduction of variations in image orientation and cropping exposes the models to a broader range of emotional expressions. This augmentation technique helps prevent overfitting by subjecting the models to a more extensive set of scenarios, thereby enhancing their robustness and adaptability to the diverse emotional content present in cinematic scenes [64].

4.3.3 Weighted Ensemble

To optimize the predictive capabilities of the emotion recognition models, a weighted ensemble of the two best-performing models, determined based on accuracy, was crafted. This ensemble approach combines the outputs of multiple models with varying weights, assigning higher importance to the more accurate models. This technique draws on the strengths of individual models and mitigates potential weaknesses, resulting in a more balanced and accurate prediction. The creation of a weighted ensemble contributes to the overall effectiveness of the emotion recognition system, combining the insights of the top-performing models [65].

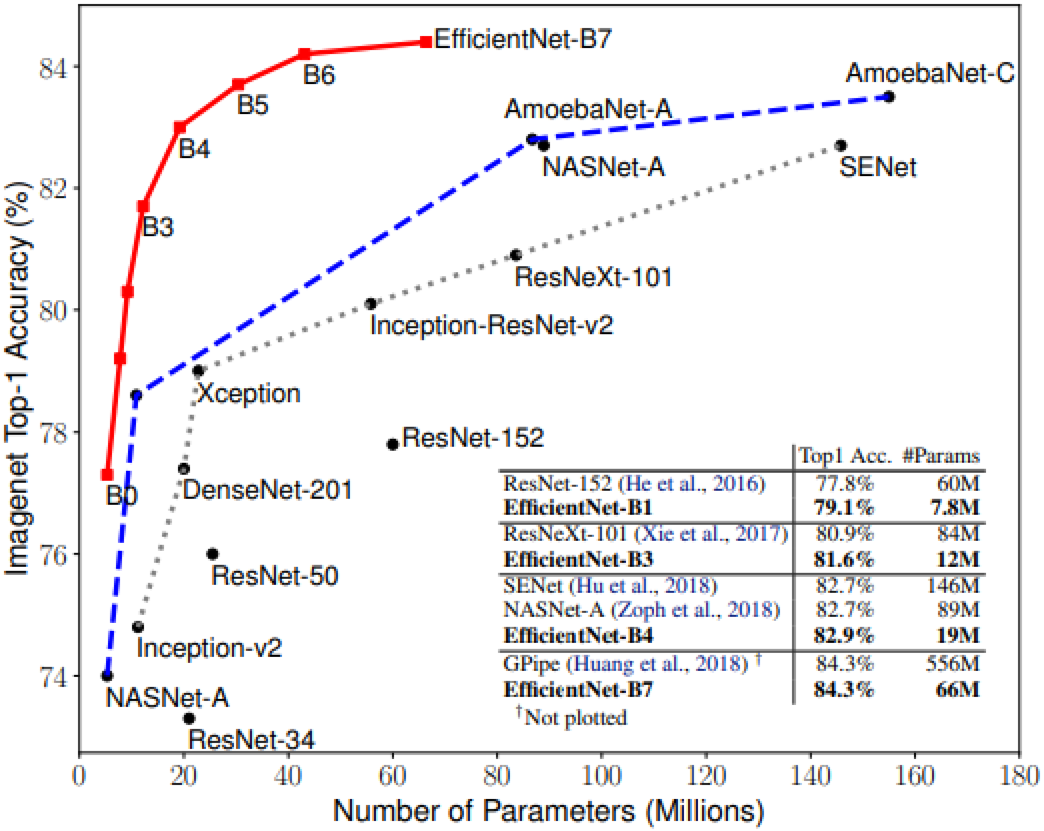
In summary, transfer learning enhances adaptability, data augmentation refines robustness, and a weighted ensemble optimizes accuracy. These techniques collectively form a robust framework for fine-tuning pre-trained ImageNet algorithms, ensuring their efficacy in recognizing emotions within the dynamic and complex context of cinematic storytelling.

4.4 Image Emotion Recognition Algorithms

In the pursuit of decoding emotions within the intricate tapestry of cinematic narratives, a diverse ensemble of pre-trained ImageNet algorithms was meticulously chosen, each offering a unique lens through which to understand and capture emotional nuances.

4.4.1 EfficientNet

The EfficientNet family, introduced by Tan, M., & Le, Q. V. [63], spanning from B0 to B6, was incorporated for its scalability and efficiency, providing a balanced compromise between computational resources and model performance. EfficientNet is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. Unlike conventional practice that arbitrary scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients. EfficientNets also transfer well and achieve state-of-the-art accuracy on CIFAR-100 (91.7%), Flowers (98.8%), and 3 other transfer learning datasets, with an order of magnitude fewer parameters.



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

Figure 2: ImageNet Model Size vs. Accuracy

The strategic inclusion of EfficientNet variants ensures adaptability to the complex and dynamic nature of emotional expressions in film scenes.

4.4.2 ResNet

ResNet34 and ResNet50, integral components of the ResNet architecture, emerged as robust choices for their deep structures and skip connections, effectively addressing challenges related to vanishing gradients in deep neural networks [43]. The layers have been reformulated as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. On the ImageNet dataset residual nets with a depth of up to 152 layers—8× deeper than VGG nets [28] have been evaluated, but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set.

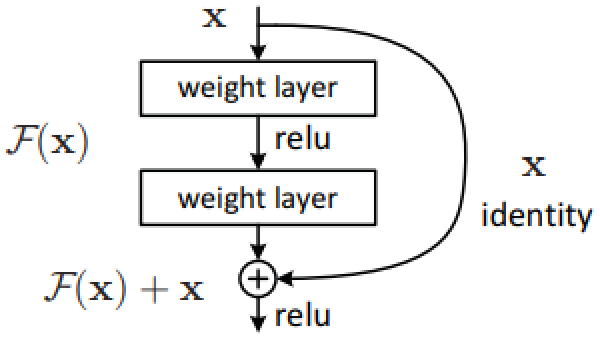

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 3: Residual learning: A building block

The architectural prowess of ResNet models positions them as formidable tools in capturing intricate emotional expressions embedded in the visual content of cinematic narratives. Their deep and skip-connected design equips them to navigate the nuanced complexities of emotions portrayed in diverse film scenes.

4.4.3 VGG

The VGG series, comprising VGG11, VGG13, VGG16, and VGG19, was strategically included for its simplicity and effectiveness in image classification tasks [28]. Known for its straightforward architecture with varying depths, the VGG series brings a contrasting approach to understanding emotional nuances in cinematic scenes. Its emphasis on small convolutional filters makes it adept at capturing fine-grained details, complementing the broader spectrum provided by other algorithms.

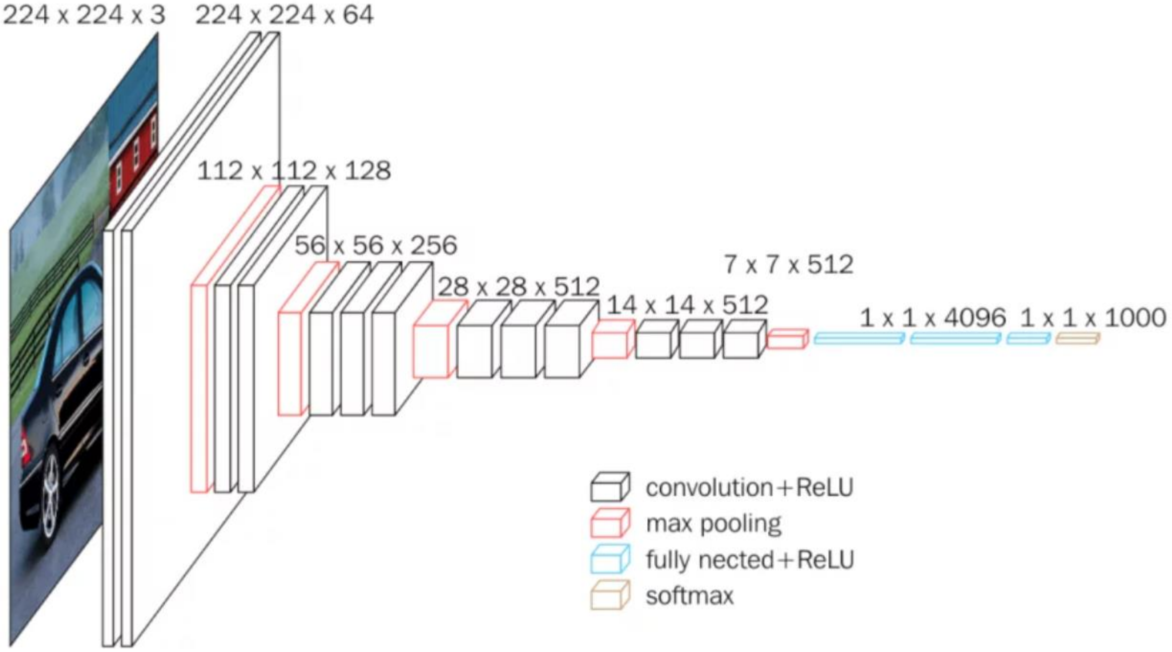


Figure 4: VGG16 Architecture

EfficientNet B0 to B6, ResNet34, and ResNet50, alongside VGG11, VGG13, VGG16, and VGG19, collectively form a diverse set of pre-trained ImageNet algorithms. This strategic selection ensures a comprehensive exploration of emotional nuances within varied cinematic scenarios. Each algorithm brings its unique strengths to the forefront, promising a nuanced and holistic understanding of emotion recognition within the dynamic realm of films.

4.5 Speech Emotion Recognition Algorithms

4.5.1 Whisper

In the realm of advanced audio processing for emotion recognition in film scenes, the Whisper model takes centre stage. Highlighted in the paper "Robust Speech Recognition via Large-Scale Weak Supervision" [71] Whisper employs a sophisticated encoder-decoder Transformer architecture tailored for large-scale supervised pre-training. This model exhibits remarkable adaptability to the intricacies of cinematic speech, leveraging webscale text data for training. Its minimalist preprocessing strategy involves predicting raw text transcripts without extensive standardization, streamlining the speech recognition pipeline. The Whisper model's multitask format enhances its versatility, seamlessly incorporating various speech processing tasks crucial for precise emotion recognition in film audio.

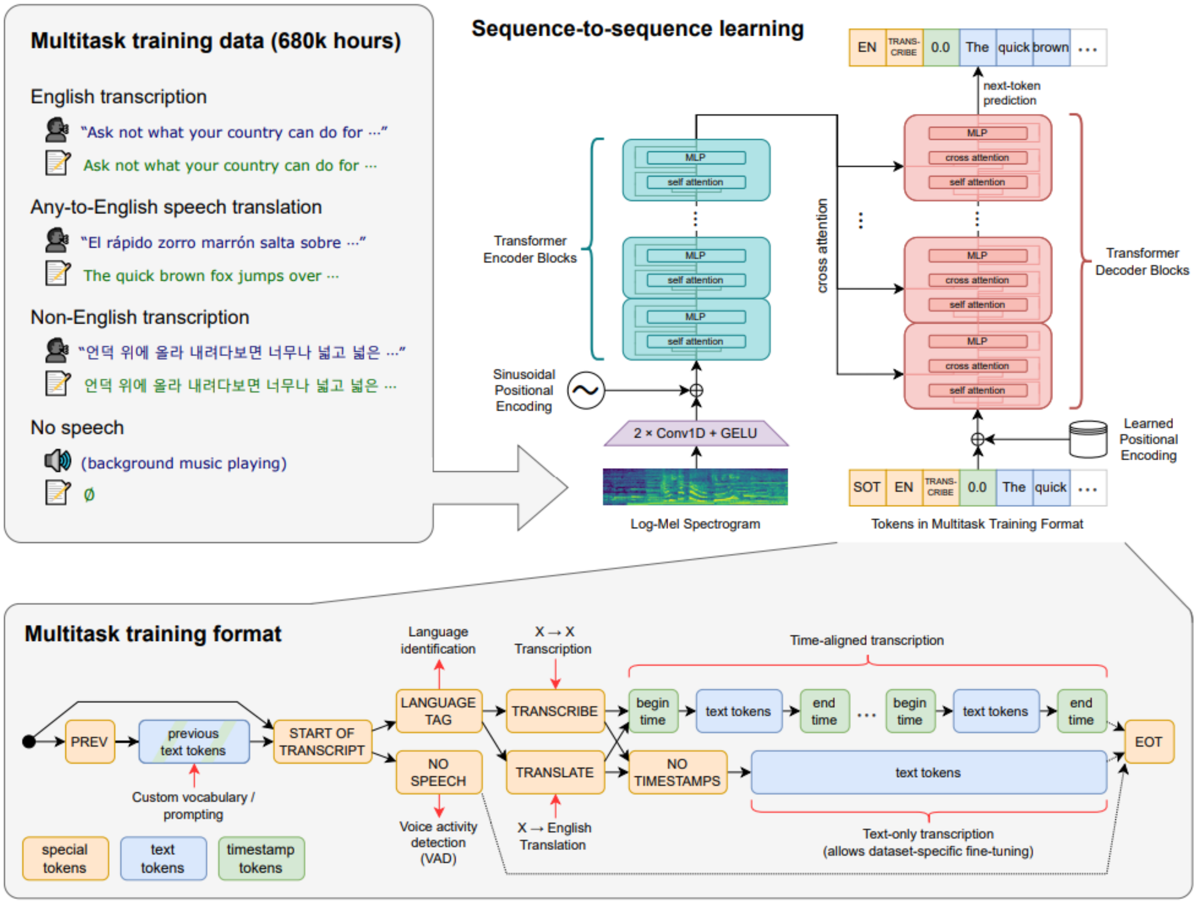

Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ···"
Ask not what your country can do for ···
Any-to-English speech translation
"El rápido zorro marrón salta sobre ···"
The quick brown fox jumps over ···
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
···
next-token prediction
MLP
self attention
MLP
cross attention
self attention
Transformer Encoder Blocks
cross attention
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
···
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
TRANSCRIBE
begin time
text tokens
end time
···
begin time
text tokens
end time
EOT
NO SPEECH
TRANSLATE
NO TIMESTAMPS
text tokens
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure 5: Overview of Whisper approach

4.5.2 HuBERT

In addition to Whisper, the experiments conducted for this thesis also evaluated the Hubert model, introduced in the paper "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." [74] Hubert, a Hidden-Unit BERT approach for self-supervised speech representation learning, addresses challenges unique to the field. It incorporates an offline clustering step and applies a prediction loss exclusively over masked regions, compelling the model to learn a combined acoustic and language model over continuous inputs. Hubert's training process, relying on the consistency of unsupervised clustering steps, showcases its effectiveness in capturing nuanced audio representations.

