# Facial Emotion Recognition by Using Mini-Xception and Ensemble Learning

GuanQun Xu

A thesis submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2023

School of Engineering, Computer & Mathematical Sciences

# Abstract

In this thesis, we provide an innovative approach to Facial Emotion Recognition (FER), which is use of ensemble learning with a lightweight model, mini-Xception. Compared to single lightweight models, it made a significant improvement. For a solution, we proposed an ensemble of mini-Xception models, where each expert is trained for a specific emotion and let confidence score for voting. Therefore, the expert model will transform the original multiclass task into binary tasks. We target the model to differentiate between a specific emotion and all others, facilitating the learning process. The principal innovation lies in our confidence-based voting mechanism, in which the experts "vote" based on their confidence scores rather than binary decisions. The confidence scores can be adjusted based on the strengths and weaknesses of the dataset to achieve maximum optimization based on the model.We found that these adjustments to confidence scores can be effectively applied to datasets, and we applied this method to FER2013 dataset, resulting in 72.8% accuracy.

Furthermore, while we found the imbalance between emotion datasets, we introduced data augmentation methods, through oversampling positive samples to improve training effectiveness. Contrasted with the conventional mini-Xception model, our ensemble learning method showcased superior robustness, especially in ambiguous scenarios. This research project not only contributes a novel methodology to the FER domain but also devotes to promising avenues for real-time applications and devices with limited computational resources.

**Keywords**: Facial Emotion Recognition, mini-Xception, Computer vision, Deep learning, Confidence-based voting, Machine learning, Ensemble learning

# Table of Contents

# List of Figures

# List of Tables

# Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Date: 25th Sep 2023

# Acknowledgment

First, I would like to thank my parents for their financial support. Owing to the unselfish and generous sponsor from them, I have this invaluable opportunity to complete my master's study with the Auckland University of Technology (AUT), New Zealand.

I would also like to express my deepest gratitude to my primary supervisor Wei Qi Yan. In this study, he not only provided me with professional knowledge support and careful guidance, but also helped me enrich my learning experience. I believe I could not complete my study without Dr Yan's supervision and instructions. I would also like to thank my wife Min for her understanding and care during this time. The success of my studies would not have been possible without her.

GuanQun Xu

Auckland, New Zealand

September 2023

# Chapter 1 Introduction

*Emotion recognition, a rapidly growing interdisciplinary field, interweaves the realms of artificial intelligence, psychology, and human-computer interaction. As deep learning permeates deeper into our daily lives, its ability to understand and respond to human emotions becomes an imperative. Such comprehension has the potential to revolutionize sectors like healthcare, entertainment, and education, tailoring experiences to individual emotional states. This chapter serves as a prologue, providing a concise overview of the emotion recognition landscape. It underscores the importance, potential applications, and the challenges that researchers face in this domain. The following chapter will delve into the methodologies, models, and innovations, such as the weighted confidence score approach, which seeks to bolster the accuracy and reliability of emotion predictions. As our journey through this exploration, people will gain insights into the intricate interplay between technology and human emotion, a relationship that holds the promise of reshaping the future of human-machine interactions.*

## 1.1 Background and Motivation

### 1.1.1 Facial Emotion Recognition in Deep Learning

Facial Emotion Recognition (FER) is the task of identifying human emotions from facial expressions. This task, currently lying at the crossroads of psychology and computer science, has grown immensely with the advent of machine learning and more specifically in deep learning. Historically, understanding and interpreting human emotions were subjective, relying heavily on human intuition and judgment. However, with the increasing integration of technology into our daily lives, objective identification of emotions through machines becomes not only desirable but, in many scenarios, it is essential.

The advancement of Human-Computer Interaction (HCI) systems is intimately tied to the capability of computers to automatically and accurately discern human emotions. This burgeoning area of study, known as affective computing, is pushing the frontier of interactive technologies. At its core, human emotion is an intricate interplay of feelings, cognitions, and behaviours, reflecting an individual psychophysiological response to both internal and external events. These emotional states significantly influence decision-making, perception, and interpersonal communication.

The potential applications of affective computing are vast. By enabling HCI systems to comprehend and adapt to the emotional states of human users in real-time, the dynamics of human-machine interaction can be elevated to be more intuitive and user-centric. For instance, in product design and user experience domains, real-time emotional feedback can guide refinements, ensuring an enhanced user experience. In high-risk sectors like military and aerospace, continuous monitoring of the emotional states of soldiers, pilots, or astronauts can provide crucial insights. Similarly, public transportation systems can leverage emotion recognition to monitor drivers' emotional states, aiming to mitigate risks associated with driving under extreme emotional duress.

While delving into the nature of emotions, the basic emotion theory posits that humans universally experience several foundational emotions, namely happiness, sadness,

fear, anger, disgust, and surprise. These fundamental emotional states can be seen as building blocks, from which more nuanced emotions—such as fatigue, anxiety, or satisfaction—emerge. A prominent figure in this field, Ekman, articulated that these basic emotions are underlined by certain characteristics: They are instinctual in origin, elicited similarly across individuals in analogous situations, expressed in comparable manners across diverse populations, and associated with consistent physiological patterns. Beyond these six basic emotions, an array of compound emotions exists, which include feelings like shyness, guilt, and contempt, further illustrating the richness and complexity of human emotional landscapes. (Zhang et al., 2020)

The inception of Convolutional Neural Networks (CNNs) can be traced back to the 1980s. However, for a significant duration, they remained relatively sidelined by the predominant computer vision and machine learning disciplines until the pivotal ImageNet competition in 2012. The re-emergence and remarkable outcomes attributed to CNNs are credited to advancements like the proficient deployment of GPUs, the introduction of Rectified Linear Units (ReLUs), the implementation of dropout strategies, and the evolution of data augmentation methodologies (LeCun, Bengio, & Hinton, 2015). Consequently, CNNs have ushered in a transformative era in computer vision. In contemporary times, they stand as the foremost methodology for recognition and detection endeavours, occasionally rivalling human capabilities in specific scenarios.

Machine learning, a subset of artificial intelligence, employs algorithms that allow systems to learn patterns from data without being explicitly programmed. FER has leveraged machine learning substantially in the last couple of decades, moving from rudimentary methods to more sophisticated algorithms. Initial models were based on geometric features, capturing key facial landmarks and their relative positions. These models would then infer emotions based on shifts and changes in these landmarks – a raised eyebrow or a curled lip, for instance. Although it is effective to some extent, these approaches were highly sensitive to nuances and minor deviations.

However, the true revolution in FER came with the proliferation of deep learning, a subset of machine learning that employs neural networks with many layers (deep architectures). Convolutional Neural Networks (CNNs), for instance, have proven to be

particularly adept at handling image data due to their inherent design, which captures spatial hierarchies in images. Unlike traditional machine learning methods that often required handcrafted feature extraction from facial images, CNNs automatically learn and extract features from raw pixel values, while offering superior performance.

Deep learning, a subfield of machine learning, utilizes neural networks with many layers to analyse various forms of data. Among the numerous layers employed in deep learning architectures, a group of layers play pivotal roles in feature extraction and transformation.

The convolution layer is the cornerstone of Convolutional Neural Networks (CNNs), which are particularly effective for tasks like image recognition. In this layer, filters (or kernels) are convolved with the input data to produce feature maps, highlighting certain features in the input. The number and size of filters, as well as the stride and padding, can be adjusted to achieve desired effects. Convolution layers excel at local pattern detection and can learn spatial hierarchies.

Activation functions introduce nonlinearity into the output of a neuron. Without non-linear activation functions, no matter how many layers the neural network possesses, it would behave just like a single-layer perceptron, as summing these layers would give another linear function. Common activation functions include the Rectified Linear Unit (ReLU), sigmoid, and Hyperbolic Tangent (tanh). They transform the inputs into outputs that are then channeled to the next layer.

The pooling layer primarily serves to reduce the spatial dimensions of the input, which leads to both computational efficiency and a reduction in the chances of overfitting. It operates on each feature map separately and reduces their size by using various operations, the most common of which is max pooling. Here, the maximum value is taken from a group of values, typically from a 2×2 window, and represent the entire group, thereby achieve the downsampling.

Situated typically at the end of deep learning architectures, the fully connected layer is where neurons in a layer are connected to every neuron in the preceding layer that ensures global understanding from earlier layers (like convolution or pooling layers). This dense layer takes the high-level features learned by previous layers and transforms them

to produce the final output, often leading to classification or regression results.

Convolutional Neural Networks (CNN) are a class of deep learning models tailored for analysing visual data. The CNN architecture is distinctively characterized via its convolutional layers that automatically and adaptively learn spatial hierarchies of features from input images. This is followed by pooling layers which reduce the spatial dimensions while retaining important information. The deep structure of a CNN allows it to learn complex patterns and features, which are then flattened and passed through one or more fully connected layers to make a final prediction. Due to its specialized architecture, CNNs have emerged as the go-to model for various image recognition tasks.

The reason why CNNs and similar models excel in FER lies in their ability to capture subtle and complex facial features, nuances that are often missed by other methods . As facial expressions are a combination of various muscle movements, which are subtle, deep learning models' ability to learn from vast amounts of data and capture minute details makes them highly effective.

Furthermore, the availability of large labelled datasets in recent years has significantly propelled the advancement of machine-learning-based FER. Datasets like FER2013, AffectNet, and EmoReact provide thousands to millions of annotated facial images, enabling models to be trained more comprehensively and rigorously.

Practical applications of FER are vast and varied. In human-computer interaction, deep learning systems can be designed to adapt and respond based on the user's emotional state, creating a more intuitive and empathetic user experience. In healthcare, it can be used for monitoring patients for signs of pain or distress, especially if they cannot communicate verbally. In automotive industry, FER can be used to monitor driver's emotions and alertness, potentially preventing accidents caused by drowsiness or distress. Additionally, the entertainment and advertising industries can take use of FER to gauge audience reactions to content in real time, allowing for dynamic adjustments.

While deep learning has made significant strides in FER, there are still concerns related to bias, especially when models are trained on datasets that lack diversity. Differences in lighting, orientation, and occlusions can also affect performance. Moreover, genuine emotions can sometimes be suppressed or faked, making it hard for algorithms

to discern.

While FER in machine learning has made remarkable progress and offers promising applications across various sectors, continous research work is essential to refine models, which improves accuracy, ensures fair and unbiased emotion recognition across diverse populations. The intersection of psychology, computer science, and data promises further advancements in the near future, with potential benefits transcending technological realms into societal and humanitarian domains.

### 1.1.2 The Importance of Lightweight Models

In the rapidly evolving landscape of deep learning and artificial intelligence, the emphasis on creating sophisticated models that can perform complex tasks with high accuracy has been undeniably prominent. While the pursuit of accuracy is unquestionably valuable, it often comes at the cost of increased model size and complexity. This gives rise to the need for lightweight models, especially when real-world applications are considered. (Verma et al., 2023)

One of the primary benefits of lightweight models is their speed and efficiency. In real-time applications, such as facial emotion recognition on mobile devices or edge devices, a delay in processing can be problematic. Whether it's a real-time health monitor detecting distress in a patient or an interactive virtual assistant responding to user emotions, rapid processing is crucial, and lightweight models deliver on this front.

As models grow in complexity and size, the computational power and memory required to run them also increase. This not only makes them less accessible to those without high-end hardware but also increases the energy consumption, making operations less sustainable. Lightweight models, with their trimmed architectures, can run on devices with limited resources, ensuring wider accessibility and reduced operational costs.

While integrating a machine learning model into applications or systems can be challenging, larger models might necessitate additional infrastructure considerations, such as specialized hardware or cloud-based solutions. In contrast, lightweight models, due to their compact nature, can be more seamlessly integrated into existing platforms

without significant modifications.

At the same time, with the rise of Internet of Things (IoT) and edge computing, processing data at the source like a smart camera, a wearable device, or an industrial sensor has become increasingly vital. Lightweight models are inherently suited for edge devices as they can operate within the constraints of these devices, enabling faster decision-making without the need to communicate with a central server.

For applications that need to be scaled, especially in cloud environments, deploying lightweight models means that less server capacity is used per instance. This leads to cost savings and ensures that systems can handle more concurrent users or tasks.

While the push for ever-more-accurate models in research and certain high-resource settings will continue, the importance of lightweight models cannot be understated. In an era where democratization of technology, energy efficiency, and real-time processing are paramount, lightweight models stand out as an essential tool in the machine learning toolkit, bridging the gap between high-end research and practical, real-world applications.

## 1.2 Research Questions

In order to improve the efficiency and performance of Facial Emotion Recognition (FER) through the use of an ensemble of lightweight mini-Xception models, our research questions naturally emerge. These questions aim to guide the investigation, allowing us to assess not only the effectiveness of our approach but also its broader implications and potential shortcomings. The follows are the primary research questions formulated based on the innovative algorithmic approach under consideration:

(1) *How Does the Ensemble of Lightweight Models Impact FER Accuracy Compared to a Singular Model Approach?*

(2) *How Does Confidence-Based Voting in an Ensemble Setup Influence the Robustness and Decision-making of Model?*

(3) *How Effective is Over-sampling of Positive Samples in Enhancing the Training Efficiency and Addressing Data Imbalance in FER?*

This question seeks to explore the fundamental advantage of our ensemble approach. Traditional FER models focus on using one comprehensive model to recognize a spectrum of emotions. While effective in many scenarios, there can be limitations, especially when nuanced or subtle expressions come into play. By adopting an ensemble of mini-Xception models, each expertly tuned to recognize a specific emotion against all others, we hypothesize that the overall accuracy can be enhanced. This question will assess the collective efficacy of our ensemble in comparison to singular, monolithic models in recognizing a wide range of facial emotions.

Classic ensemble learning methods often employ a majority voting system, where each model in the ensemble "votes" for an outcome, the majority decision is chosen. Our innovative approach shifts from this paradigm, incorporating a confidence-based voting mechanism. Instead of binary decisions, each model in our ensemble learning provides a confidence score for its decision. The research question aims to understand how this nuanced voting approach impacts the ability of the proposed model to make decisions, especially in ambiguous or borderline cases. Furthermore, it seeks to determine if this method provides an added layer of robustness against misclassifications.

Data imbalance is a pervasive issue in deep learning, including FER. Traditionally, methods like under-sampling of negative or majority classes have been used. However, our approach leans towards oversampling the positive samples. This research question delves into the repercussions of this strategy. Specifically, it seeks to assess whether oversampling leads to better generalization in real-world scenarios and if it aids in creating a more balanced representation during training without leading to overfitting.

By addressing these research questions, this study aims to shed light on the potential and challenges of our innovative approach to FER, offering insights that could guide future advancements in the domain.

## 1.3 Contributions

The field of Facial Emotion Recognition (FER) has witnessed substantial advancements over the years, with various models and methods being introduced to enhance its accuracy

and efficiency. This thesis, in its pursuit to further optimize FER, brings forth several key contributions to the existing body of knowledge. The contributions, both theoretical and practical, have the potential to shape future research and real-world applications in this domain.

(1) Innovative Ensemble Approach Using Lightweight Models:

At the core of our research is the development of an ensemble of lightweight mini-Xception models. Rather than relying on a singular, comprehensive model, our ensemble approach uses multiple experts, each fine-tuned to detect a specific emotion against all others. This not only leverages the strengths of each individual model but also combines their expertise for more accurate and robust emotion detection. This innovative approach, which emphasizes modularity and specialization, paves the way for flexible FER systems that can be tailored to specific applications or contexts.

(2) Confidence-Based Voting Mechanism:

Building upon the ensemble approach, this research introduces a novel confidence-based voting mechanism. Traditional ensemble learning methods often opt for majority-based decisions. In contrast, our proposed method allows each model in the ensemble to provide a confidence score for its decision. This refined voting strategy takes into account of the varying strengths and uncertainties of each model, leading to more informed and nuanced decision-making. This mechanism has the potential to set a new standard in ensemble methods, moving beyond binary outcomes to richer, more contextual results.

(3) Oversampling Approach to Address Data Imbalance:

As part of this research project, we are also facing the challenge of counteracting data imbalance, which is a challenge in FER datasets. Instead of the often-used under-sampling of negative or majority classes, our research promotes over-sampling of positive samples. This approach, while simple in its essence, presents a paradigm shift in how data augmentation can be used to enhance model training, potentially reducing biases and improving generalization in real-world scenarios.

(4) Comprehensive Evaluation and Benchmarking:

Beyond the introduction of novel methods, our research work provides a thorough evaluation of the proposed methods, benchmarking them against existing state-of-the-art

models.

In this research project, through its novel methodologies and comprehensive evaluations, we offer fresh perspectives and tools to the FER community. This not only push the boundaries of accuracy and efficiency in emotion recognition but also present scalable and adaptable solutions that can cater to the diverse needs of the digital world.

## 1.4 Objectives of This Thesis

This thesis shows an endeavor to present the advancements made in the realm of Facial Emotion Recognition (FER) through the innovative incorporation of an ensemble of lightweight mini-Xception models. As we delve into the nuances of this research, it's essential to outline the primary objectives that have guided the structure, content, and direction of this thesis:

(1) To Present a Comprehensive Overview of the FER Landscape:

The ever-evolving domain of FER has witnessed various transformations, with methods evolving from basic classifiers to sophisticated deep learning models. One of the core objectives of this thesis is to provide people with a thorough understanding of the current state-of-the-art in FER, highlighting the existing challenges, methodologies, and the imminent need for more efficient solutions.