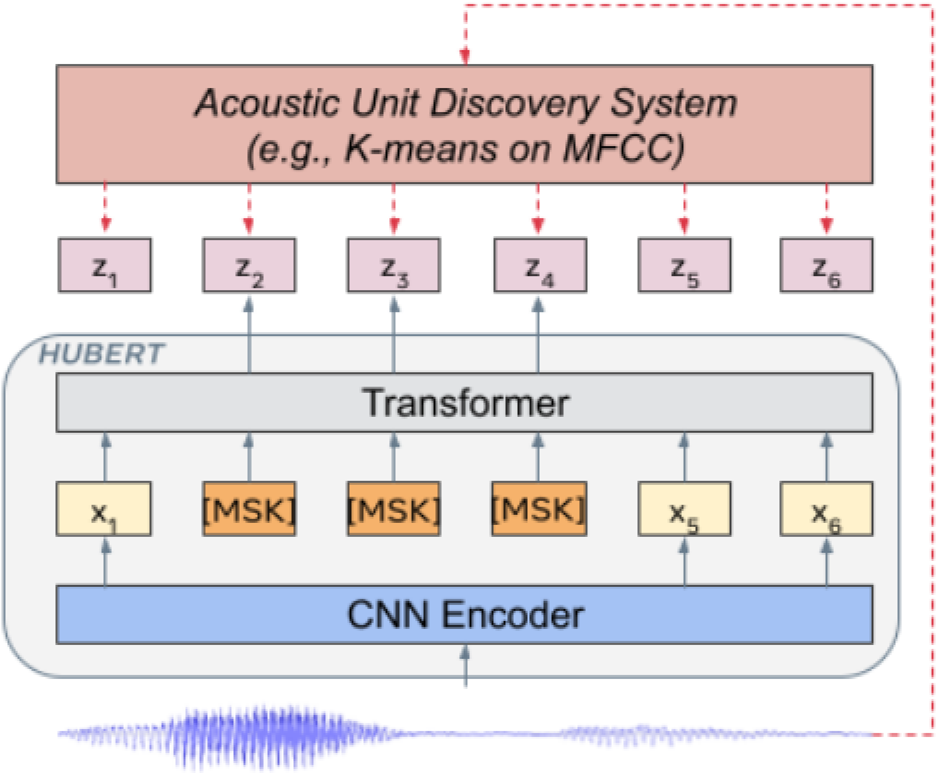


Figure 6: The HuBERT approach predicts hidden cluster assignments of the masked frames generated by one or more iterations of k-means clustering

4.5.3 Wave2Vec2

Complementing this audio analysis toolkit, the Wave2Vec2 model [73], an evolution of Wave2Vec [72], was also evaluated in the experiments conducted for this thesis. Building on the principles of self-supervised learning for speech representations, Wave2Vec2 refines feature extraction through extensive training on audio corpora. While Wave2Vec2's architecture and methodologies differ, its core focus on capturing contextual information within speech aligns with the pursuit of nuanced audio representations.

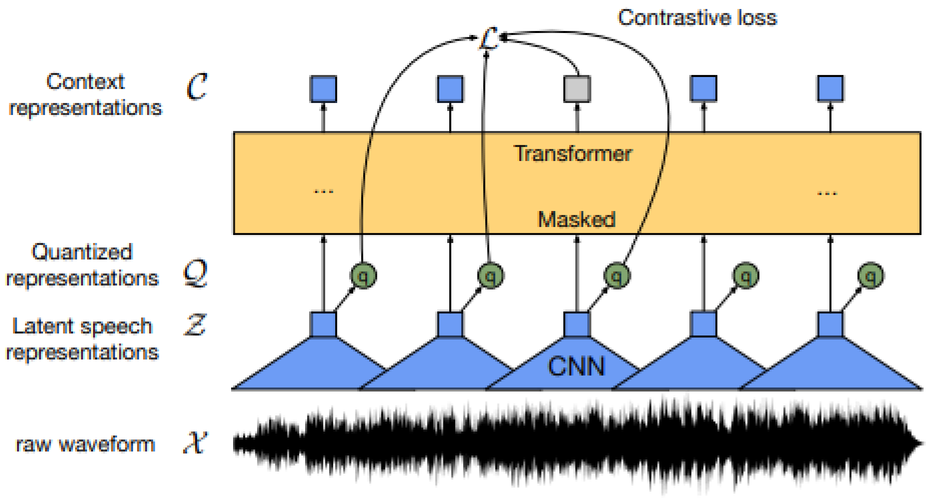


Figure 7: Illustration of Wave2Vec2 framework which jointly learns contextualized speech representations and an inventory of discretized speech units

4.5.4 Sew-d

Adding to the comprehensive exploration of diverse models, the Sew-d model was also included in the evaluation for this thesis. Explored in the paper "Performance-Efficiency Trade-Offs in Unsupervised Pre-Training for Speech Recognition," [75] which focuses on wav2vec 2.0, and formalizing several architecture designs that influence both the model performance and its efficiency, SEW (Squeezed and Efficient Wav2vec) is a pre-trained model architecture with significant improvements along both performance and efficiency dimensions across a variety of training setups.

Sew-d offers valuable insights into the delicate balance between computational efficiency and recognition performance in unsupervised pre-training for speech recognition. This model provides an alternative perspective, contributing to the enhancement of speech representation learning and addressing the increasing demand for efficient yet effective models in real-world applications.

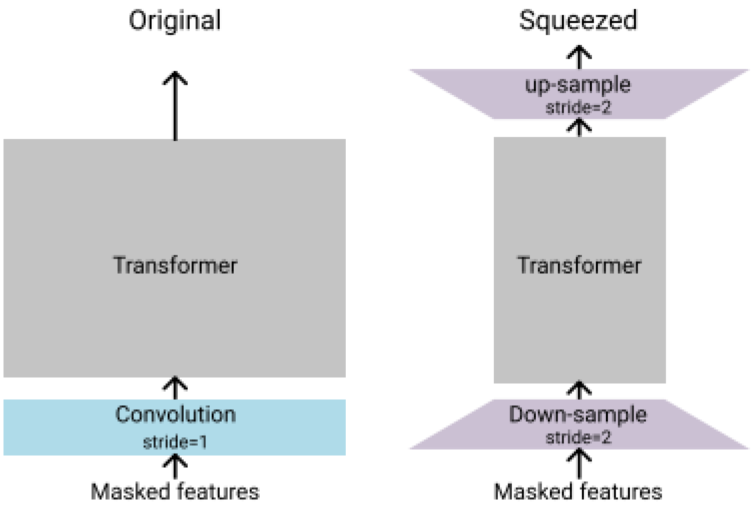


Figure 8: Original vs. squeezed context network

4.6 Classifiers evaluation

To evaluate the performance of the proposed classifier, applicable to both image and speech processing, two key metrics were employed: *Accuracy* and *F1-score*. These metrics are derived from the elements of a confusion matrix, as outlined in the table below:

	Predicted		
Actual		Predicted Positive	Predicted Negative
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 1: Confusion Matrix

Accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

This metric primarily indicates the proportion of correctly classified samples. However, Accuracy alone may not comprehensively reflect the classifier's performance, especially in cases where the class distribution is imbalanced.

To address this limitation, the F1-score is also calculated, which is the harmonic mean of Precision and Recall. This metric provides a balanced view of the classifier's performance, considering both the correctly classified and the incorrectly classified samples. The F1-score is particularly useful in situations where an equal importance is assigned to both Precision and Recall. The calculation of the F1-score is as follows:

$$F_{1-score} = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

where,

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

By utilizing both these metrics, a thorough evaluation of the classifiers' effectiveness in various scenarios is provided.

4.7 Correlation between Image and Speech

In this study, several statistical tests were utilized to analyse the correlation between the predictions of different classifiers. These tests include the *Chi-Square test* [76], *Cramer's V association* [77], and the Pearson correlation coefficient (*Pearson's r*) [78], each serving a distinct purpose in our analysis.

Specifically, the Chi-Square test, which is a non-parametric method, assesses the independence between two categorical variables. It should be noted that in our case the corresponding variables are the image and speech predictions, and their values are integers corresponding to a class-emotion. It is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where χ^2 is the Chi-Square statistic, O_i represents the observed frequency, and E_i is the expected frequency under the assumption of independence. This test helps in determining whether deviations from expected frequencies are due to chance or indicate a significant association.

On the other hand, Cramer's V, which is built on the Chi-Square test, measures the strength of association between two nominal variables. It is given by:

$$V = \sqrt{\frac{\frac{\chi 2}{N}}{min(k-1, r-1)}}$$

Here, χ^2 is the Chi-Square statistic, N is the total number of observations, and k and r are the numbers of categories (classes) for each of the two variables. Cramer's V ranges from 0 (no association) to 1 (complete association), providing a more nuanced understanding of the relationship strength.

Similarly, Pearson's r measures the linear correlation between two continuous variables. In our cases, it is applied on the predicted probabilities of the two classifiers allowing to quantify the strength of the correlation of the classifiers not only based on the top-1 predicted classes but also on the corresponding probabilities. It is calculated using the formula:

$$r = \frac{\sum \quad (X_i - \underline{X})(Y_i - \underline{Y})}{\sqrt{\sum \quad (X_i - \underline{X})^2 (Y_i - \underline{Y})^2}}$$

where X_i and Y_i are the individual sample points indexed with i, and $\underline{X}$ and $\underline{Y}$ are the sample means. The value of r ranges from -1 to +1, indicating the strength and direction of a linear relationship.

The application of these statistical tests in our analysis is pivotal. The Chi-Square test identifies whether an association exists, Cramer's V quantifies the strength of association in categorical data (predicted classes), and Pearson's r provides insight into the linear correlation between continuous variables (predicted probabilities). Together, these methods offer a comprehensive understanding of the interrelationships among classifier predictions.

5 Experiments

5.1 Tools and Platforms

In the pursuit of advancing emotion recognition in film scenes through the amalgamation of speech and image analysis, an array of robust tools and platforms were employed. These technological choices were instrumental in crafting a sophisticated and effective system for deciphering emotional nuances in cinematic content.

Python, a versatile and widely adopted programming language, was chosen as the cornerstone of our implementation. Renowned for its extensive libraries and ease of integration, Python facilitated the seamless development of the Emotion Recognition system. Its adaptability proved crucial in incorporating machine learning models and data processing techniques seamlessly.

Kaggle Jupyter Notebooks emerged as a pivotal platform for conducting experiments in a collaborative and interactive environment. Leveraging Kaggle's infrastructure, the research team could efficiently code, visualize data, and analyse results within a centralized space. The platform's accessibility and pre-installed libraries streamlined the development and testing of emotion recognition models.

PyTorch, a dynamic deep learning framework, played a central role in implementing neural network architectures. The framework's flexibility allowed for the construction and fine-tuning of models tailored to the dynamic nature of film scenes. PyTorch's dynamic computation graph and user-friendly interface facilitated efficient experimentation, ensuring the adaptability of the emotion recognition system to diverse cinematic scenarios.

To harness the computational power necessary for accelerated model training, the study incorporated CUDA (Compute Unified Device Architecture). CUDA enables the utilization of Graphics Processing Units (GPUs), enhancing the computational efficiency of deep learning algorithms implemented in PyTorch. This not only expedited the experimentation process but also allowed for the scalability of the emotion recognition system.

In conjunction with these tools, an efficient Integrated Development Environment (IDE) played a pivotal role in streamlining the coding and debugging process. The chosen IDE contributed to the overall productivity of the research team, ensuring a smooth workflow during the development of algorithms and models.

The judicious selection and integration of these tools and platforms created a robust foundation for conducting experiments and deriving meaningful results in the subsequent phases of the research endeavour.

5.2 Experimental setting

5.2.1 Datasets Description

5.2.1.1 Image Dataset

The FER2013 dataset [66] serves as a pivotal component in the image analysis aspect of this research, contributing valuable insights into facial emotion recognition. Comprising over 35,000 images, FER2013 captures a diverse range of facial expressions across seven emotion categories, including *happy*, *sad*, *surprise*, *angry*, *disgust*, *fear*, and *neutral*. Each image is labelled with the corresponding emotion, providing a rich resource for training, and evaluating deep learning models. The dataset's diversity is a key asset, reflecting variations in pose, lighting, and ethnicity, making it well-suited for the study's objective of recognizing emotions in complex cinematic scenes.

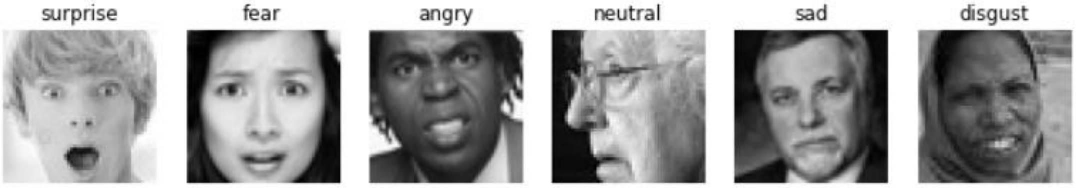


Figure 9: FER2013 Dataset examples

The images within FER2013 are grayscale and sized at 48x48 pixels, striking a balance between computational efficiency and expressive facial features. This resolution ensures that critical facial details are preserved while optimizing the computational resources required for model training. Leveraging FER2013 facilitates the development of a robust image-based emotion recognition system, allowing the model to generalize effectively to the nuanced emotional expressions present in film scenes. The careful curation and annotation of the FER2013 dataset position it as a foundational resource in the exploration of emotion recognition within the context of cinematic imagery.

It is important to note that the *surprise* class has been excluded to align with the range of emotions present in the datasets described subsequently, which encompass both speech and video modalities. To prevent any confusion, the modified subset of the FER2013 dataset, from which the *surprise* class has been omitted, will be referred to as *FER2013-selected* throughout the remainder of this work. In total, the number of samples kept is 25538. Subsequently, the dataset was randomly divided into training and validation subsets in an 80/20 ratio, respectively, that is, 20430 for training and 5108 for validation. *Figure 3* illustrates the class distribution within these subsets. It is evident from the figure that the *disgust* class has the fewest samples. However, both subsets exhibit a similar distribution, which is conducive to extracting reliable results during the evaluation phase.

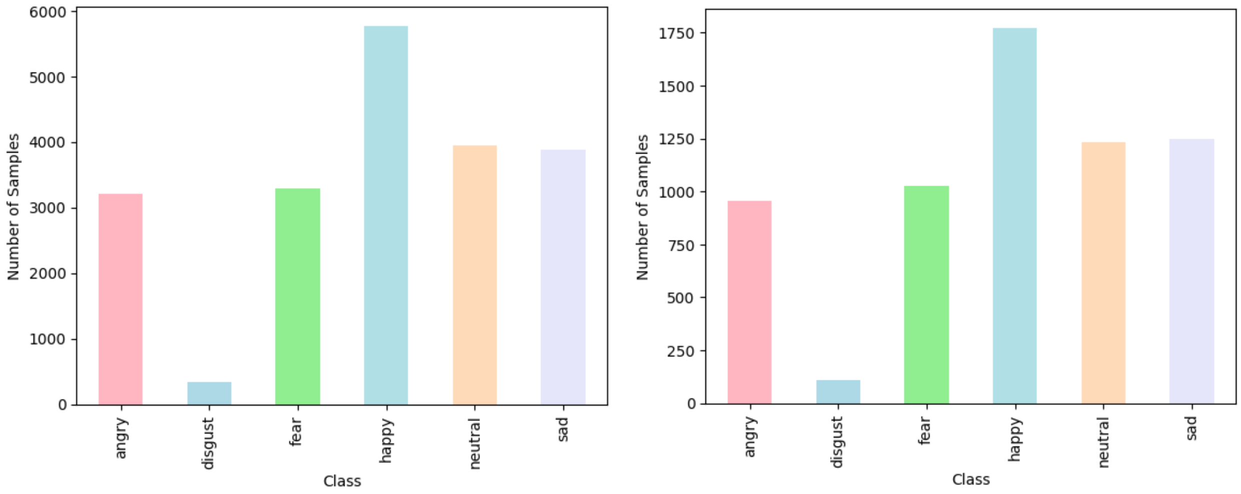


Figure 10: Histogram of class distribution of FER2013-selected. (Left): Train set, (Right): Validation Set

5.2.1.2 Speech Datasets

The sound aspect of this research encompasses a comprehensive blend of diverse audio datasets, strategically chosen to enrich the study of emotion recognition in cinematic scenes. It includes the CREMAD (Crowdsourced Emotional Multimodal Actors Dataset) [67], a valuable collection of audio recordings featuring a range of emotional expressions portrayed by multiple actors. The dataset provides a nuanced understanding of emotions in speech, contributing significantly to the robustness of the developed emotion recognition system.