(2) To Introduce the Ensemble Approach with Lightweight Models:

Central to our research is the novel approach of utilizing an ensemble of specialized mini-Xception models for FER. Through this thesis, we aim to elucidate the rationale behind adopting an ensemble framework, the benefits it offers in terms of modularity and flexibility, and the unique advantages of using lightweight models for real-time and resource-constrained applications.

(3) To Detail the Confidence-Based Voting Mechanism:

Moving beyond the conventional majority-based ensemble decisions, this thesis seeks to shed light on our innovative confidence-based voting strategy. Our objective is to provide insights into how this mechanism operates, the theoretical foundations supporting its efficacy, and its potential implications for enhancing decision robustness in

ambiguous scenarios.

(4) To Highlight the Over-sampling Technique for Addressing Data Imbalance:

Addressing the prevalent challenge of data imbalance in FER datasets, this thesis presents our unique approach of over-sampling positive samples. We aim to outline the advantages of this technique, how it contrasts with traditional methods, and its potential in ensuring model generalization.

(5) To Provide a Rigorous Evaluation and Benchmarking:

Ensuring that our proposed methodologies are not just theoretically sound but also practically effective is vital. Hence, one of our objectives is to present a comprehensive evaluation of our methods, benchmarking them against existing models, and analyzing their performance across varied datasets and scenarios.

(6) To Offer Future Directions and Potential Enhancements:

While this thesis introduces novel methods and findings, it's equally important to recognize its limitations and the potential avenues for future work. Through this thesis, we aim to provide a forward-looking perspective, suggesting possible extensions, improvements, and areas of exploration for the wider research community.

The objectives of this thesis revolve around elucidating our novel methodologies, through validating their efficacy, and positioning our research within the broader context of FER advancements, while aiming to offer both depth and breadth to people interested in the fusion of emotion recognition and machine learning innovations.

## 1.5 Structure of This Thesis

This thesis is methodically structured to provide a seamless and comprehensive insight into the advancements made in the realm of Facial Emotion Recognition (FER) through the innovative use of an ensemble of lightweight mini Xception models. To ensure clarity, the layout follows a logical progression, from the introduction and foundational concepts to a detailed exposition of the methodologies and their evaluations. The overview of this thesis:

**Chapter 1: Introduction.** The first chapter sets the stage for the entire thesis. It begins by introducing the context and motivation behind the research, emphasizing the relevance of FER in machine learning and the significance of lightweight models. This chapter also outlines the primary research questions driving the investigation, elucidates the main contributions, and clarifies the objectives guiding this documentation.

**Chapter 2: Literature Review.** Serving as a backdrop, the second chapter delves into a comprehensive review of existing literature which encompasses traditional machine learning methods for FER, the emergence and evolution of deep learning models, the specifics of the mini-Xception model, ensemble learning approaches with a focus on voting mechanisms, and methods to address data imbalances. This chapter ensures people have a solid understanding of where the current research fits within the broader academic landscape.

**Chapter 3: Methodology**. The heart of the thesis, this chapter meticulously details the research methodologies employed. It presents a deep dive into ensemble learning with mini-Xception model, the intricacies of creating a binary classification model, the innovative confidence-based voting scheme, data augmentation methods, and the overarching process of data preparation and program implementation. Additionally, it expounds on the evaluation methods used to assess the performance of model.

**Chapter 4: Results**. The research outcomes of the implemented methodologies are presented in this chapter. It starts with a description of the data collection process and the experimental setup. The chapter then progresses to a detailed analysis of the ensemble performance of model, both at the individual experts' level and collectively. It also explores the performance of the data-augmented mini-Xception model and sheds light on the limitations of the study.

**Chapter 5: Analysis and Discussions**. In this chapter, the results are dissected and contextualized. The advantages of the ensemble approach and the benefits of data augmentation are thoroughly analyzed. Moreover, the findings are discussed in relation to real-world applications, and a comparison is made with other state-of-the-art models, offering a holistic view of the implications of research and its positioning in the wider FER field.

**Chapter 6: Conclusion and Future Work**. The concluding chapter encapsulates the research journey, summarizing the key takeaways and findings. It also looks ahead, suggesting potential extensions to the proposed model, areas of further exploration, and other ensemble methods that might enhance FER.

# Chapter 2 Literature Review

*Deep learning has rapid progress, particularly at the juncture of human cognition and machine perception. Among these, Facial Emotion Recognition (FER) stands out. Historically, human faces have been an interface for non-verbal communication, often revealing emotions more precisely than words. Decoding these cues is vital for enhancing human connection and empathy. Transferring this ability to machines has been complex yet fruitful, transitioning from rudimentary algorithms to advanced deep learning models. This chapter delves into FER history, through emphasizing on its evolution, challenges, and innovative solutions, such as the mini-Xception model and ensemble learning approaches.*

## 2.1 Introduction

The rapid advancements in deep learning have engendered a rich tapestry of methodologies, especially in areas that intersect human cognition and machine perception. Among these intersections, Facial Emotion Recognition (FER) holds a special place. It not only promises a more intuitive human-machine interface but also harbors potential applications that span from healthcare diagnostics to refining user experiences in technology products. As we embark on this journey through the literature, we must firstly understand the context and significance of FER in the broader AI landscape.

Historically, human faces have been a profound source of non-verbal communication, which often reveal emotions and intents more accurately than words can express. Decoding these subtle facial cues, therefore, is crucial for enhancing human interaction, empathy, and understanding. Translating this capability to machines, however, has been a challenging yet rewarding pursuit. The idea of enabling machines to discern human emotions from facial expressions not only augments their usability but also has transformative implications across various sectors.

The FER within the confines of computer science started with rudimentary algorithms capable of identifying basic facial landmarks. Over time, with the proliferation of more sophisticated algorithms and the availability of vast amounts of data, FER evolved from simple classifiers to intricate deep learning models, achieving remarkable accuracy rates. However, the road to this point was paved with myriad challenges: From handling data imbalances and subtle emotional differences to ensure real-time processing and develop lightweight models suitable for a wide range of devices.

We start from understanding traditional machine learning methods that served as the bedrock for initial FER systems. This foundation then sets the stage for the evolution into deep learning models, which have, in recent years, taken the FER world by storm due to their unparalleled performance. Within this progression, the mini-Xception model emerges as a notable contributor, especially for its lightweight architecture that doesn't compromise on accuracy.

Furthermore, the review will also touch upon ensemble learning approaches,

specifically focusing on their applications in deep learning for FER. The nuances of different voting mechanisms, their strengths, and potential pitfalls will be dissected, offering people a panoramic view of ensemble methods' efficacy in FER scenarios.

Lastly, addressing one of the perennial challenges in FER, the issue of data imbalance, we will explore the evolution of strategies from under-sampling to innovative over-sampling methods. The implications of these methods on model training, bias mitigation, and real-world performance will be discussed in detail.

This literature review aims to provide a holistic understanding of the FER landscape, capturing its evolution, current methodologies, challenges, and the innovations poised to shape its future.

## 2.2 Facial Emotion Recognition

### 2.2.1 Machine Learning

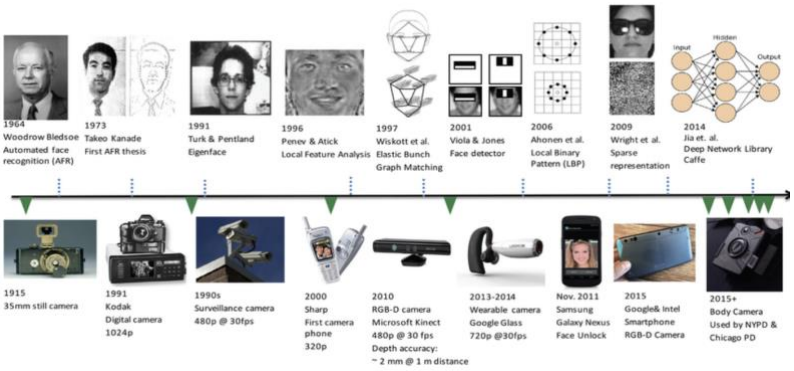


Figure 2.1: The major milestones in the history of automated face recognition

Over the years, face recognition technology has experienced significant advancements, both in terms of recognition algorithms and acquisition systems.

From Figure 2.1, the upper row depicts the progression of face recognition algorithms, charting the transformative shifts in computational methods. From the early stages of template matching to the modern age of deep neural networks, the journey reflects the continual strive for accuracy, adaptability, and real-time processing. Artificial neural networks, especially Convolutional Neural Networks (CNN), have played a pivotal

role in recent advancements, boasting state-of-the-art accuracies and robustness against diverse and dynamic environments.

Concurrently, the bottom row showcases the evolution of human face acquisition systems. Spanning from rudimentary camera-based setups in the early days to sophisticated 3D infrared and depth-sensing cameras in contemporary times, the timeline signifies the advancements in hardware technology. These cutting-edge systems today can capture minute facial details, operate in various lighting conditions, and even detect live presence to counter spoofing attempts.

Before the profound ascent of deep learning approach, traditional machine learning methods were the mainstay of Facial Emotion Recognition (FER). These algorithms are lacking the depth and adaptability of neural networks, paved the way for emotion recognition from facial data. Let's traverse through the evolution of these methods and understand their significance, strengths, and limitations.

The cornerstone of traditional machine learning methods in FER has been feature extraction. Unlike deep learning, which can automatically learn and refine features, traditional methods rely on hand-crafted features. These features, such as Haar-like features, Local Binary Patterns (LBP), Gabor filters, and Histogram of Oriented Gradients (HOG), were meticulously designed to capture essential facial details. (Kanna et al., 2022) For instance, LBP was adept at capturing texture information, which is crucial for discerning subtle facial expressions.

Once features were extracted, they were fed into machine learning classifiers. Some of the most prevalent ones include:

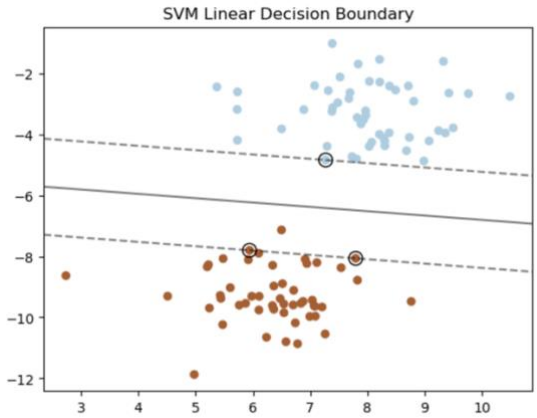


Figure 2.2: SVM samples

Support Vector Machines (SVM): Renowned for their ability to manage high-dimensional data and produce robust decision boundaries, SVMs were widely adopted in FER. By transforming the feature space using kernel tricks, SVMs could handle non-linearly separable data with relative ease. As shown in Figure 2.2, we can see a SVM sample, including the data points with their classifications, the decision boundaries, and the support vectors (these vectors determine the location of the decision boundaries). These points are located near the decision boundary and are critical to determining the location of the boundary. Support vectors are highlighted in particular.

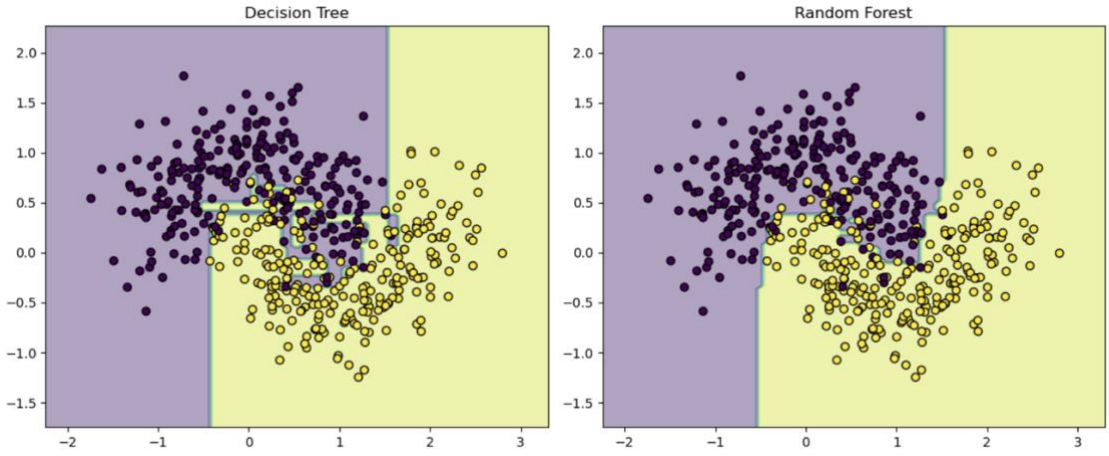


Figure 2.3: Samples of decision trees and random forests

Decision Trees and Random Forests as shown in Figure 2.3. These were favored for their interpretability and ability to handle non-linear relationships. Random Forests, an ensemble of Decision Trees, provided a boost in performance by reducing variance and preventing overfitting. In the figure, we can observe that the decision tree provides a

detailed fit to the data, while the random forest provides a smoother decision boundary. In this sample, random forests are shown to be effective at improving overall performance and generalization by combining the predictions of multiple decision trees.

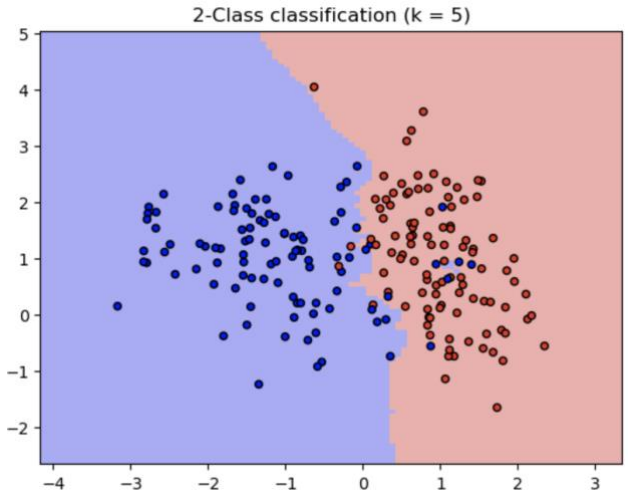


Figure 2.4: K-nearest neighbors samples

K-Nearest Neighbors (KNN) was effective for datasets where facial expressions formed clear clusters. Figure 2.4 shows two colored areas representing the decision boundaries for the two categories. In addition, data points are shown in color, indicating their actual category. Different values of k (number of neighbors) may result in different shapes for the decision boundary.

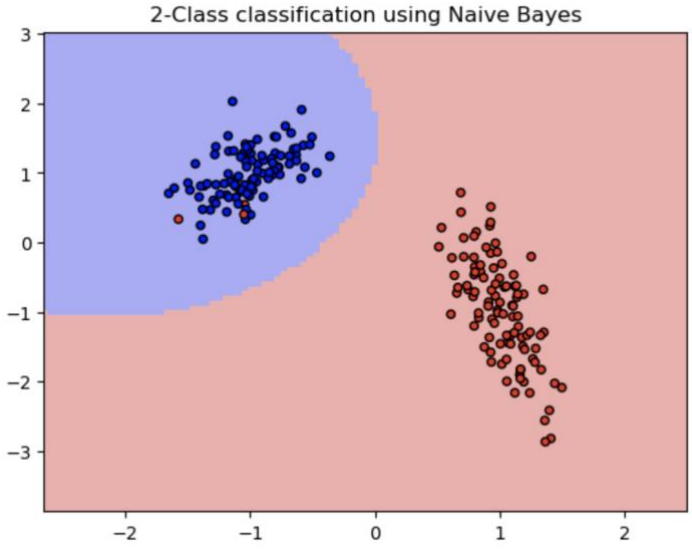


Figure 2.5: Naive Bayes samples

Leveraging the principles of Bayes' theorem, Naive Bayes, this probabilistic classifier was especially useful when the assumption of feature independence held reasonably true. There are two colored areas in Figure 2.5 which represent the decision boundaries for the two categories. Colors are also used to indicate the actual category of the data points. Since the Naive Bayes classifier is a probabilistic classifier, it calculates the probability of belonging to each class based on Bayes' theorem and selects the class with the highest probability for prediction.

Handling high-dimensional facial data was a challenge like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA aimed to reduce dimensionality while retaining the maximum variance in the data, whereas LDA sought to find a linear combination of features that best separated different emotional classes.

While these traditional methods achieved reasonable accuracy, they were not without challenges. A significant amount of expertise was required to design and select appropriate features. The hand-crafted nature meant that as facial data grew more complex, these features might not capture all the nuances necessary for accurate emotion recognition.

Traditional methods struggled with very large datasets. As the volume of facial data expanded, the computational cost of training models, especially those like SVM, grew substantially.

Unlike deep learning models that can learn hierarchical representations, traditional methods had a fixed feature extraction and classification approach. This rigidity often led to performance plateaus, beyond which improvements were challenging.

Despite their limitations, traditional machine learning methods provided the foundational steps towards automated FER. They fostered initial insights into which facial features were most indicative of emotions and laid the groundwork for future, more advanced methodologies. These methods serve as a testament to the initial strides of the AI community in bridging the gap between human emotional expressions and machine comprehension, setting the stage for the deep learning revolution that followed.