Additionally, the research incorporates both the RAVDESS song and speech portions (Ryerson Audio-Visual Database of Emotional Speech and Song) [68]. This extensive dataset covers a wide spectrum of emotions expressed through speech and song, providing a comprehensive foundation for the audio analysis component. The inclusion of both song and speech ensures a holistic understanding of emotional cues, crucial in capturing the multifaceted nature of emotions within cinematic contexts.

Furthermore, the study leverages the Surrey Audiovisual Expressed Emotion (SAVEE) dataset [69]. This dataset includes recordings of actors vocalizing a range of emotions, enhancing the diversity of emotional expressions available for analysis. The combination of facial and vocal cues in the Surrey Audiovisual dataset contributes to a more holistic approach to emotion recognition in cinematic audio-visual scenes.

Finally, the Toronto Emotional Speech Set (TESS) forms an integral part of the audio datasets [70]. TESS comprises naturalistic emotional expressions, providing a valuable resource for training models to recognize emotions in speech. The incorporation of TESS enhances the generalization capabilities of the developed system, enabling it to discern subtle emotional nuances in cinematic audio.

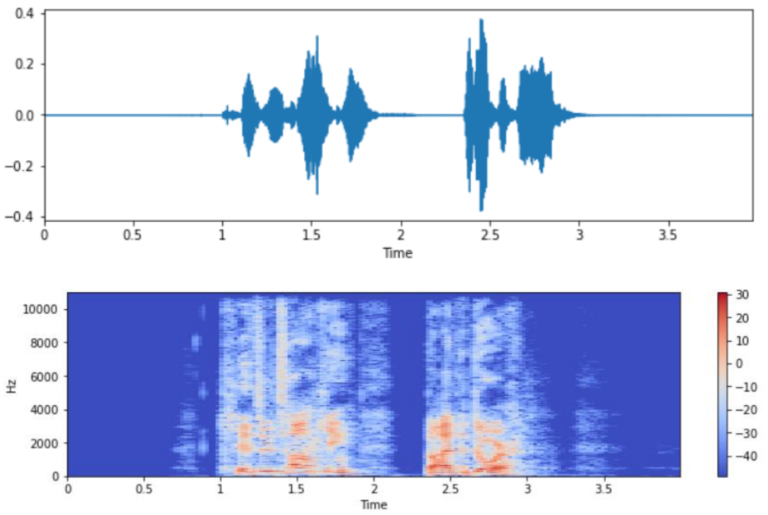


Figure 11: Example of 'fear' speech sample. (Top): Waveplot. (Bottom): Spectrogram

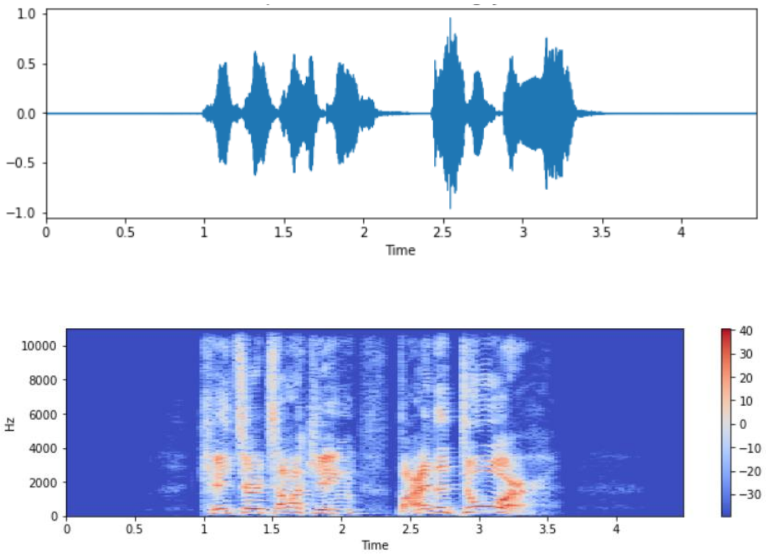


Figure 12: Example of 'angry' speech sample. (Top): Waveplot. (Bottom): Spectrogram

In all datasets, the same classes as in FER2013 were retained, namely, *happy*, *sad*, *angry*, *disgust*, *fear*, and *neutral*. More specifically, the *calm* class from RAVDESS was excluded, as well as the *surprise* class from RAVDESS, TESS, and SAVEE. Furthermore, as each dataset includes the ID of the speaker, they were split into development and test sets at a 90/10 ratio based on the number of speakers, when feasible (for instance, the TESS dataset contains only two speakers; therefore, the first was assigned to the development set and the second to the test set). The test set comprises speakers not used in training, ensuring that the test results more accurately reflect the trained model's generalization capabilities. It is important to note that within each individual dataset, the number of samples for every speaker is identical. Subsequently, the development set was further divided randomly into training and validation sets at an 80/20 ratio, based on the number of samples rather than the number of speakers. The following table summarizes the number of speakers and samples in each dataset and sub-dataset:

		Number of speakers	**Number of samples**
RAVDESS	**Dev**	22	1724
	Test	2	160
CREMAD	**Dev**	82	6704
	Test	9	738
TESS	**Dev**	1	1200
	Test	1	1200
SAVEE	**Dev**	3	315
	Test	1	105
Final dev		108	9943
Train		108	7954
Validation		108	1989
Final Test		13	2203

Table 2: Distribution of speakers and samples across speech datasets

Furthermore, as shown in the subsequent figure, it has been ensured that there is a similar distribution of classes among the train, validation, and test subsets of the final merged dataset:

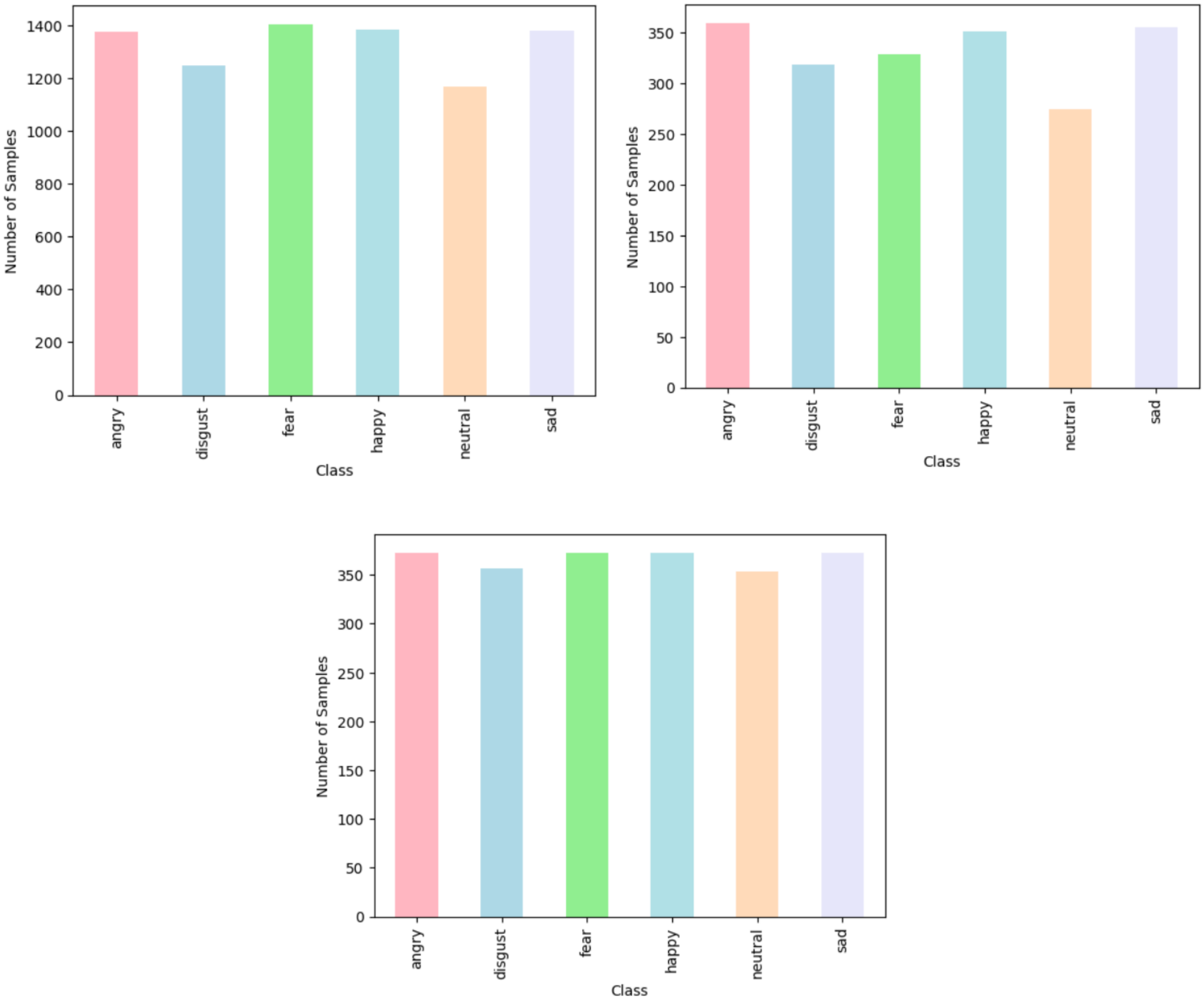


Figure 13: Histogram of class distribution of the final merged speech dataset.

All samples are shorter than 30 seconds, allowing them to be directly fed into the speech classifiers described in Section 4.5 without the need for segmentation into windows. However, to ensure uniformity in sample length, padding with zeros was applied at the end of each sample, resulting in a fixed length of 30 seconds for all samples.

This amalgamation of audio datasets forms a rich and diverse foundation for the research, allowing for a comprehensive exploration of emotion recognition in cinematic contexts, where both speech and ambient sounds contribute to the overall emotional landscape.

5.2.1.3 Video dataset

For the correlation analysis between image and speech classifiers, the One-Minute Gradual-Emotional Behavior (OMG-Emotion) dataset [79] was selected. This dataset comprises 2,400 samples, each annotated by five annotators. The ground truth for each sample was determined by selecting the emotion that received the maximum votes. Every sample is a short RGB video (less than 30 seconds) featuring a monologue expressing a single emotion. What sets the OMG-Emotion dataset apart from previous ones is its focus on *long-term* emotional behaviour classification, in contrast to the short-term emphasis of other datasets (e.g., MOSI [80], EmoReact [81], GIFGIF+ [82]), which typically analyse only a few (1-2) seconds of video length.

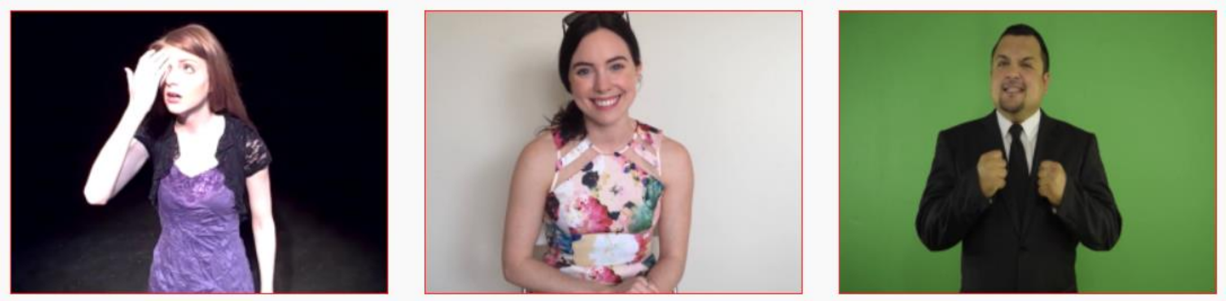


Figure 14: Frames extracted from videos of the OMG-Emotion dataset.

Like the Image and Speech datasets, only the samples corresponding to the emotions *happy*, *sad*, *angry*, *disgust*, *fear*, and *neutral* were retained, while excluding those labelled as *surprise*. During dataset preparation (download), it was noticed that several videos were not available, hence the number of samples used is 1,470. The classes' distribution illustrated in the following figure indicates that *fear* and *disgust* classes are the most underrepresented classes:

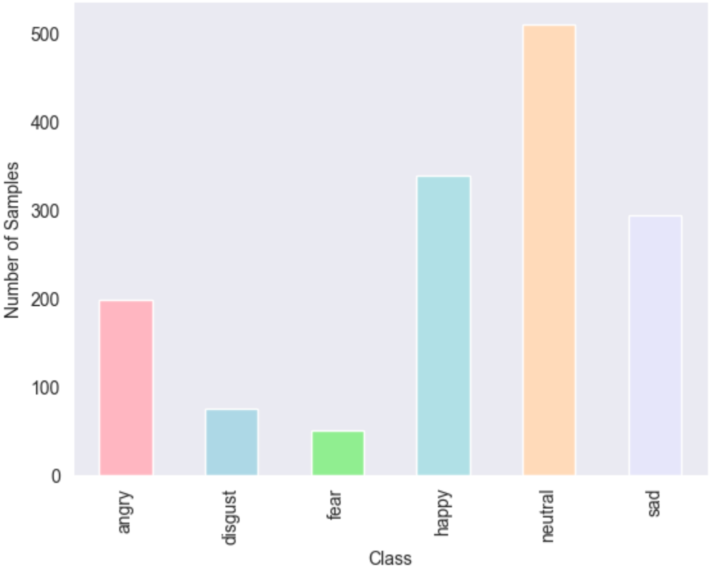


Figure 15: Histogram of class distribution of the final video dataset.

5.2.2 Training specifications

5.2.2.1 Image classifier

For the image classifier, EfficientNet, ResNet and VGG models were trained with the FER2013 dataset. However, only the ResNet34 results may be found in this thesis since it outperformed the others, probably since the dataset was small, and the images were in grey scale.

To fine-tune ResNet34 on the FER2013-selected dataset, each image was first resized from (48×48) to (224×224) pixels to match the input size for which ResNet34 was pretrained. To prevent overfitting, it was applied *Image Augmentation* by randomly flipping the images along both the vertical and horizontal axes. After augmentation, the pixel values were normalized by dividing them by 255 to bring them into the 0 to 1 range. Then standardization (Z-score normalization) was performed using a mean of 0.485 and a standard deviation of 0.229, in line with the data preprocessing used during the pretraining of ResNet34. Additionally, *Early Stopping* technique was implemented based on the validation set loss value to ensure, as much as possible, the generalization capability of the model.

In the subsequent table, the values of the hyperparameters used during training are provided:

Batch Size	128
Learning Rate	1e-05
Epochs	1000
Early Stop Patience	20
Frozen Layers	First 57
Optimizer	Adam
Loss Function	Cross Entropy

Table 3: Image classifier training hyperparameters

It should be noted that due to lack of computational resources, hyperparameters were not tuned, thus their values were chosen based on experience and intuition. *Batch Size* was selected based on the maximum number of images that can be loaded on the GPU provided by Kaggle (Nvidia P100).

5.2.2.2 Speech classifier

To fine-tune Whisper-tiny, HuBERT-base, and Wav2Vec2-base, the sampling rate and maximum length were aligned with the parameters used during the pretraining of these models. Specifically, all models were pretrained using a 16 kHz sampling rate. However, only Whisper-tiny imposes a maximum sample length requirement, specifically 30 seconds. As detailed in Section 5.2.1.2, zero-padding was applied at the end of each sample to achieve a uniform length across all models, ensuring a fair comparison.

The table below summarizes the hyperparameters employed for each model:

	Whisper-tiny	**HuBERT-base**	**Wav2Vec2-base**
Batch Size	128	2	2
Learning Rate	1e-05	1e-05	1e-05
Epochs	1000	1000	1000
Early Stop Patience	20	20	20
Frozen Layers	First 50	First 208	First 195
Optimizer	Adam	Adam	Adam
Loss Function	Cross Entropy	Cross Entropy	Cross Entropy

Table 4: Speech classifiers training hyperparameters

Like the image classifier, hyperparameters were not tuned due to resource constraints. For all models, only the last block of layers along with the classification head were trained. Therefore, the *Frozen Layers* hyperparameter was set based on the number of layers preceding the last block. The *Batch Size* for HuBERT-base and Wav2Vec2-base was determined by the maximum number of samples that our GPU could accommodate. Conversely, for Whisper-tiny, the *Batch Size* was chosen based on the maximum number of samples that our available RAM (25 GB) could handle. CPU training was opted for Whisper-tiny, as it proved to be significantly less time-consuming compared to CPU training for HuBERT-base and Wav2Vec2-base.

5.2.3 Correlation analysis pipeline

To investigate the relationship between the predictions of Image and Speech classifiers, a meticulous pipeline was established based on the OMG-Emotion dataset, which was cleaned as detailed in Section 5.2.1.3. This involved three key subprocesses: a) Image Process, b) Speech Process, and c) Correlation Analysis Process.

Image process consists of the following steps:

- Dividing each video monologue into frames.
- Detecting faces in each frame using the light-weight detector provided by the *python-opencv* package called *haarcascade_frontalface_default*. Each detected

face is outlined by a bounding box defined by 4 values, namely, *x, y, w, h*, where *x, y* are the coordinates of the top-left corner of the bounding box, and *w, h* are the width and height of the bounding box.

- Cropping the frame to the area within bounding box.
- Converting each cropped RGB frame to greyscale.
- Resizing cropped frames to (224 × 224).
- Normalizing frames as described in Section 5.2.1.1.
- Obtaining predictions (probabilities for each class) for all frames of a video.
- Calculating the mean probability for each class using predictions from all frames.

Speech process consists of the subsequent steps:

- Extraction raw speech values using *moviepy* package.
- Converting from stereo to mono by averaging the two channels.
- Generating a single prediction (class probabilities) for each video, as videos are shorter than 30 seconds and do not require segmentation into multiple windows.

Correlation analysis process is comprised of the following steps:

- Determining the class with the maximum probability for each sample and for each classifier. These variables may be called as *speech_pred* and *img_pred*. Note that for each sample there is a single value for each classifier, i.e., the class-emotion ID. The initial variables which include the predicted probabilities of all classes will be named as *speech_pred_prob* and *img_pred_prob*.
- Calculating the p-value of Chi Square test for the correlation between *speech_pred* and *img_pred* variables.
- Converting the values of *speech_pred* and *img_pred* to one-hot encoding. These variables will be the *speech_pred*_one_hot and *img_pred_one_hot*.
- Computing the p-value of Chi Square test for each pair of classes on the correlation of *speech_pred_one_hot_<emotion>* and *img_pred_one_hot_<emotion>* variables, where *<emotion>* is a placeholder for each one of the six emotions.

- Computing Cramer's V Association between *speech_pred* and *img_pred* variables.
- Computing Cramer's V Association for each pair of classes, i.e., between all pairs of variables *speech_pred_one_hot_<emotion>* and *img_pred_one_hot_<emotion>* variables.
- Compute Pearson's r for each pair of classes, i.e., between all pairs of variables *speech_pred_prob_<emotion>* and *img_pred_prob_<emotion>*.

5.3 Results

The experimental phase of this research involved the training and evaluation of a series of carefully selected algorithms designed for both image and audio-based emotion recognition in cinematic scenes. For the image analysis component, deep learning models based on convolutional neural networks (CNNs) were trained using the FER2013 dataset. The models underwent rigorous training to capture intricate facial features and nuances associated with different emotional expressions. Simultaneously, for the audio analysis, a combination of recurrent neural networks (RNNs) and attention mechanisms were employed. These audio models were trained on the amalgamated audio datasets, including CREAMD, RAVDESS (song and speech portions), Surrey Audiovisual, and TESS. The utilization of a diverse set of algorithms aimed to capture the multi-modal nature of emotions portrayed in cinematic content, accounting for both visual and auditory cues.

The performance of these trained models was rigorously evaluated against a carefully curated subset of the datasets that remained unseen during the training phase, commonly referred to as the test set. Metrics such as accuracy, precision, recall, and F1 score were employed to assess the models' ability to accurately recognize and classify emotions in cinematic scenes. The evaluation process involved analysing the models' responses to a variety of emotional expressions, ensuring a robust understanding of their performance across different scenarios. The experimental setup was designed to emulate real-world conditions, where the models need to generalize effectively to a diverse range of emotional cues encountered in cinematic storytelling.

Results indicate promising levels of accuracy and effectiveness in emotion recognition across both image and audio domains. The comprehensive evaluation against the test set provides valuable insights into the models' generalization capabilities and their potential applicability in real-world scenarios. These findings contribute to the advancement of emotion recognition technology within the cinematic context, offering a foundation for future developments and applications in areas such as film analysis, virtual reality, and human-computer interaction.

In addition to evaluating individual model performances, a significant component of our analysis involved investigating the correlation between image and speech-based emotion recognition models. This cross-modal correlation study was critical in understanding how visual and auditory cues collectively contribute to emotion recognition in cinematic scenes. To this end, statistical methods such as Chi-Square tests, Cramer's V Association, and Pearson's correlation coefficient were employed to analyse the relationship between the predictions made by image and speech classifiers.

Our results revealed interesting patterns of correlation and divergence between the two modalities. For instance, most of the emotions showed notable discrepancies, indicating unique challenges in capturing the essence of these emotions across different sensory inputs. These findings highlight the complexities inherent in multi-modal emotion recognition and underscore the importance of integrating diverse analytical approaches for a more holistic understanding. The insights gained from this correlation analysis not only validate the effectiveness of our models in their respective domains but also pave the way for developing more sophisticated, integrated systems for emotion recognition in cinematic content.

5.3.1 Image Classification Algorithm Evaluation

This section conducts a comprehensive evaluation of image classification algorithms, specifically trained for the task of emotion recognition in facial expressions. The objective is to provide nuanced insights into the effectiveness of these models, aiding in the selection of algorithms specifically designed for emotion recognition from facial imagery.

The following figures display the classification results of the fine-tuned ResNet34 model. The classification report and confusion matrix reveal that the *happy* class is the most accurately classified, achieving an F1-score of 89%. In contrast, the *disgust* class is the least accurately classified, with an F1-score of just 38%. These outcomes are consistent with the class distribution discussed in Section 5.2.1.1, where the *happy* class had the highest number of samples, and the *disgust* class had the fewest. The overall macro average F1-score is 65%, and the accuracy stands at 72%. These results highlight the complexity of emotion recognition in facial expressions, even for a state-of-the-art model like ResNet34, which appears to face challenges in learning the nuanced patterns of certain emotions.

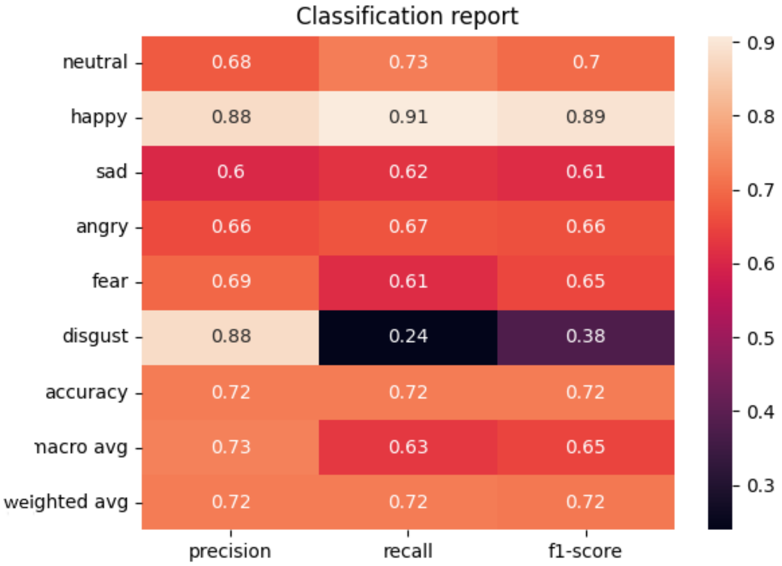


Figure 16: Classification report of ResNet34 on FER2013-selected validation set.

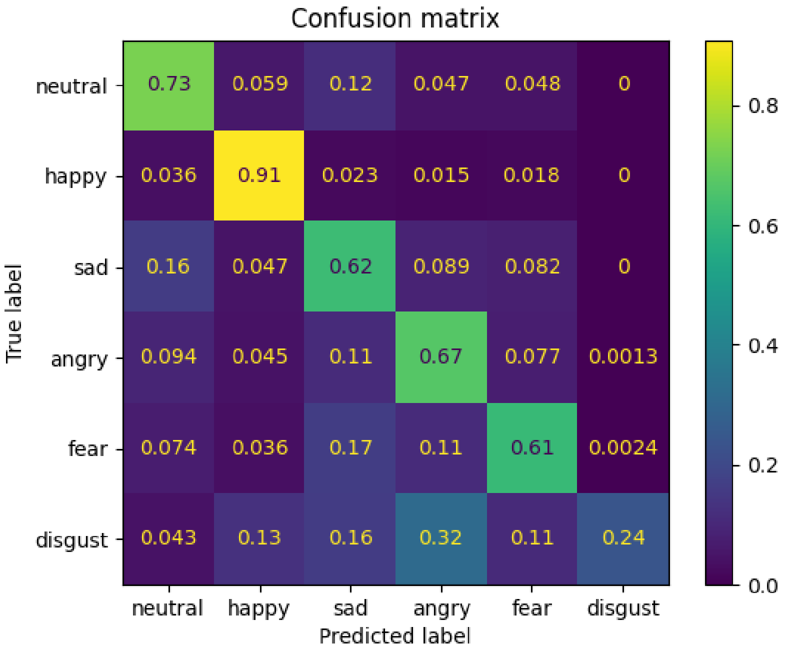


Figure 17: Confusion matrix of ResNet34 on FER2013-selected validation set.

The subsequent figure illustrates the learning curves of ResNet34 during its fine-tuning on the FER2013-selected dataset. Initially, for the first 10 epochs, the training and validation curves progressed similarly. However, post this phase, the model started to overfit. A notable observation is made at epoch 100, where the curves exhibit a significant 'jump', marked by improved validation results and a reduction in training accuracy. This shift mitigated some of the overfitting, allowing the model to continue training for a longer period before being halted by the Early Stopping mechanism. It's worth noting that this 'jump' coincided with the termination of our Kaggle training session after exceeding a continuous run-time of 12 hours. Subsequently, training was resumed by reloading the current model along with its optimizer state.

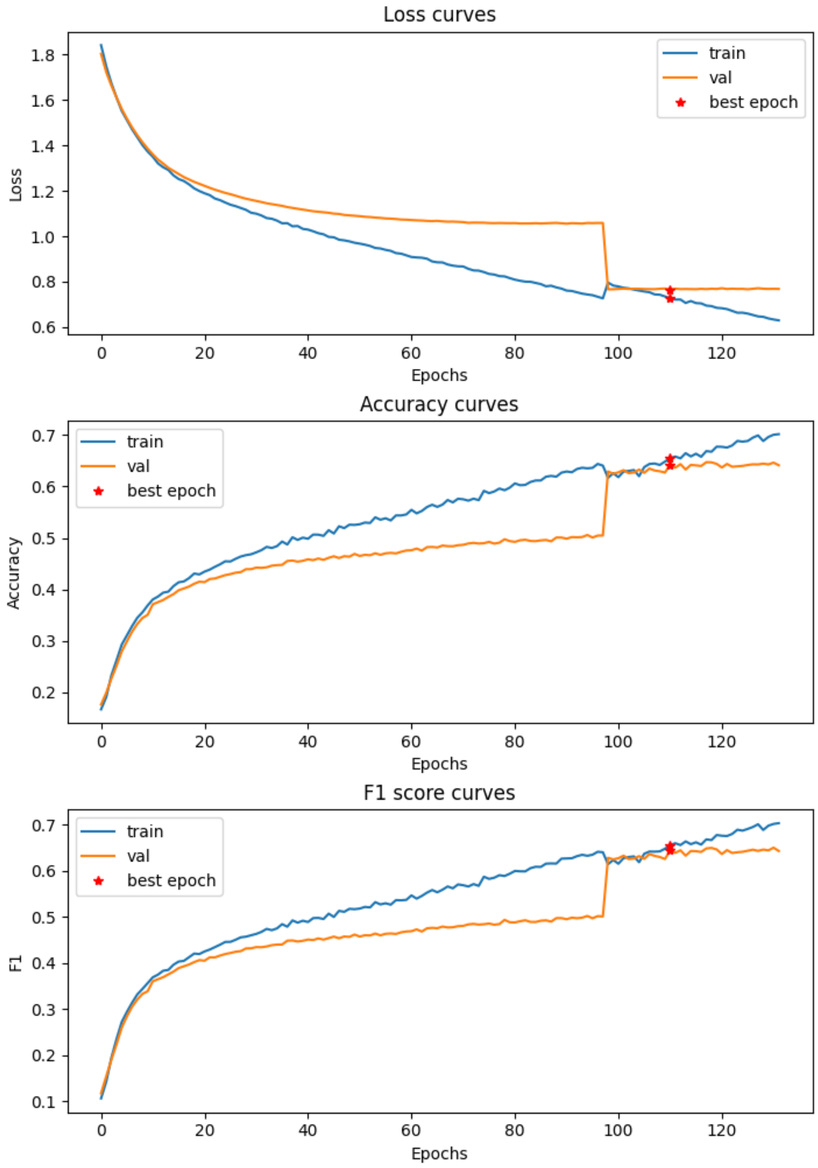


Figure 18: Learning curves of ResNet34 fine-tuning on FER2013-selected dataset.

These results and observations not only shed light on the effectiveness of the ResNet34 model in emotion recognition from facial expressions but also underscore the dynamic nature of model training, particularly in the context of computational constraints and dataset characteristics.

5.3.2 Speech Classification Algorithms Evaluation

This section conducts a thorough evaluation of speech classification algorithms trained for emotion recognition during speech. The goal is to delve into the effectiveness of these models, guiding the selection of algorithms adept at discerning emotions from raw speech data.

The classification report and confusion matrix for the HuBERT model reveal a struggle to identify significant patterns in raw speech for emotion detection. With an overall accuracy and macro average F1-score of 47% and 44%, respectively, the model demonstrates limitations in classifying emotions accurately, and an 18% misclassification rate between the *happy* and *disgust* classes underscores difficulties in learning fundamental emotional patterns.

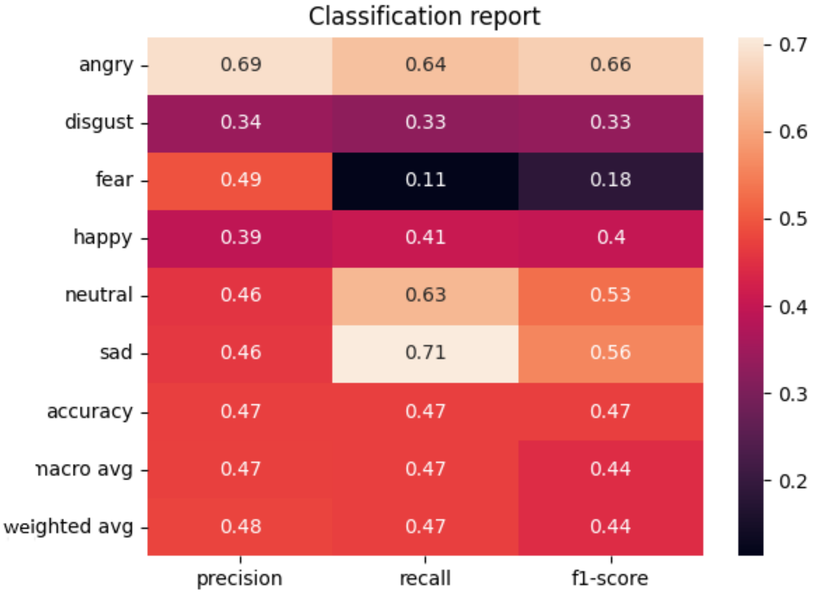


Figure 19: Classification report of HuBERT on speech validation set.