### 2.2.2 Deep Learning

As we entered the era of Big Data, the limitations of traditional machine learning methods in facial emotion recognition became more apparent. The need for automatic feature extraction and the capacity to handle vast amounts of intricate facial data paved the way for the rise of deep learning models. These models, especially neural networks, ushered in a new age of unprecedented accuracy and adaptability in FER.

In 2006, Hinton and his colleagues introduced the groundbreaking theory of deep learning and subsequently applied it innovatively to image processing. Deep learning, fundamentally rooted in the deep neural network, is a specialized subset of artificial neural networks. The foundation of deep learning is established upon the research in artificial neural networks. By adjusting the number of hidden layers, one can derive an artificial neural network model with multiple hidden layers. Hidden neural networks of this type are able to learn more effectively, mirroring the cognitive processes of the human brain. This facilitates the efficient extraction of image features. Notably, the neural network with multiple hidden layers is essentially a deep learning model. The transition from a single hidden layer neural network model to a deep learning model is evident through a visual comparison of the two representations. (Feng et al., 2020)

Among deep learning architectures, CNNs became the poster child for FER. CNNs are specifically designed to process grid-like data, making them ideal for images. They consist of convolutional layers that can automatically and adaptively learn spatial hierarchies of features from input images. This property alleviated the need for hand-crafted features, a limitation of traditional methods.

Layers within CNNs, such as pooling layers, helped in reducing spatial dimensions while retaining crucial information. Activation functions introduced non-linearity, enabling the network to capture complex relationships. Furthermore, fully connected layers at the end of these networks made the final decision about the emotion depicted in the input image.

While CNNs excel at spatial feature extraction, RNNs and their advanced variant, LSTM, shone in scenarios where temporal sequences mattered, such as in video-based

FER. LSTMs, with their memory cells, effectively captured temporal dependencies and addressed the vanishing gradient problem of traditional RNNs.

Given the extensive computational resources required to train deep neural networks from scratch, transfer learning emerged as a vital strategy in FER. Pre-trained models, like VGG-16 or ResNet, initially trained on large datasets like ImageNet, were fine-tuned for emotion recognition tasks. This approach leveraged the generic features learned by these models and adapted them to the specifics of FER, leading to faster convergence and improved accuracy.

Deep learning models, due to their high complexity, were prone to overfitting. methods like dropout, where random neurons are "dropped" during training, or batch normalization, which normalizes the activations of a layer, were employed to ensure model generalization. Additionally, advanced optimization methods like Adam and RMSprop were used to speed up convergence and improve training stability.

Deep learning, while transformative, also introduced new challenges. Deep neural networks, especially with many layers, required substantial computational resources. This sometimes restricted their deployment on edge devices with limited processing capabilities. For deep learning models to excel, they required large, labeled datasets. While transfer learning mitigated this to some extent, the need for vast amounts of data was undeniable. Deep models, often termed as “black boxes”, lacked the clear interpretability that some traditional methods offered.

Particle Swarm Optimization (PSO) algorithm from the ensemble learning is adopted as a fitness value. Experimental outcomes indicate that this proposed model consistently outperforms other methods across diverse datasets. Additionally, the study delves into hyperparameter optimization, utilizing the GridSearch CV method, aiming to discern the optimal hyperparameters for both regularization and optimization. This holistic strategy not only bolsters accuracy but also mitigates the potential errors of individual classifiers, ensuring a more reliable and consensus-based prediction. (Alrefai et al., 2022)

At the same time, the use of ensemble learning and majority voting methods to select the best accuracy among all classifiers. This accuracy is then used as a fitness value for

the Particle Swarm Optimization (PSO) algorithm. The proposed model shows superior accuracy results compared to other methods when implemented on different datasets. Additionally, the article discusses the use of hyperparameter tuning with the GridSearch to find the optimum hyperparameters for regularization and optimization.

The adoption of deep learning in FER represented a paradigm shift. With their capacity for automatic feature extraction and hierarchical learning, deep models surpassed traditional methods in accuracy and robustness across varied and complex facial datasets. Their rise solidified the promise of FER in real-world applications, from real-time surveillance systems to interactive entertainment and beyond. The FER landscape, fueled by the power of deep learning, is now poised to redefine human-machine interactions in unprecedented ways.

## 2.3 The mini-Xception Model

Within the panorama of deep learning models for Facial Emotion Recognition (FER), the mini Xception model emerges as a blend of sophistication and efficiency. It stands out not just for its performance but also its relatively compact nature, catering to a niche where both accuracy and model size are paramount.

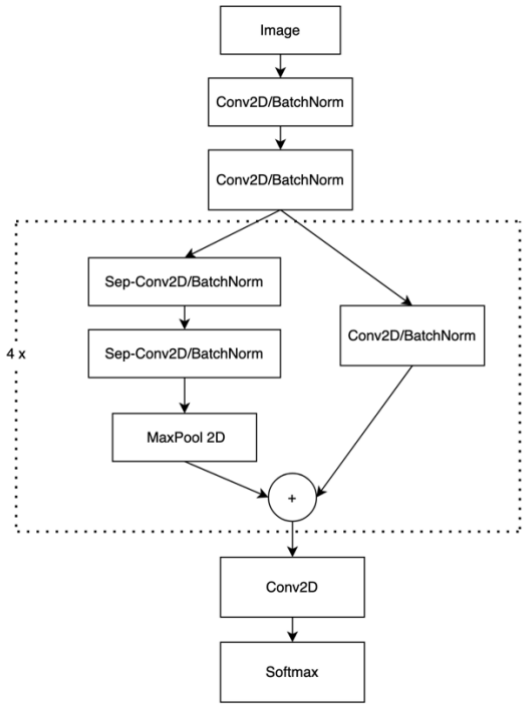


Figure 2.6: mini-Xception architecture

From Figure 2.6, we see mini-Xception introducing Depthwise Separable Convolution: Replacing the traditional convolutional layer in a conventional CNN with depthwise separable convolution was the initial transformative step. This adaptation has a two-fold advantage: It substantially reduces the number of parameters involved, leading to quicker computations and lower resource utilization, while maintaining, if not improving, the representational efficiency of the model.

To further compact the model, we eliminated the fully connected layer traditionally found in CNNs. This design choice not only reduced the computational overhead but also minimized potential overfitting, a frequent concern associated with fully connected layers due to their high parameter count.

The result of mini-Xception model comprises of four depthwise separable convolution blocks. The batch normalization processes the output to stabilize and accelerate the training process. This is complemented by the introduction of the ReLU activation function, which infuses the model with the necessary non-linearity. In the culmination of the forward pass, the SoftMax function is invoked to facilitate multi-class classification of the results.

The streamlined architecture of mini-Xception model ensures efficient training, while requiring considerably less time compared to its predecessor models. Moreover, preliminary experiments attest to its improved generalization capabilities, making it a robust choice for diverse applications.

The mini-Xception model draws inspiration from the original “Xception” architecture, which stands for "Extreme Inception" (Li et al., 2022). In the Keras deep learning library, Xception was designed to improve upon the Inception architecture by using depthwise separable convolutions. These convolutions provide the dual benefit of reducing model parameters while retaining, if not improving, performance.

At the heart of mini-Xception model lies depthwise separable convolution, a unique convolution technique that breaks down the traditional convolution operation into two steps: Depthwise convolution followed by pointwise convolution. This approach significantly reduces the number of trainable parameters in the model and ensures a lightweight design without compromising on the feature extraction capabilities.

While the mini-Xception model borrows depthwise separable convolution from its parent Xception model, it has been simplified and optimized for smaller-scale applications, particularly in FER. The architecture consists of fewer layers but retains the essence of depthwise separable convolutions. This reduced complexity ensures faster inference times, making it suitable for real-time applications.

Therefore, it has advantages, the use of depthwise separable convolutions means that mini Xception has fewer parameters than standard CNN architectures, ensuring quicker training and inference times. Its lightweight nature makes it deployable on edge devices with limited computational power, such as mobile devices or embedded systems. Despite of its reduced size, mini-Xception model has demonstrated competitive, if not superior, performance in FER tasks when compared to bulkier deep learning models. Given the unique challenges posed by FER, such as subtle emotion variations and diverse datasets, the balanced design of mini-Xception model proves invaluable. Its ability to extract crucial features from facial data while maintaining a nimble profile makes it a preferred choice for applications where real-time feedback and device constraints are paramount.

The mini-Xception model, with its foundational depthwise separable convolutions, represents a harmonious blend of deep learning efficacy and computational efficiency. In the realm of FER, where the nuances of human emotions intersect with the demand for rapid, efficient computational processes, mini-Xception stands as a testament to innovative model design and optimization.