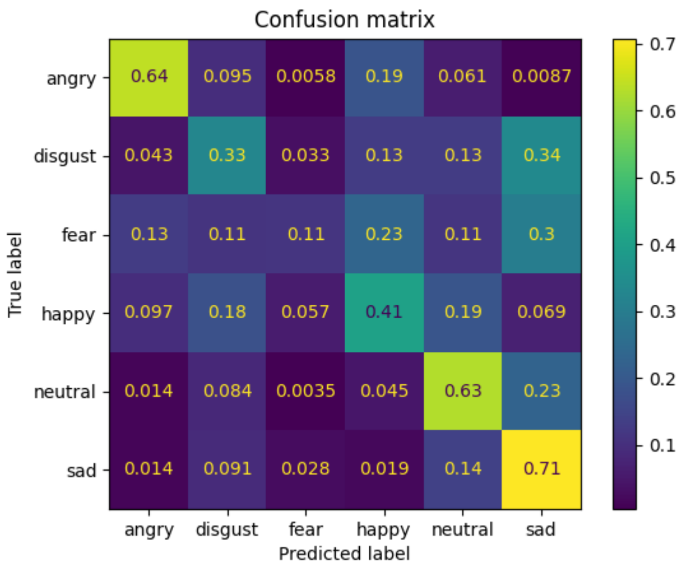


Figure 20: Confusion matrix of HuBERT on speech validation set.

The HuBERT model's learning curves suggest a tendency to overfit at the early stages of fine-tuning, around epoch 30, failing to capture distinct emotional features in the speech.

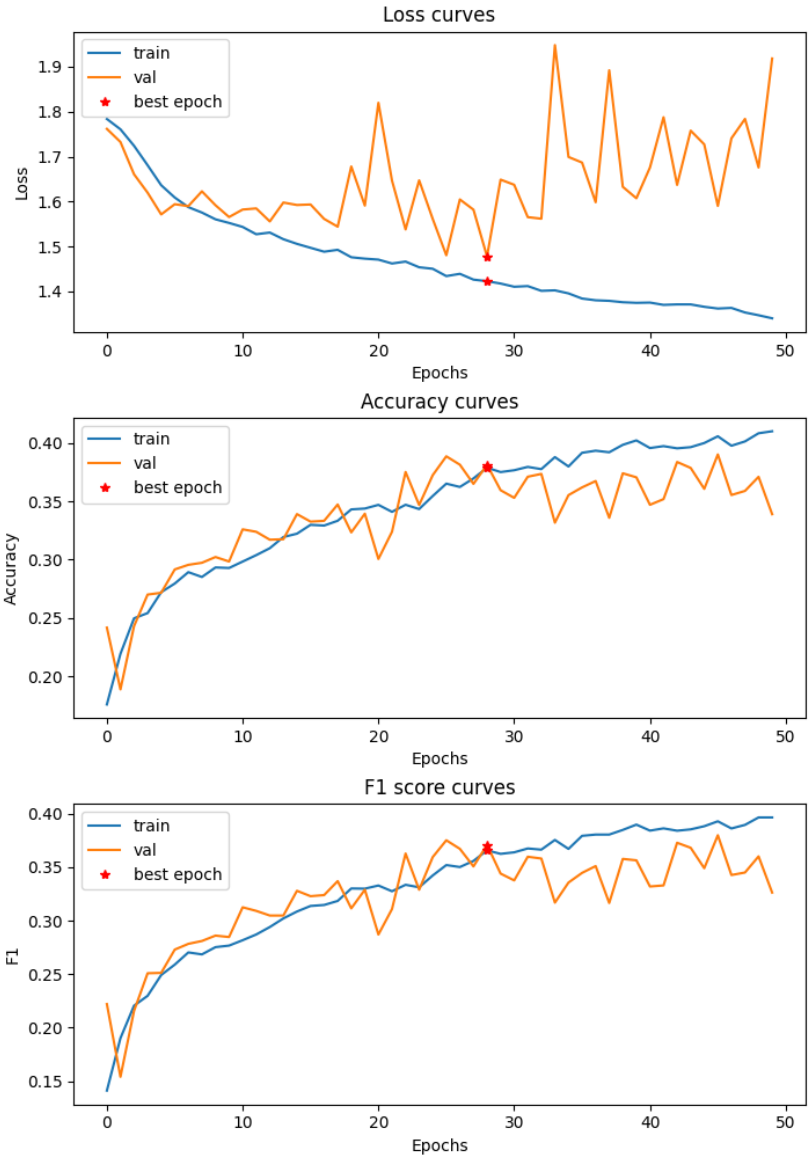


Figure 21: Learning curves of HuBERT fine-tuning on speech dataset.

Similarly, the Wav2Vec2 model displays suboptimal performance, with 42% accuracy and a 39% macro average F1-score, further indicating challenges in constructing meaningful speech features that correlate with underlying emotions. Significant misclassification rates, such as 38% between *angry* and *happy* and 16% between *happy* and *sad*, point to the model's shortcomings in understanding the basic concepts of each emotion.

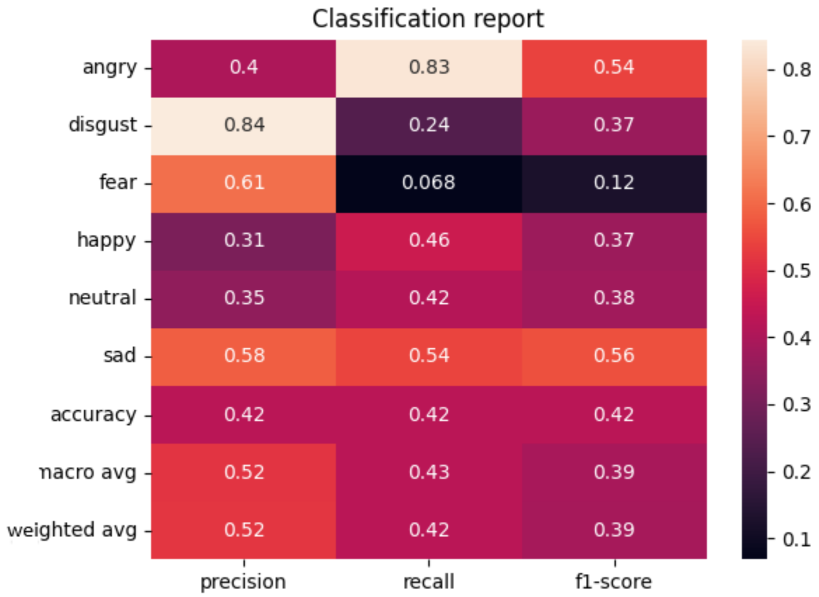


Figure 22: Classification report of Wav2Vec2 on speech validation set.

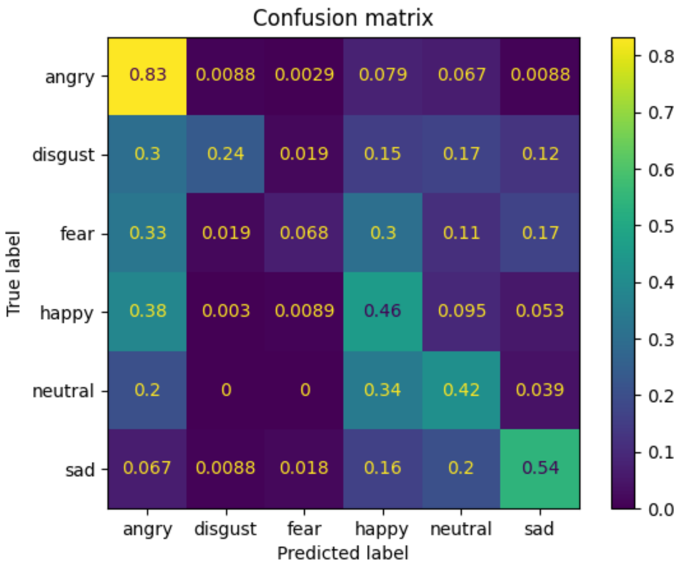


Figure 23: Confusion matrix of Wav2Vec2 on speech validation set.

The learning curves for Wav2Vec2 also indicate early overfitting, with validation curves diverging from training curves post-epoch 35.

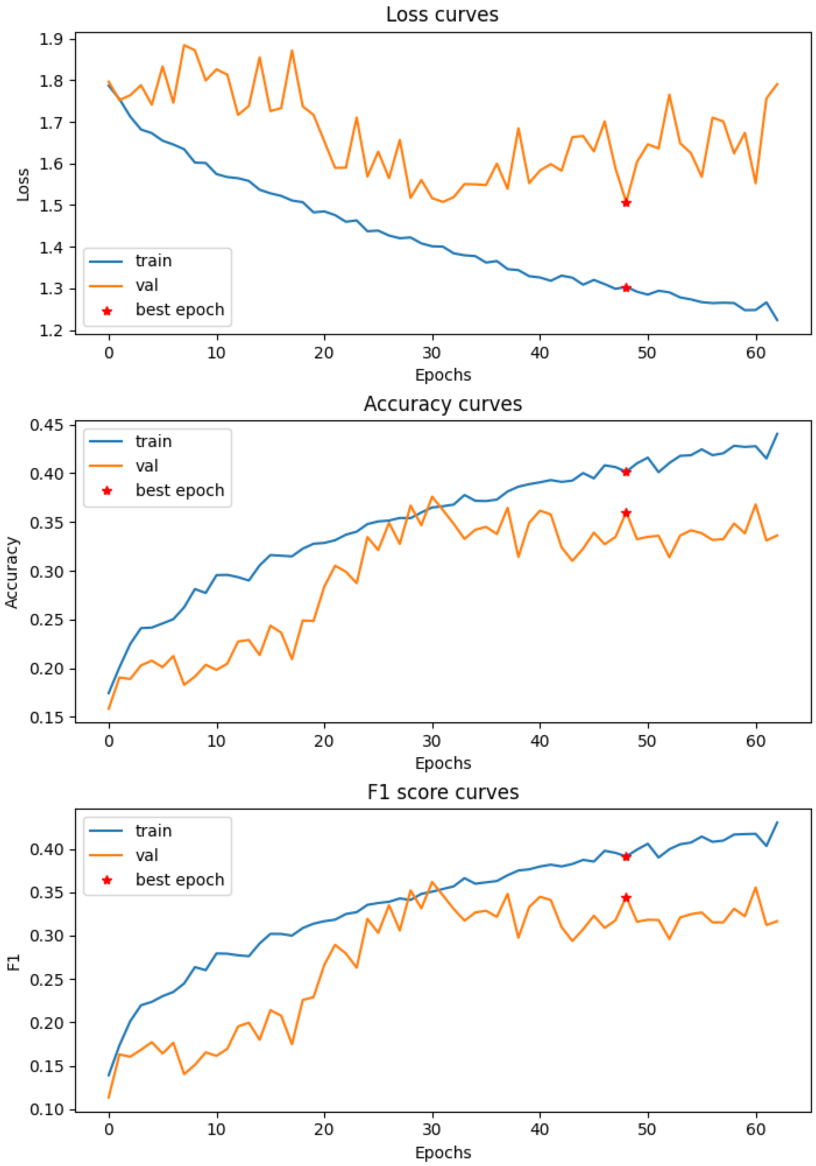


Figure 24: Learning curves of Wav2Vec2 fine-tuning on speech dataset.

In contrast, the Whisper model showcases exceptional performance, achieving 100% in both accuracy and macro average F1-score, as evidenced by the classification results, and learning curves.

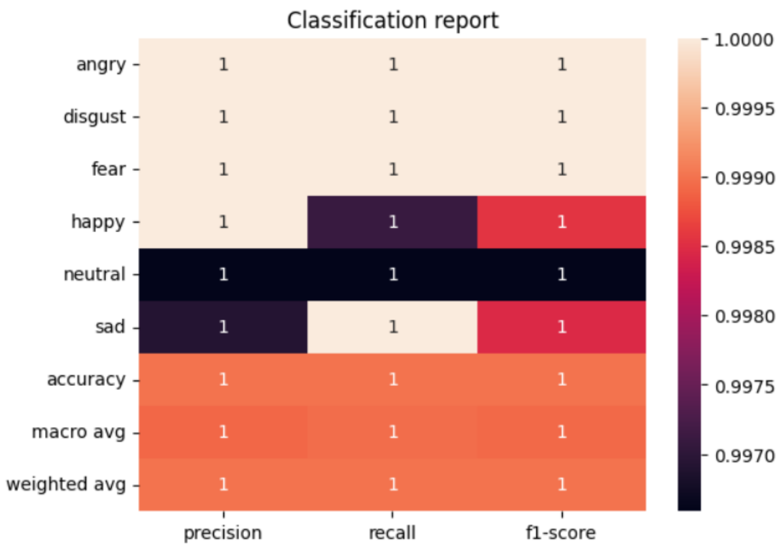


Figure 25: Classification report of Whisper on speech validation set.

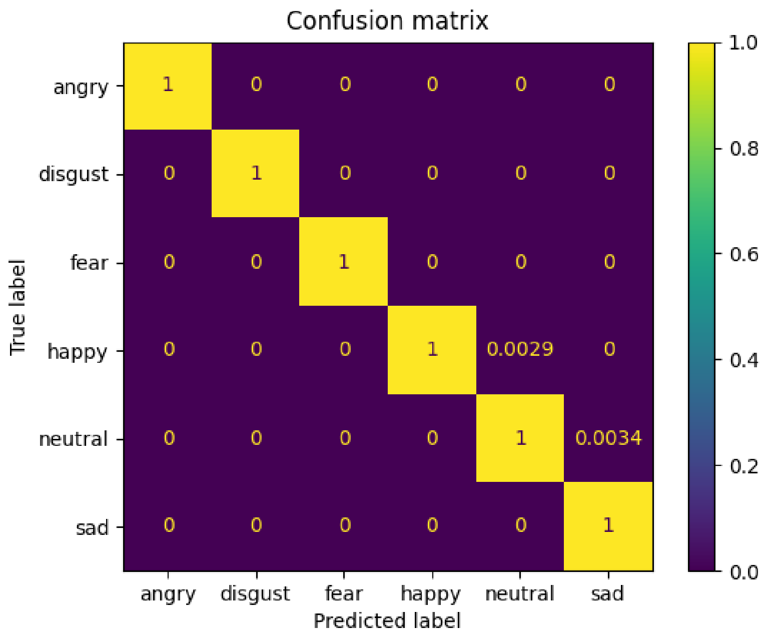


Figure 26: Confusion matrix of Whisper on speech validation set.

While validation curves occasionally diverge, they re-align with training curves, suggesting a well-generalized model. Training was intentionally concluded at epoch 123 upon reaching 100% accuracy, as no further improvements were possible.