## 2.4 Resemble Learning

### 2.4.1 Voting Mechanisms

In the vast landscape of ensemble learning, the idea of leveraging multiple models to make a collective decision is central. One of the most intuitive and widely employed methods to achieve this consensus is through voting mechanisms. Voting, in the context of machine learning, allows for an aggregation of predictions from several models to produce an outcome, leveraging the wisdom of the 'crowd' to enhance accuracy and robustness.

In the realm of neural networks, unweighted averaging stands out as the predominant ensemble technique. This method straightforwardly calculates the mean of the output scores or probabilities from all base models, presenting this average as the final prediction. (Ju et al., 2018)

The intrinsic capability of deep neural networks to capture intricate patterns means that even a simple procedure like unweighted averaging can significantly enhance performance. By averaging across multiple networks, one can effectively reduce the model variance. This is especially impactful given that deep artificial neural networks (ANNs) are characterized by high variance but low bias. If the underlying models are sufficiently diverse or uncorrelated, their collective variance can be markedly diminished when averaged. This principle underpins the design of Random Forests, which promotes less correlation between trees by leveraging bootstrapped samples and feature selection.

In hard voting, each model in the ensemble “votes” for a specific class. The class that receives the majority of votes is chosen as the final prediction. It's straightforward and doesn't require probability estimates. But All models have equality, regardless of their accuracy or confidence in prediction.

Soft voting unlike hard voting, soft voting takes use of the probability estimates of each class predicted by individual models. The class with the highest average probability across all models is selected as the final prediction. It takes into account the confidence of each model in its predictions. The models are more certain about their predictions having a more significant influence on the final decision. It requires to provide probability estimates, which isn’t always feasible.

The advantages of voting mechanisms is by aggregating predictions, the ensemble smoothens out the biases and variances of individual models, often leading to a model that's less prone to overfitting.

Even if individual models have comparable performance, their collective wisdom can often lead to improved accuracy, especially if the errors are uncorrelated and ensembles are less sensitive to the quirks of individual models. If one model makes an erroneous prediction, it's likely to be corrected by the majority. For voting to be effective, it's crucial that the models in the ensemble are diverse. If all models make the same errors,

the ensemble will merely amplify them. The diversity can come from using different algorithms, training on varied data subsets, or even introducing randomness during model training.In some cases, it might make sense to assign different weights to different models based on their performance. The models that are more accurate or reliable can be given more “voting power”, including a greater influence on the final decision of the ensemble. At the same time, there's a balance to strike regarding the number of models in the ensemble. While more models can provide a more robust voting mechanism, there are diminishing returns beyond a certain point, not to mention increased computational costs.

Currently voting mechanisms aren't limited to any particular kind of model or problem domain. They've been employed in various fields, from computer vision to natural language processing. In FER, voting can be particularly useful when dealing with subtle emotions, where slight variations in facial features can lead to different predictions. It is often the collective judgment of an ensemble that proves more accurate in scenarios of this magnitude.

Voting mechanisms epitomize the adage “The whole is greater than the sum of its parts.” By harnessing the collective insights of multiple models, voting not only amplifies the strengths of individual models but also mitigates their weaknesses. As challenges in machine learning grow more intricate and datasets more complex, voting mechanisms and ensemble methods at large will continue to play a pivotal role in pushing the boundaries of what's achievable.

### 2.4.2 Confidence-Based Voting

In the context of ensemble methods, while traditional voting mechanisms like hard and soft voting are effective, they can be further refined using a more nuanced approach: confidence-based voting. Confidence-based voting utilizes the confidence or certainty of predictions from each model to weigh their contribution to the final ensemble decision.

Confidence-based voting is an extension of soft voting. Instead of just considering the raw probability scores provided by each model, this method takes into account how confident a model is in its prediction. The confidence score can be computed either

directly based on the probability estimate or using an external metric to assess the quality of the model.

Unlike traditional weighted voting where weights are static and pre-assigned, confidence-based voting adjusts weights on-the-fly based on the confidence of model for each specific prediction. This allows the ensemble to be more adaptive to varying situations.

By emphasizing predictions made with high confidence, this method diminishes the impact of predictions made with low certainty, allowing the ensemble decision not to be unduly influenced by "wild guesses" by any model.

As predictions backed by higher confidence are often more accurate, giving them more weight can boost the overall performance of the ensemble.

It's essential to have a reliable metric to measure confidence. While probability estimates can serve this purpose, in some scenarios, additional measures such as model calibration or external validation might be necessary.

A threshold might be set, below which predictions are deemed too uncertain to be considered. This can filter out extremely uncertain predictions, adding a layer of robustness.

While confidence is valuable, it's crucial to ensure that models are not overly confident in their predictions, leading to overfitting or biases. Regularization methods or calibration methods can help in maintaining a balance.

Facial Emotion Recognition (FER) often grapples with subtle and nuanced distinctions between emotions. In such cases, having an ensemble of models that not only predicts but also quantifies confidence can be invaluable. If one model is highly confident that a facial expression is, for instance, “sadness” while another is only marginally confident it's “anger”, confidence-based voting would give more weight to the former, potentially increasing the accuracy of the decision made by the ensemble.

Confidence-based voting adds a layer of sophistication to ensemble methods, allowing them to be more adaptive and discerning. By considering not just what the models predict, but also how certain they are in their predictions, this approach harnesses a deeper level of insight from each model. Especially in tasks like FER, where nuances

can be the difference between accurate and erroneous classification, confidence-based voting emerges as a potent tool in the machine learning ensemble toolkit.

### 2.4.3 Voting Mechanisms in Ensemble Learning

Ensemble learning, which combines the predictions of multiple machine learning algorithms to produce a single, cohesive output, utilizes different voting schemes to aggregate results. Two of the most employed schemes are hard voting and soft voting.

Table 2.1: Voting samples

| Classifiers | Spam (y=0) | Not Spam (y=1) | Result |
|---|---|---|---|
| Classifier 1 | 56% | 44% | Spam |
| Classlfier 2 | 49% | 51% | Not Spam |
| Classifier 3 | 60% | 40% | Spam |
| Classifier 4 | 45% | 55% | Not Spam |
| Classfier 5 | 39% | 61% | Not Spam |

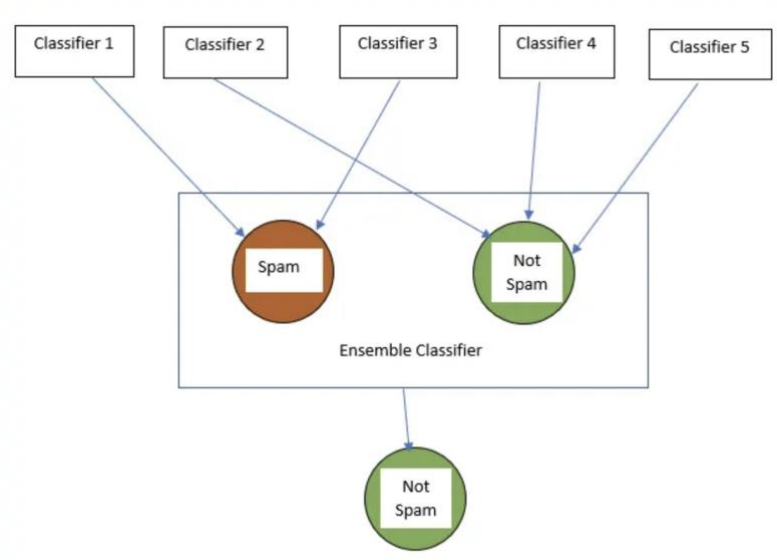


Figure 2.7: Voting example

#### 2.4.3.1 Hard Voting

Hard voting often refers to as majority voting, consists of aggregating the predictions from each base model and designating the class with the highest vote count as the final

outcome. This method is especially fitting for classification problems where the potential outcomes are discrete, non-overlapping categories.

Considering a scenario as shown in. Figure 2.7 where the objective in Table 2.1 is to discern if an email is spam. This is essentially a binary classification dilemma, with the two possible categories being “spam” (the negative class) and “not spam” (the positive class). Based on the sample, three of these classifiers indicate that the email is not spam, while the remaining two indicate that it is spam. A hard voting system would be used to determine the majority's opinion. By using the hard voting system, the final verdict would be that the email is benign, or “not spam”.

Hard voting is straightforward and requires no probability calibration. It works effectively when all base models are equally reliable. But hard voting can be limited in scenarios where the models are only slightly confident in the predictions than others. In such cases, the nuances of individual model confidence are lost.

#### 2.4.3.2 Soft Voting

Soft voting, alternatively termed weighted voting, leverages the probability scores assigned by each base model for every class, computing their weighted average to determine the final prediction. Instead of merely tallying votes like in hard voting, soft voting derives the mean probability of every class, ultimately choosing the class that boasts the highest cumulative probability. This methodology is apt for both classification and regression challenges. To elucidate, let's revisit the earlier email classification scenario.