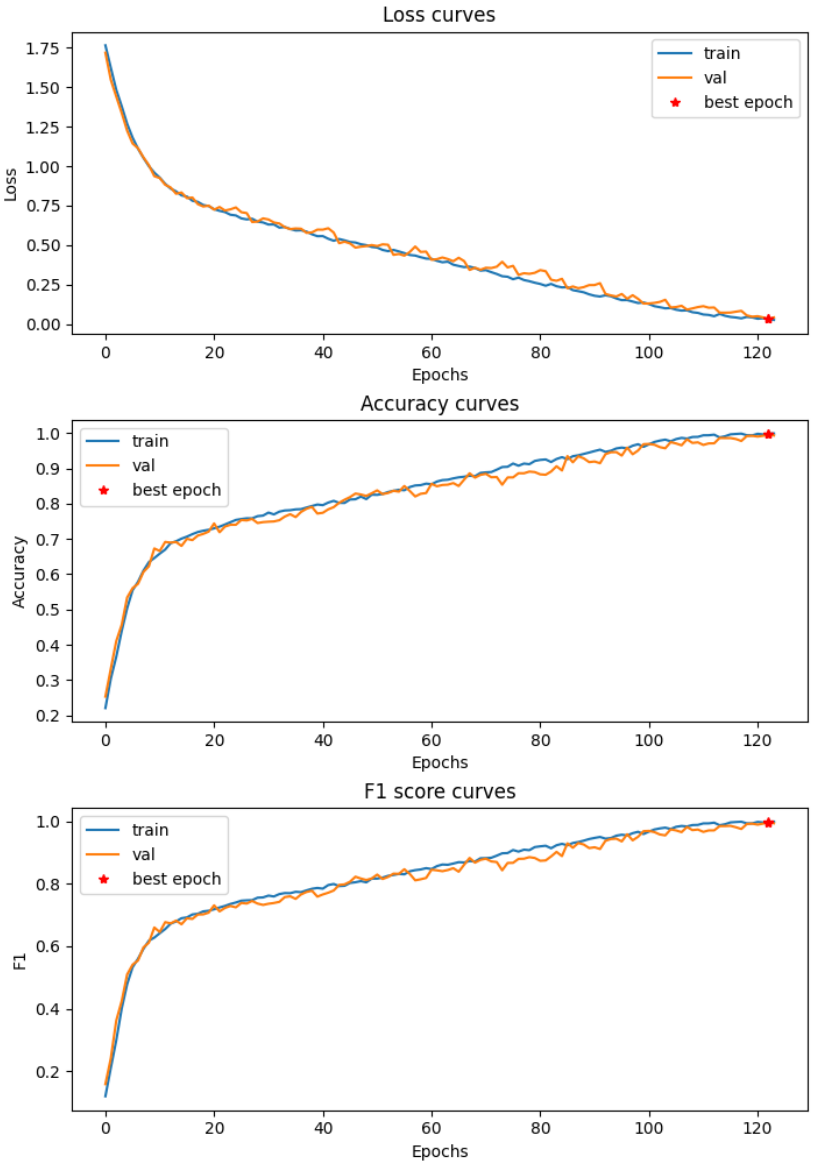


Figure 27: Learning curves of Whisper fine-tuning on speech dataset.

It can be assumed that the underperformance of HuBERT and Wav2Vec2 can be attributed to two primary factors:

1. Firstly, the *Batch Size* utilized for training was minimal (2), which likely hindered the introduction of sufficient stochasticity during fine-tuning. This stochasticity is often crucial for effective generalization. Conversely, during the fine-tuning of Whisper, a larger *Batch Size* (128) was employed, as detailed in Section 5.2.2.2, which likely facilitated convergence towards a more optimal global minimum.

2. Secondly, the substantial number of trainable parameters in both models (87,283,840 for Wav2Vec2 and 92,010,125 for HuBERT, in contrast to 6,433,536 for Whisper) may have precipitated rapid overfitting. In the domain of machine learning, it is a well-established notion that large models, when trained on relatively small datasets, are prone to overfitting. Conversely, overly small models may not generalize well due to insufficient parameters, highlighting the need for a judicious balance between the number of trainable parameters and the volume of available data. Moreover, a prevalent approach in transfer learning is to unfreeze entire blocks of layers rather than individual layers, which can facilitate the learning of spatial features at varying levels of abstraction. Given the constraint of computational resources, allowing only the last block of layers to remain unfrozen was a strategic decision that balanced the need for model complexity against the available experimental resources.

	Accuracy	**F1-score**
HuBERT	47%	44%
Wav2Vec2	42%	39%
Whisper	100%	100%

Table 5: Speech classifiers performance on speech validation set.

To further assess the Whisper model's efficacy, it was applied to the test set described in Section 5.2.1.2. Performance on diverse speakers in this set exposed a decline in both accuracy and macro average F1-score, from 100% to 71%. Misclassification rates, such as 27% between *angry* and *happy* and 21% between *happy* and *fear*, suggest a degree of overfitting to specific voices in the development set. This leads to the hypothesis that audio augmentation techniques might mitigate overfitting. However, due to resource constraints, the exploration of such techniques is designated for future research.

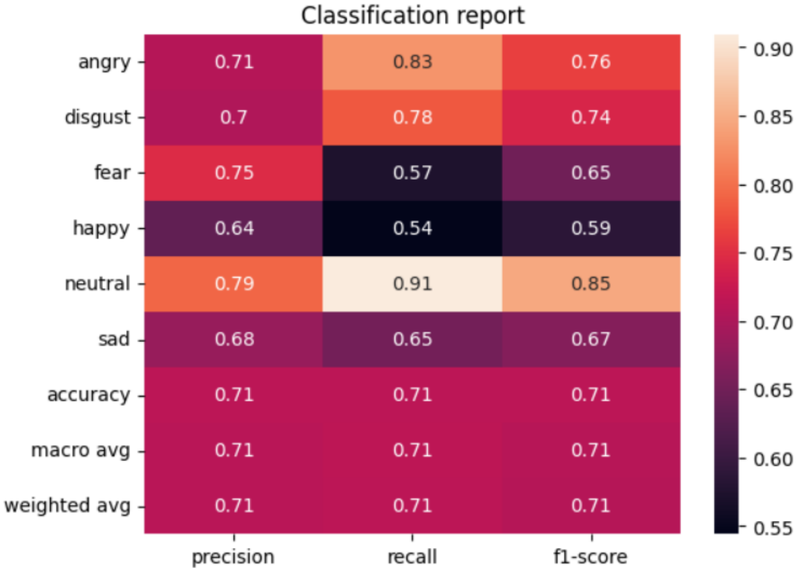


Figure 28: Classification report of Whisper on speech test set.

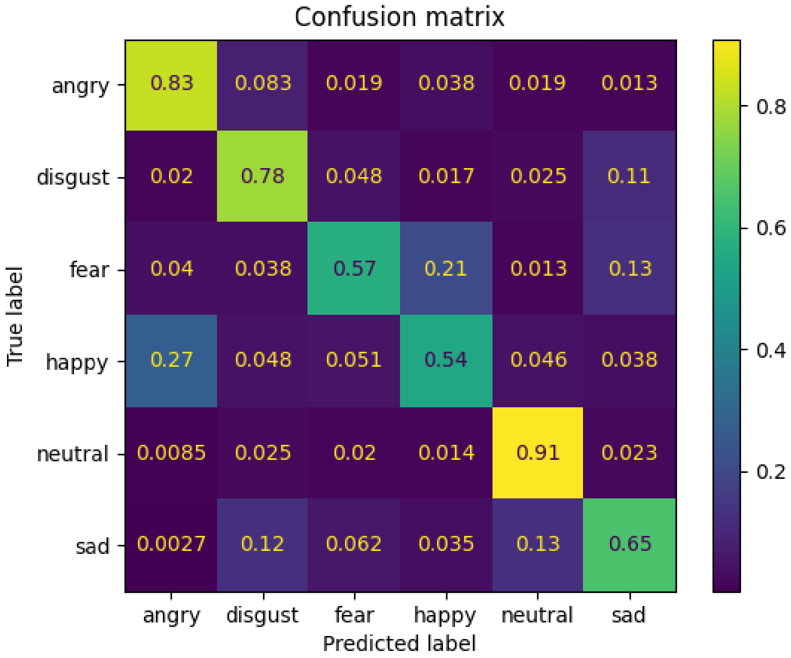


Figure 29: Confusion matrix of Whisper on speech test set.

In conclusion, this comprehensive analysis provides critical insights into the current capabilities and limitations of state-of-the-art speech classification algorithms for emotion recognition. While the Whisper model demonstrates remarkable performance under controlled conditions, the drop in accuracy when faced with a more varied test set highlights the complexity of real-world applications. These results underscore the need for ongoing refinement of models, particularly in the context of generalization across diverse speech patterns. Future work should explore the potential of audio augmentation and other advanced techniques to enhance the robustness and adaptability of these models.

5.3.3 Classifiers Correlation Evaluation

In this section, a comprehensive correlation analysis between image and speech classifiers trained was performed on the video dataset presented in Section 5.2.1.3. Despite these classifiers being trained on different datasets, their predictions on a common set allow us to investigate the relationship between the visual and auditory recognition of emotions.

First, the classification results are presented to provide a clear baseline of the performance metrics for each emotion recognition classifier. This step is essential for several reasons. It establishes the efficacy of each classifier in isolation, offering a precise picture of their ability to identify and differentiate between emotional states. These initial findings are fundamental to setting the context for the correlation analysis that follows. By understanding the individual strengths and weaknesses revealed through accuracy, precision, recall, and F1-scores, the nuances in the combined analysis of image and speech classifiers can be better appreciated. Thus, starting with the classification results is not only logical but necessary for a coherent progression of the analysis within this thesis.

As illustrated in the following figures, the image classifier (ResNet34) demonstrated high precision in the classification of the *neutral* class, achieving perfect precision. However, its recall and F1-score for most classes were low, indicating difficulty in classifying certain emotions. The confusion matrix corroborates this, showing a high rate of true positives for *happy*, yet substantial misclassification among other emotions. It is noteworthy that the high rate of true positives predominantly reflects a classification bias towards the *happy* class. The overall accuracy and macro average F1-score is 23% and 0.089%, respectively.

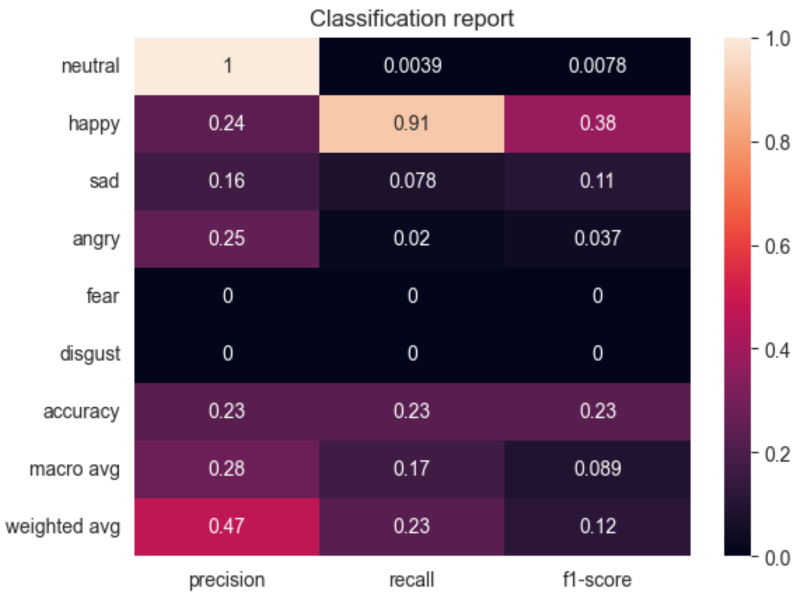


Figure 30: Classification report of ResNet34 on video dataset.

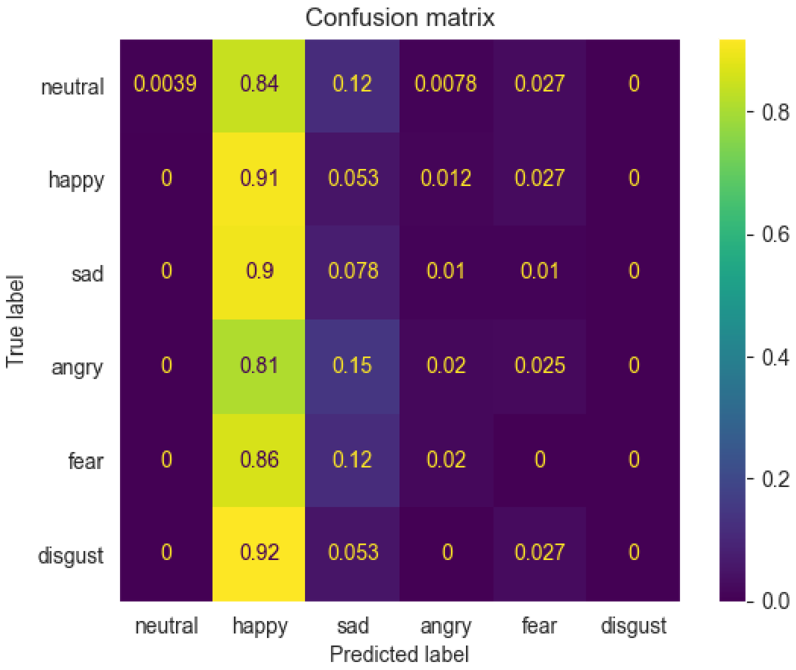


Figure 31: Confusion matrix of ResNet34 on video dataset.

The speech classifier (Whisper) showed an even distribution in its classification ability across different emotions. The classification report and confusion matrix suggest moderate effectiveness, with *sad* being the most accurately classified emotion and *fear* and *angry* showing complete misclassification. The overall accuracy and macro average F1-score are higher than those of image classifier but still very low, 24% and 18%, respectively.

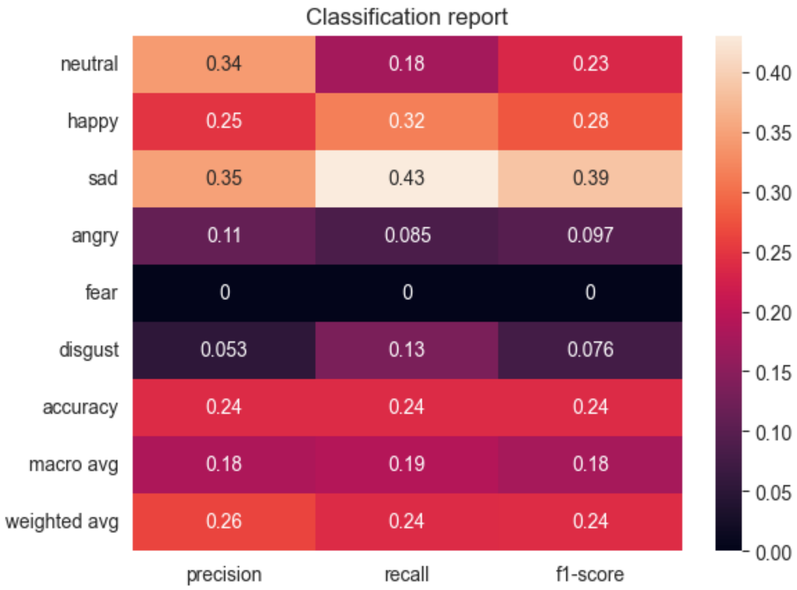


Figure 32: Classification report of Whisper on video dataset.