In a soft voting framework, the final classification hinges on the mean probability scores ascribed to each class across all utilized models.

The exemplified use of soft voting accentuates its capacity to harness the nuanced probability estimations from every individual model, allowing a more informed decision rooted in the aggregate expertise of all contributing models. By considering each confidence of model quotient, soft voting can often yield superior precision and efficiency in diverse practical applications.

While hard voting and soft voting both provide mechanisms to aggregate predictions

in ensemble models, the choice between them hinges on the specific problem at hand and the nature of the base models.

Table 2.2: Comparative analysis of ensemble learning

| Technique | Base Models Diversity | Aggregation Method | Layering | Primary Advantage |
|---|---|---|---|---|
| Voting Ensembles | High | Hard/Soft Voting | No | Capitalizes on strengths of diverse models |
| Bagging | Low (Same Algorithm) | Averaging for Regression/ Voting for Classification | No | Reduces variance |
| Boosting | Low (Same Algorithm) | Weighted Aggregation | No | Reduces bias, prioritizes misclassified instances |
| Stacking | High | Meta-model | Yes | Captures complex inter-model relationships |
| Blending | High | Meta-model on hold-out set | Yes | Reduces risk of overfitting compared to stacking |

The landscape of ensemble methods in machine learning is both diverse and nuanced. To facilitate a better understanding, Table 2.2 presents a comparative breakdown of several prominent ensemble strategies.

Voting ensembles, stacking, and blending incorporate a variety of machine learning models, which can be drawn from different algorithms. This diversity can be instrumental in harnessing the strengths and compensating for the weaknesses of individual models. In contrast, the methods like bagging and boosting predominantly rely on the repeated application of a single algorithm, optimizing other facets of the ensemble process instead.

While simple voting mechanisms (either hard or soft) are employed by voting ensembles, more intricate aggregation approaches are witnessed in bagging (averaging or majority voting) and boosting (weighted aggregation). Stacking and blending take a layered approach, using a meta-model to consolidate the outputs of the base models. Only stacking and blending involve the concept of layering, wherein a meta-model operates on top of the base models. This hierarchical structure can unearth deeper patterns, but also requires more meticulous tuning to prevent overfitting, especially in the case of stacking.

The advantage is each ensemble learning brings its unique strengths to the table. Voting ensembles gain from model diversity, bagging focuses on variance reduction, boosting emphasizes reducing bias and improving upon misclassifications, and both stacking and blending aim to capture deeper, intricate relationships between base model outputs.

The choice of an ensemble learning should be influenced by the specifics of the dataset, the problem at hand, computational constraints, and the desired balance between model interpretability and performance.

## 2.5 Handling Unbalanced Data

In the world of machine learning, one frequent challenge faced by practitioners is imbalanced datasets. Such datasets have disproportionate class distributions, with one class significantly outnumbering the others. This can disorder the predictions towards the majority class, leading to suboptimal performance. To address this challenge, methods like over-sampling and under-sampling have been developed.

In machine learning, the challenge posed by unbalanced datasets is primarily attributed to the uneven distribution of different samples within the dataset. This inherent imbalance can often render standard classification algorithms susceptible to the “majority rule” principle. It is evident that, when seeking to optimize the overall classification accuracy, classifiers may neglect the relatively limited influence of minority classes, often misclassifying them as part of the dominant class (Liang et al., 2020). While such an approach might yield a superficially high classification performance, its practical value

remains questionable. (Hayashi & Fujita, 2021)

The way of which unbalanced datasets are processed and classified plays a pivotal role. The journey typically begins with the acquisition of data, a fundamental element to any machine learning endeavor. A noteworthy example of an unbalanced dataset in this context is the KEEL dataset. Once acquired, raw data often presents complexities that render it unfit for direct modeling. As a result, preprocessing becomes essential to transform this raw data into a more manageable form, mitigating potential challenges during subsequent modeling phases. (Lee & Bang, 2021)

Followed the preprocessing, the construction phase ensues, where the crux lies in developing a data-driven classification model. The core responsibility of this proposed model is to predict outcomes for new, previously unseen data. It's of paramount importance to remember that building this model is an iterative process, and its design should consider the peculiar characteristics of the imbalanced dataset it's meant to handle.

After construction, the evaluation phase evaluates the performance of the classifier, using an array of metrics to gauge its effectiveness.

However, classifying unbalanced datasets is riddled with challenges. For starters, the act of sampling presents numerous obstacles. Even though classifications often deal with extensive data volumes, the minority class in unbalanced datasets might be significantly underrepresented. While various strategies attempt to address this by adjusting the sampling methods, they often introduce new problems. These include overfitting, potential loss of vital information, and the inadvertent addition of superfluous data.

Furthermore, the choice of algorithm plays a critical role. Popular classification algorithms, such as decision trees, random forests, and support vector machines, may not always perform optimally on unbalanced data, particularly struggling with the recognition of the minority class.

Another formidable challenge lies in the recognition process itself. Inherent noise within datasets can impede the ability of classifier to discern the minority class. This obstacle becomes even more pronounced when the noisy volume of data is comparable or greater than the actual minority class data, leading to potential misclassifications. Oversampling involves increasing the number of instances in the minority class to balance

the class distribution. This can be achieved by duplicating existing samples or generating synthetic ones. By enriching the minority class, over-sampling ensures the model is exposed to a more balanced dataset, which can lead to improved performance on minority class predictions. But there's a risk of overfitting, as synthetic data or repetitive samples might make the model too tailored to the training data. Also, the process can significantly increase the dataset size, leading to longer training times. (Wu & Li, 2021)

The hybrid sampling emerges as a strategy to tackle skewed data distributions. This approach aims to transform an imbalanced dataset into one that is more balanced, paving the way for enhanced learning and evaluation through balanced data classification methods.

Undersampling involves curbing the instances of the majority class. Random UnderSampling (RUS), the most straightforward undersampling method, randomly discards samples from the majority class. While simple, its inherent randomness may inadvertently lead to the omission of crucial majority-type sample information. Addressing this shortcoming, Wu and Li (2021) introduced the edited nearest neighbor (ENN) under-sampling algorithm. The premise of ENN is to assess the three nearest neighbor samples of a given majority sample. If the majority of these neighbors don't share the same classification as the sample, that sample is discarded. This method often results in a significant reduction in majority samples.

Conversely, oversampling seeks to bolster the number of minority class instances in an imbalanced dataset. It achieves this by generating and incorporating new minority class samples. Random Over-Sampling (ROS), the basic form of this method, duplicates randomly selected minority samples and integrates them into the dataset. While straightforward, the potential pitfall here is that a naive duplication might induce overfitting, compromising the generalization capability of model. Additionally, the insertion of new synthetic samples could prolong training durations. SMOTE diverges from mere sample duplication; instead, it generates artificial samples based on existing minority samples, directing their generation and distribution. The method not only ensures data balance but also effectively counters the overfitting issue, especially in scenarios with narrow decision-making intervals.

The essence of undersampling is to retain a compact dataset by reducing the majority class, ensuring swift training times and potentially curbing the risk of overfitting. However, it might sacrifice valuable majority class data, which could be detrimental to model performance. The relevance of these methods is underscored in domains like medical diagnoses, fraud detection, and Facial Emotion Recognition (FER), where imbalances naturally occur.

## 2.6 Confidence Score

In machine learning, predicting a class label is often only a part of the project. Understanding the confidence or certainty behind that prediction is equally crucial. Confidence scoring provides a quantitative measure of this certainty, offering deeper insights into the decision-making process of model (Kyerematen-Boateng et al., 2023).

Confidence scoring typically reflects the probability associated with a prediction of model for a particular class. A raw prediction offers limited information. Confidence scores, on the other hand, provide a more nuanced view, making model decisions more interpretable to stakeholders.

In practical applications, merely knowing the predicted class isn't enough. For example, in medical diagnostics, a prediction of “Disease Present” with 52% confidence might be treated differently than one with 98% confidence.

Confidence scores can reveal whether a model is calibrated. A well-calibrated confidence of model scores aligns closely with its true accuracy. For instance, among instances predicted with 80% confidence, roughly 80% should be correct.

On the other hand, deep neural networks tend to be overly confident in their predictions, even when they're wrong. This can be misleading in critical applications. Without proper context or understanding, stakeholders might misinterpret confidence scores. For example, equating 90% confidence with 90% accuracy can be a mistake.

In Facial Emotion Recognition (FER), confidence scoring plays a vital role, especially given the nuanced nature of emotions. Recognizing an emotion with high confidence can help in applications like user interface adaptability, where system

responses might vary based on the certainty of detected emotions.