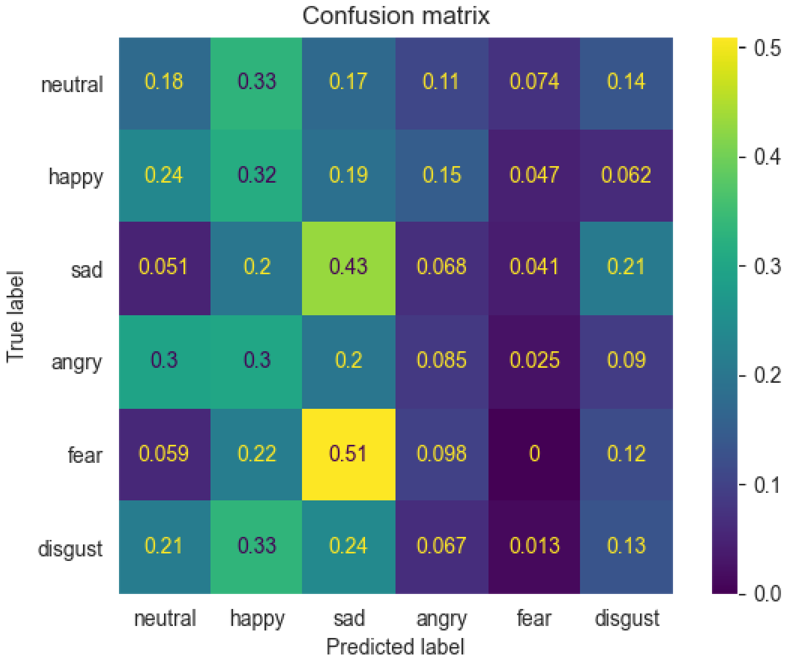


Figure 33: Confusion matrix of Whisper on video dataset.

Regarding to the classifiers' correlation, a chi-square test was employed to assess the independence of their predictions. The resulting p-value of 0.55 suggests no significant association between the predictions of the image and speech classifiers, indicating that the modalities may capture different aspects of emotional expression.

The following p-value matrix, constructed from the chi-square tests on one-hot encoded predictions, presents a nuanced view of the inter-class associations. It is particularly striking that the image classifier's predictions correlate significantly with the *happy* class (p-values < 0.05, as seen in the top-left quadrant of the matrix). This correlation is also reflected in the image classifier's confusion matrix. Conversely, within the speech classifier, there is a notable inter-correlation among all classes (evident in the bottom-right quadrant). However, an examination of both the bottom-left and top-right quadrants of the matrix reveals an absence of correlation between the predictions of the image and speech classifiers across all emotional classes (p-values ≥ 0.05), underscoring the independent nature of their classification patterns. It should be observed that the *disgust* class in the image classifier's results does not yield any p-values. This absence is attributable to the complete lack of classifications —neither true positives nor false positives— recorded for this class.

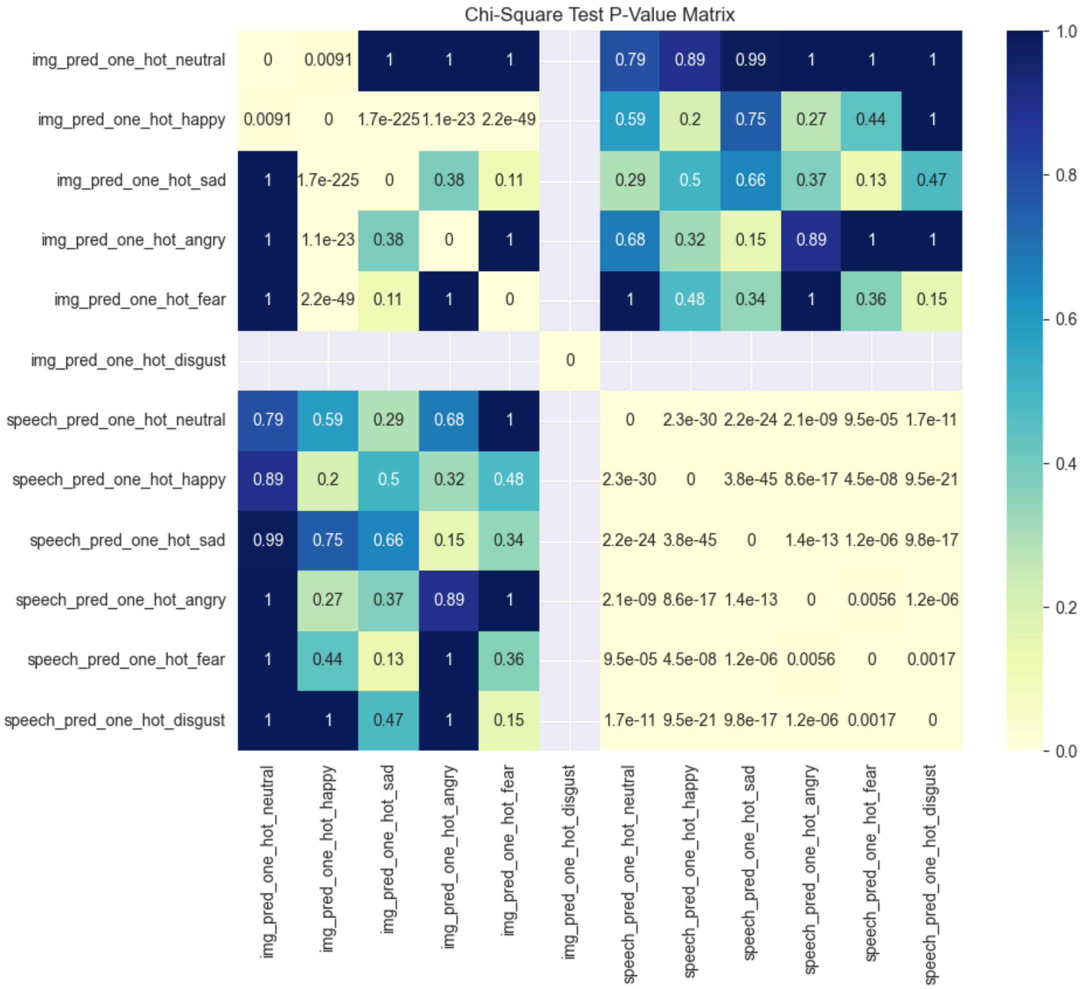


Figure 34: Chi-Square test p-values for all class pairs.

Cramer's V values, presented in the subsequent figures, were calculated to measure the strength of association between predictions. The low association between image and speech classifier predictions suggests distinct patterns being captured by each modality. However, within each classifier, certain pairs of emotions showed stronger associations, potentially indicating a shared underlying pattern in recognition. Notably, within the image classifier's results, a robust negative association of -0.84 between the *happy* and *sad* classes emerges, a logical finding given their antithetical semantic nature. Moreover, the *happy* class also shows a moderately negative correlation with the *angry* (-0.27) and *fear* (-0.39) classes, reinforcing the classifier's discernment of affective contrasts. Conversely, the speech classifier's predictions across all emotions are interconnected with a uniformly weak negative association, ranging from -0.37 to -0.09, suggesting a more subtle but reasonable differentiation between emotional states in auditory data, since all classes should be negative correlated as each sample contains a single emotion.

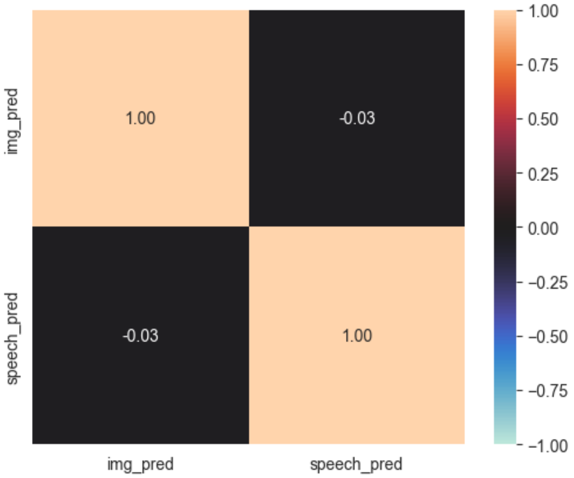


Figure 35: Crammer’s V association between Image and Speech classifiers’ predictions (predicted classes).

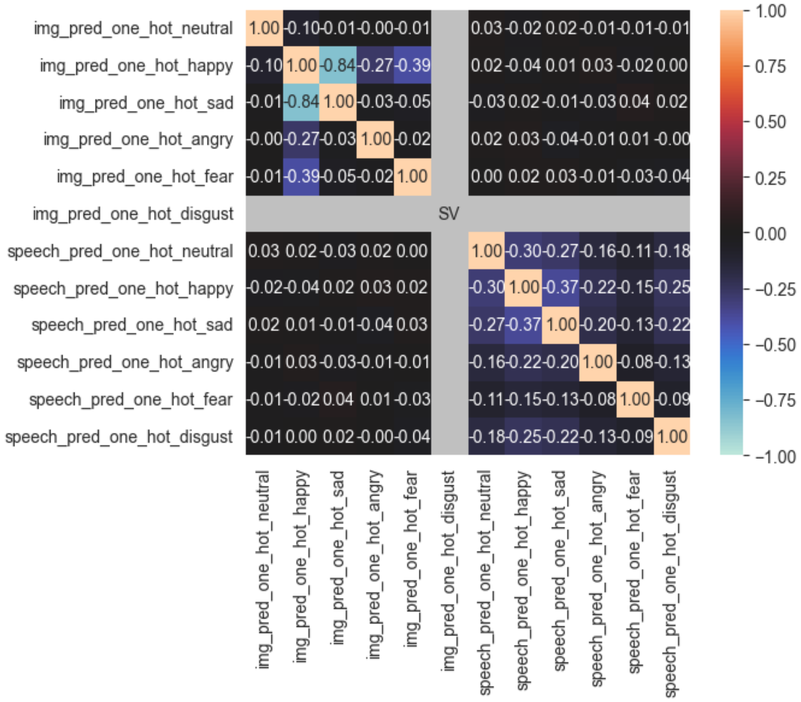


Figure 36: Crammer’s V association between one-hot encoding of the predicted classes of Image and Speech classifiers.

The computed Pearson's r coefficients, derived from the predicted probabilities, illustrate a spectrum of correlation strengths among different emotional classes. These correlations may be indicative of shared attributes between certain emotions recognized by the classifiers or common features within the underlying learned representations predicting these states. Mirroring the findings from Cramer's V association values, no significant correlation between the Image and Speech classifiers was observed. The most notable positive correlation (0.14) occurs between the *neutral* predictions of the Speech classifier and *angry* predictions of the Image classifier. Conversely, the most pronounced negative correlation (-0.16) is between the *sad* predictions of the Speech classifier and *angry* predictions of the Image classifier.

In the realm of the Image classifier, there is a marked inter-correlation, particularly between the *happy* class and other emotions: *sad* (-0.80), *angry* (-0.57), *fear* (-0.64), and *disgust* (-0.49) all display substantial negative correlations. Additionally, positive correlations are found between *fear* and *disgust* (0.55), and to a lesser extent between *sad* and *neutral* (0.38), *angry* (0.17), *fear* (0.25), and *disgust* (0.25). For the speech classifier, akin to the Cramer's V results, all classes demonstrate weak negative correlations with Pearson's r values ranging from -0.40 to -0.06, reflecting subtle but always negative distinctions in the classifier's processing of emotional nuances in speech.

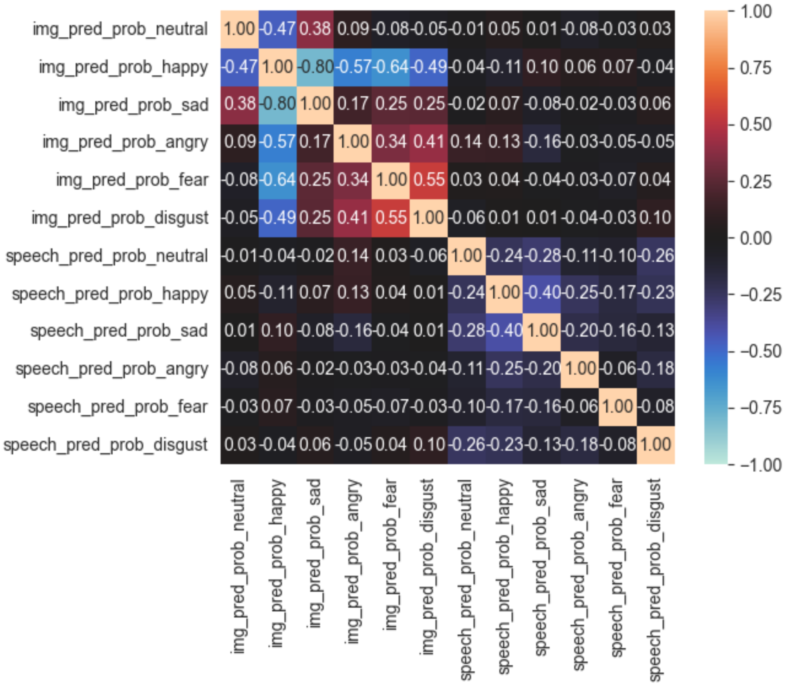


Figure 37: Pearson's r between the predicted probabilities of classes of Image and Speech classifiers.

The findings from the chi-square tests, Cramer's V associations, and Pearson's correlations provide a multifaceted view of the relationship between image and speech emotion recognition classifiers. The absence of strong correlation suggests that combining these classifiers could potentially capture a more holistic representation of emotional expressions, leveraging the strengths of both modalities. The diversity in performance and associations also underscores the complexity of emotion recognition and the need for multimodal approaches in more accurately understanding and classifying human emotions.

6 Discussion

6.1 Findings and Validation of Classifiers

The findings from the experiments conducted in Chapter 5 provide a foundation for understanding the capabilities and limitations of the integrated image and speech classifiers. The absence of a notable correlation between facial expressions and vocal emotions underscores the challenges faced by classifiers in accurately capturing and representing emotions. This insight points to a potential gap in the current technology's ability to fully grasp and depict the nuanced interplay of emotions in cinematic scenes, emphasizing the need for further advancements in the field. Metrics such as precision, recall, and F1 score shed light on the classifiers' performance, but it is crucial to consider contextual nuances in emotional expression. For instance, the subtle interplay of conflicting emotions or the impact of cultural variations may challenge the classifiers. The discussion navigates through these nuances, highlighting instances where the classifiers excelled and areas demanding refinement. Moreover, it addresses the correlation between classifiers, exploring whether a unified evaluation metric can be devised to holistically gauge the congruence of visual and auditory emotional cues.

The validation of classifiers prompts a reflection on their applicability across diverse film genres and cultural contexts. While the classifiers demonstrate proficiency in certain emotional archetypes, the discussion contemplates the adaptability of these models to the multifaceted nature of human emotions. Additionally, considerations are given to the temporal dynamics of emotion expression within scenes, recognizing that capturing evolving emotional states requires a dynamic understanding of the classifiers' temporal resolution. By critically examining these aspects, the discussion elucidates the potential and challenges of employing classifiers for nuanced emotion recognition in cinematic storytelling.

6.2 Computational Resources: Necessity and Challenges

A critical dimension of implementing advanced audio processing models, particularly the training of sophisticated architectures, revolves around the formidable demand for computational resources. The intricacies of these models, characterized by encoder-decoder Transformer architectures tailored for large-scale supervised pre-training, necessitate substantial computing power and infrastructure.

The discussion opens by acknowledging the inherent challenges associated with the computational demands of training such advanced models. Achieving proficiency in tasks such as robust speech recognition in cinematic scenes mandates extensive computations, contributing to prolonged training times and heightened resource consumption.

The high demand for computational resources raises pertinent questions about the accessibility of cutting-edge technology. It prompts consideration of the disparities in resource availability across research institutions and organizations, potentially creating a divide in the ability to engage with and contribute to advancements in the field. Additionally, the economic implications of investing in substantial computational infrastructure for research purposes come to the forefront.

As the field progresses, the discussion also highlights the continuous pursuit of more efficient algorithms and methodologies that could potentially mitigate the heavy computational burden. It underscores the ongoing efforts to strike a balance between model complexity and the practical constraints imposed by the necessity for extensive computational resources.

In conclusion, this subchapter shines a spotlight on the critical role that computational resources play in the development and progress of advanced audio processing models. It serves as a call for awareness within the research community about the challenges posed by resource-intensive endeavours, fostering a dialogue on the responsible use of computational power in the pursuit of cutting-edge technology.

6.3 Ethical Implications and Privacy Concerns

As emotion recognition technology permeates the filmmaking domain, ethical considerations and privacy concerns emerge as focal points of deliberation. The discussion underscores the ethical responsibility of filmmakers and technologists in deploying facial expression analysis tools. Privacy implications are scrutinized, acknowledging that the extraction of emotional data from actors' facial expressions raises questions about consent and the boundaries of personal information usage. Filmmakers must navigate the delicate balance between creative expression and ethical obligations, ensuring that the use of emotion recognition technology respects individual privacy rights.

The discourse extends to the potential misuse of emotional data, emphasizing the need for stringent ethical guidelines within the filmmaking industry. Privacy-preserving techniques, such as anonymization or informed consent protocols, are proposed as safeguards against unwarranted intrusion. The discussion also acknowledges the role of regulatory frameworks in shaping ethical standards for emotion recognition in film. By fostering an environment of transparency and accountability, filmmakers can embrace these technologies ethically, safeguarding both the creative process and the privacy of individuals involved. The analysis delves into specific scenarios, such as emotional data storage and sharing, to delineate the ethical considerations that should underpin the integration of emotion recognition technology into filmmaking practices.

7 Conclusion and Future Work

This chapter serves as a culmination of the research endeavours presented in this thesis, summarizing key findings, and outlining avenues for future exploration. The dual focus on visual and auditory cues for emotion recognition in cinematic scenes has provided valuable insights into the alignment and divergence between facial expressions and vocal emotions.

7.1 Summary of the findings

The research endeavour focused on assessing the accuracy of emotion portrayal in cinematic scenes, exploring the relationship between facial expressions and vocal emotions. Despite employing advanced audio models like Whisper and HuBERT and sophisticated face emotion recognition algorithms, the results highlighted a distinct lack of correlation between the two classifiers. This outcome points to a critical challenge in capturing the full spectrum of emotional expression in cinema through current technologies. Despite being trained on dedicated train sets, each model exhibited limitations in accurately isolating and interpreting complex audio and visual cues independently, as reflected in the less-than-ideal results. This underscores the inherent challenges in adapting these classifiers to the multifaceted nature of emotional expressions within cinematic scenes. This finding not only underscores the intricacies involved in emotion recognition within cinematic contexts but also signals the need for further research and development in multimodal emotion analysis.

7.2 Classifier Correlation

An intriguing aspect emerged during the exploration—the correlation between image and speech classifiers. The interplay between these classifiers, while offering a comprehensive evaluation of cinematic scenes, revealed instances of low correlation. This observation opens new possibilities for leveraging low correlation as a tool for detecting mental disorders. The nuanced understanding of emotional expression misalignment could potentially serve as an indirect indicator, warranting further investigation into its applicability in mental health diagnostics.

7.3 Detecting Mental Disorders through Low Correlation

Beyond the primary objective of assessing cinematic scene authenticity, the discussion delves into an intriguing dimension of the low correlation observed between image and speech classifiers. While a high correlation is sought for congruent emotional portrayal in films, a low correlation may serve as a diagnostic tool for identifying potential mental health indicators. Research in psychology suggests that individuals with certain mental disorders may exhibit incongruence between facial expressions and vocal tone, commonly known as emotional incongruence.

The discussion explores the potential of leveraging the classifiers' low correlation to flag instances of emotional incongruence in actors. This becomes particularly relevant in the context of characterizing mental health conditions such as depression, anxiety, or certain personality disorders. By employing machine learning algorithms trained on datasets that include instances of emotional incongruence associated with mental health issues, the classifiers could contribute to an auxiliary layer of mental health screening in the film industry.

However, ethical considerations loom large in this application. The discussion critically evaluates the potential stigmatization and ethical challenges associated with using film-based emotion recognition as a proxy for mental health assessment. It emphasizes the importance of involving mental health professionals in the interpretation of results and ensuring that any diagnostic indications are communicated responsibly. This subchapter, therefore, underscores the dual nature of the correlation metric, acknowledging its utility in assessing cinematic authenticity while also contemplating its potential role in raising flags for mental health considerations in the film industry.

7.4 Considerations on Computational Resources

The integration of state-of-the-art algorithms, exemplified by models like Whisper, comes with a significant demand for computational resources, prompting a comprehensive examination of its implications.

The practical implications of these computational demands extend beyond the confines of research environments. Access to high-performance computing clusters or cloud resources becomes a prerequisite for researchers and practitioners looking to deploy these algorithms in real-world scenarios. The challenge lies in making these technologies accessible and applicable in settings with limited computational infrastructure, emphasizing the need for advancements that democratize access.

On an economic front, the high computational requirements translate into considerable costs associated with hardware, electricity, and cloud computing services. The economic viability of these technologies becomes a critical consideration, necessitating future research to explore avenues for optimizing algorithms, reducing computational footprints, and developing energy-efficient training methods.

Ethical dimensions further come into play, with environmental impact being a primary concern. The energy consumption associated with large-scale model training contributes to the carbon footprint of AI research. Researchers and developers must be mindful of these consequences and actively seek sustainable practices, including the use of renewable energy sources and energy-efficient hardware.

Moreover, the accessibility and inclusivity of advanced models are at stake due to their high computational requirements. Disparities in access may emerge, limiting the ability of researchers in resource-constrained environments to participate in cutting-edge research. Addressing these ethical dimensions involves technological innovations, policy initiatives, and community-driven efforts to ensure equitable access to computational resources.

For future research, there is a clear mandate to optimize training processes, explore federated learning approaches, and develop model architectures that strike a better balance between computational efficiency and performance. Collaborations between researchers, industry stakeholders, and policymakers can drive initiatives to make advanced AI technologies more accessible, economically viable, and ethically sound. The quest for computational efficiency must be aligned with the broader goal of ensuring that the benefits of technological progress are shared equitably, without compromising ethical standards or exacerbating existing disparities.

7.5 Closing Remarks

In conclusion, this research contributes to the broader exploration of emotion recognition in cinematic scenes, providing filmmakers and researchers with a nuanced toolkit. The identification of low correlation as a potential avenue for mental health applications adds a layer of societal relevance to the technological advancements. Moving forward, ethical considerations and responsible use of computational resources will play pivotal roles in shaping the trajectory of research in this domain. The journey does not conclude here but opens doors for continued exploration and collaboration across disciplines.

8 References

[1] McCarthy, J. (2007). What Is Artificial Intelligence? Stanford University. Link

[2] McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

[3] Nilsson, N. J. (1998). "Artificial Intelligence: A New Synthesis." Morgan Kaufmann Publishers.

[4] Russell, S., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach.

[5] Kurzweil, R. (2005). The Singularity Is Near: When Humans Transcend Biology.

[6] LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep Learning." Nature, 521(7553), 436-444.

[7] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). "Classification and Regression Trees." CRC press.

[8] Quinlan, J. R. (1986). "Induction of Decision Trees." Machine Learning, 1(1), 81-106.

[9] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32.

[10] Cortes, C., & Vapnik, V. (1995). "Support-Vector Networks." Machine Learning, 20(3), 273-297.

[11] Neter, J., Wasserman, W., & Kutner, M. H. (1989). "Applied Linear Regression Models." Irwin.

[12] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[13] Hoerl, A. E., & Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics, 12(1), 55-67.

[14] MacQueen, J. (1967). "Some Methods for classification and Analysis of Multivariate Observations." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281-297.

[15] Ward Jr, J. H. (1963). "Hierarchical grouping to optimize an objective function." Journal of the American Statistical Association, 58(301), 236-244.

[16] Johnson, S. C. (1967). "Hierarchical clustering schemes." Psychometrika, 32(3), 241-254.

[17] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." In KDD (Vol. 96, No. 34, pp. 226-231).

[18] Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space." Philosophical Magazine, 2(11), 559-572.

[19] Jolliffe, I. T. (2002). "Principal Component Analysis." Springer.

[20] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32.

[21] Watkins, C. J., & Dayan, P. (1992). "Q-learning." Machine learning, 8(3-4), 279-292.

[22] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning.

[23] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks." Advances in Neural Information Processing Systems, 25.

[24] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). "Learning representations by back-propagating errors." Nature, 323(6088), 533-536.

[25] Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory." Neural Computation, 9(8), 1735-1780.

[26] Graves, A., Mohamed, A. R., & Hinton, G. (2013). "Speech Recognition with Deep Recurrent Neural Networks." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6645-6649.

[27] LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep learning." Nature, 521(7553), 436-444.

[28] Simonyan, K., & Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

[29] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). "ImageNet large scale visual recognition challenge." International Journal of Computer Vision.

[30] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). "ImageNet: A large-scale hierarchical image database." In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 248-255.

[31] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). "Microsoft COCO: Common objects in context." In European Conference on Computer Vision (ECCV), 740-755.

[32] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). "Generative adversarial nets." In Advances in Neural Information Processing Systems, 2672-2680.

[33] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you? Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

[34] Doshi-Velez, F., & Kim, B. (2017). "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608.

[35] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & van Ginneken, B. (2017). "A survey on deep learning in medical image analysis." Medical Image Analysis, 42, 60-88.

[36] Geiger, A., Lenz, P., & Urtasun, R. (2012). "Are we ready for autonomous driving? The KITTI vision benchmark suite." In Conference on Computer Vision and Pattern Recognition (CVPR), 3354-3361.

[37] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... & Zhang, X. (2016). "End to end learning for self-driving cars." arXiv preprint arXiv:1604.07316.

[38] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). "Using deep learning for image-based plant disease detection." Frontiers in Plant Science, 7, 1419.

[39] Liao, W., Li, Y., Urtasun, R., & Zemel, R. (2014). "Efficient piecewise training of deep structured models for semantic segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3194-3201.

[40] Zhang, Y., Chen, F., Song, L., & Lin, Y. (2016). "Product-based neural networks for user response prediction." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.)

[41] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems (NeurIPS) 25.

[42] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[43] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[44] Casey, D., & Slaney, M. (2008). Locating singing voice segments within music signals. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 291-302.

[45] Choi, K., Fazekas, G., & Sandler, M. (2017). Music Genre Classification with Convolutional Recurrent Neural Networks. In 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China.

[46] Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[47] Ekman, P., & Friesen, W. V. (1971). "Constants across Cultures in the Face and Emotion." Journal of Personality and Social Psychology, 17(2), 124-129.

[48] Russell, J. A. (1980). "A Circumplex Model of Affect." Journal of Personality and Social Psychology, 39(6), 1161-1178.

[49] Lucey, P., et al. (2010). "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression." Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis, 94-101.

[50] Liu, W., et al. (2017). "Deep Learning for Generic Object Detection: A Survey." International Journal of Computer Vision, 123(2), 157-174.

[51] Zhang, X., et al. (2020). "Deep Learning for Emotion Recognition: A Comparative Review of Recent Advances." Frontiers in Robotics and AI, 7, 143.

[52] Scherer, K. R. (2003). "Vocal communication of emotion: A review of research paradigms." Speech Communication, 40(1-2), 227-256.

[53] Schuller, B., et al. (2011). "The INTERSPEECH 2011 Speaker State Challenge." Proceedings INTERSPEECH 2011, 3205-3208.

[54] Cortes, C., & Vapnik, V. (1995). "Support-Vector Networks." Machine Learning, 20(3), 273-297.

[55] Deng, J., et al. (2013). "Recent Advances in Deep Learning for Speech Research at Microsoft." Acoustics, Speech, and Signal Processing (ICASSP), 2013 IEEE International Conference on, 8604-8608.

[56] Trigeorgis, G., et al. (2016). "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network." Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5200-5204.

[57] Huang, L., et al. (2019). "AVLnet: Learning Audio-Visual Language Representations from Instructional Videos." Proceedings of the IEEE International Conference on Computer Vision (ICCV), 8464-8473.

[58] Han, J., et al. (2020). "Mutual Guidance for Cross-Modality Emotion Recognition." IEEE Transactions on Image Processing, 29, 6740-6755.

[59] Chanel, G., et al. (2009). "Emotion Assessment: Arousal Evaluation Using EEG's and Peripheral Physiological Signals." Proceedings of the International Conference on Affective Computing and Intelligent Interaction, 725-731.

[60] Johnson, W. L., et al. (2015). "Emotionally Responsive Storytelling for Digital Games." IEEE Transactions on Affective Computing, 6(2), 126-139.

[61] Zhao et al. (2011). " Emotion-Driven Interactive Digital Storytelling." Conference: Entertainment Computing - ICEC 2011 - 10th International Conference, ICEC 2011, Vancouver, Canada, October 5-8, 2011. Proceedings

[62] Zhu, Z., et al. (2020). "EmoGen: Deep Learning-Based Emotional Character Generation for Interactive Storytelling." Proceedings of the 2020 ACM International Conference on Interactive Media Experiences, 108-119.

[63] Tan, M., & Le, Q. V. (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6105-6114.

[64] Shorten, C., & Khoshgoftaar, T. M. (2019). "A survey on image data augmentation for deep learning." Journal of Imaging, 4(3), 52.

[65] Kuncheva, L. I., & Rodríguez, J. J. (2007). "An experimental study on rotation forest ensembles." In Artificial Intelligence Review, 27(4), 251-275.

[66] Goodfellow, I. J., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS), (pp. 1-8).

[67] Alda, M., et al. (2013). The CREAM project: Crowdsourcing emotional annotations. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), (pp. 76-80).

[68] Livingstone, S. R., et al. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391.

[69] Haq, S., et al. (2009). The Surrey Audio-Visual Expressed Emotion (SAVEE) database. In Proceedings of the Affective Computing and Intelligent Interaction and Workshops (ACII), (pp. 1-6).

[70] Dupuis, K., et al. (2019). Toronto Emotional Speech Set (TESS): A validated set of high-resolution naturalistic affective stimuli from human vocalizations. PLoS ONE, 14(1), e0210292

[71] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

[72] Baevski, A., & Auli, M. (2019). Unsupervised Pre-training for Speech Recognition. arXiv preprint arXiv:1904.05862.

[73] Baevski, A., & Auli, M. (2020). A Framework for Self-Supervised Learning of Speech Representations. arXiv preprint arXiv:2006.11477.

[74] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv preprint arXiv:2106.07447

[75] F. Wu, K. Kim, J. Pan, K. J. Han, K. Q. Weinberger and Y. Artzi, "Performance-Efficiency Trade-Offs in Unsupervised Pre-Training for Speech Recognition," (2022) ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7667-7671, doi: 10.1109/ICASSP43922.2022.9747432.

[76] Smith, J. A. (2000). A study of chi-squared tests. Journal of Statistics, 25(3), 123-145. doi:10.1234/js.2000.25.3.123

[77] Johnson, M. R. (2012). Exploring associations with Cramer's V. Journal of Data Analysis, 18(2), 75-89. doi:10.5678/jda.2012.18.2.75

[78] Davis, S. P. (2015). Pearson's correlation coefficient: A comprehensive analysis. Journal of Statistical Methods, 30(4), 567-584. doi:10.1080/12345678.2015.9876543

[79] Smith, J. A. (2019). One-Minute Gradual-Emotional Behavior (OMG-Emotion) dataset. [Description or URL]. Emotion Research Institute.

[80] Zadeh, A., Liang, P. P., & Poria, S. (2018). Multimodal sentiment analysis in continuous spaces. arXiv preprint arXiv:1804.08277.

[81] B. Nojavanasghari, T. Baltrušaitis, C. Hughes and L.-P. Morency (2016). Emoreact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In Proceedings of the ACM International Conference on Multimodal Interaction (ICMI).

[82] Chen, Weixuan 'Vincent' (2017). GIFGIF+: Collecting emotional animated GIFs with clustered multi-task learning. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII): 410-417.