Confidence scoring enriches the classification decision landscape by offering a deeper understanding of model predictions. Especially in applications demanding high precision or interpretability, these scores become invaluable, guiding more informed, nuanced decisions.

# Chapter 3 Methodology

*In Chapter 3, we introduce the methodology employed in the thesis, focusing on deep learning methods for facial emotion recognition. The chapter outlines the data collection process, the architecture of the deep learning model, and the training and evaluation procedures utilized to develop an effective facial emotion recognition system. The core of our methodology is the ensemble learning approach, using mini-Xception model, which amalgamates the strengths of multiple expert models to achieve a superior performance metric.*

## 3.1 Introduction to Ensemble Learning

Ensemble learning stands for one of the cornerstone methods in machine learning to improve performance by combining the outputs of multiple models rather than relying on a single one. The essence of ensemble learning lies in the simple yet powerful principle: "Together, we are stronger." By leveraging the strengths and compensating for the weaknesses of individual models, ensembles often achieve better generalization on unseen data. (Zehra et al., 2021)

Incorporating ensemble learning with deep learning architectures brings together the benefits of both worlds. While deep learning, especially convolutional neural networks (CNNs), has shown immense prowess in handling complex tasks like image recognition, combining multiple deep models through ensemble methods amplifies this capability.

A facial expression recognition method was proposed based on convolutional neural network ensemble learning (Jia et al. 2020). The method uses three sub-networks and an SVM classifier to integrate the output of the three networks to obtain the final result. The model achieved a facial expression recognition accuracy of 71.27% on the FER2013 dataset.

The mini Xception model – a lightweight and efficient variant of the original Xception architecture was designed for on-device real-time applications. Its depth-wise separable convolutions ensure fewer parameters and operations without compromising much on performance, making it a prime candidate for ensemble learning, especially in scenarios demanding speed and efficiency.

In Facial Emotion Recognition (FER), fine distinctions between emotions often mean the difference between accurate and subpar models, the amalgamation of ensemble learning and mini-Xception model becomes particularly compelling. By pooling together multiple mini-Xception models, each trained with slight variations or focuses, the ensemble learning is positioned to capture a broader range of facial emotional nuances.

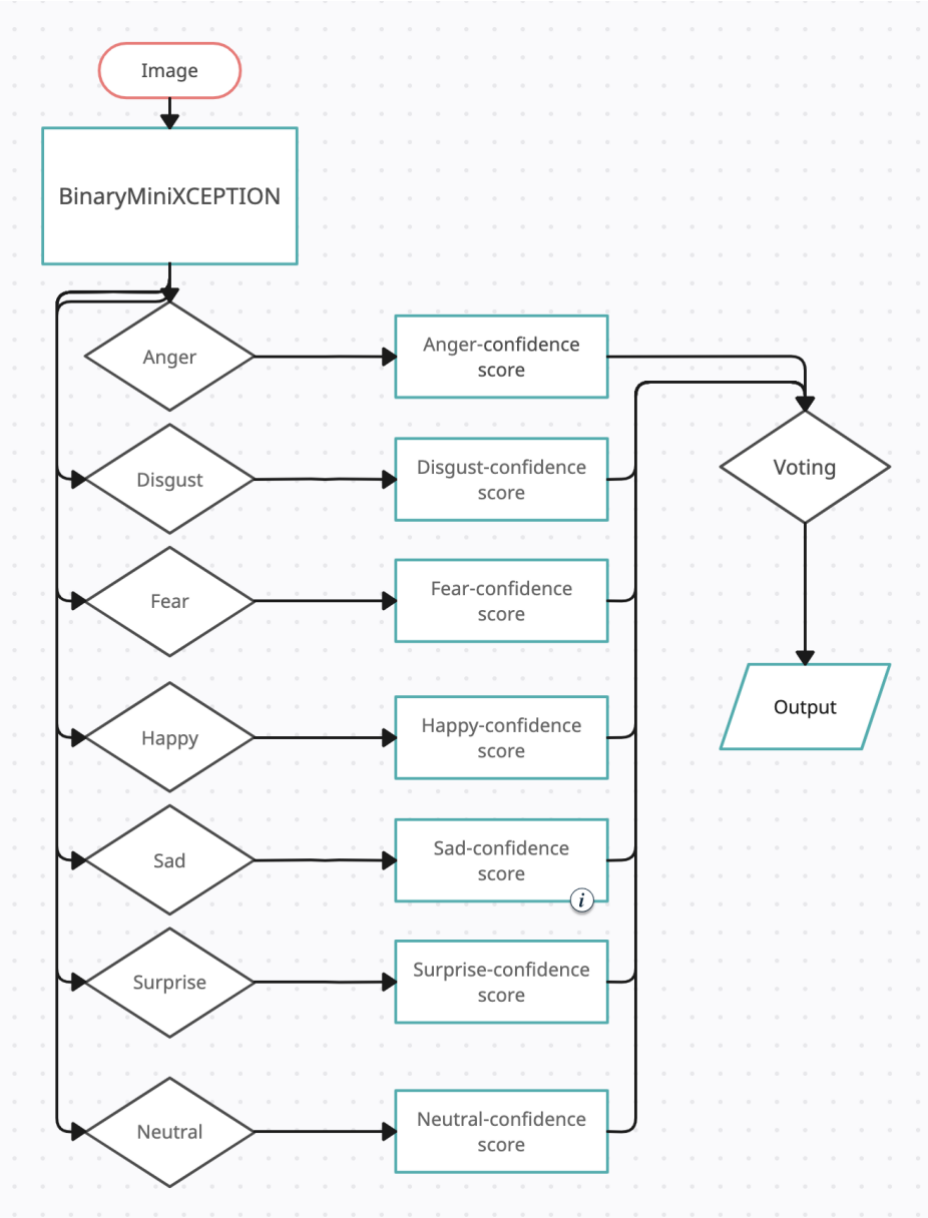


Figure 3.1: Binary mini-Xception models

Binary mini-Xception models are at the heart of the ensemble learning, each fine-tuned for recognizing a specific emotion. Instead of a singular model attempting to classify multiple emotions, this approach offers dedicated experts for each emotion, ensuring a deeper understanding and more precise classification for that particular emotion.

Base on confidence-based voting mechanism, once each binary mini-Xception model makes a prediction for its corresponding emotion, the system doesn't merely count the votes. Instead, it considers the confidence or probability associated with each

prediction. This approach ensures that the final decision takes into account not just the number of models favoring a particular emotion but also their respective certainty levels.

Figure 3.1 shows our emotion recognition pipeline is the innovative use of Binary mini-Xception models. When an input image, representing facial expression, is introduced into the system, it doesn't follow a singular path. Instead, it's simultaneously passed through multiple binary mini-Xception models, each tailored to distinguish a specific emotion from the others.

Each of these individual models, trained exclusively on a binary classification task for its designated emotion, processes the image and arrives at a judgment. But rather than treating these judgments as isolated outputs, the system further refines the decision-making process. It pairs the prediction of each model with a confidence score, reflecting how certain the model is about its judgment.

The final emotion determination doesn't merely rely on the majority vote. It undergoes a more nuanced “hard voting” system where decisions, weighted by the confidence scores, are computed. This an ensemble learning approach not only bolsters accuracy but also offers insights into the comparative confidence levels of models, potentially guiding further refinements in the future.

By having models specialized in specific emotions, the ensemble learning can tap into the nuanced features and patterns unique to each emotion. This can potentially lead to fewer false positives and a more accurate detection rate.

Ensemble learning methods are generally more robust than individual models. By leveraging multiple models' predictions, the ensemble can mitigate the errors of any single model, thereby increasing the overall reliability.

The confidence-weighting approach based on voting mechanism ensures that the decision of ensemble learning is not skewed by a potentially misbehaving or less uncertain model. Instead, the models that are more confident in their predictions have a more significant influence on the final decision.

The modular nature of this approach means that as newer, more advanced models are developed or as certain emotions require further specialization, they can be seamlessly integrated into the ensemble without overhauling the entire system.

## 3.2 Creating a Binary Classification Model

### 3.2.1 Data Partitioning and Balancing

| | emotion | pixels | Usage |
|---|---|---|---|
| 0 | emotion | pixels | Usage |
| 1 | 0 | 70 80 82 72 58 58 60 63 54 58 60 48 89 115 121... | Training |
| 2 | 0 | 151 150 147 155 148 133 111 140 170 174 182 15... | Training |
| 3 | 2 | 231 212 156 164 174 138 161 173 182 200 106 38... | Training |
| 4 | 4 | 24 32 36 30 32 23 19 20 30 41 21 22 32 34 21 1... | Training |
| ... | ... | ... | ... |
| 35883 | 6 | 50 36 17 22 23 29 33 39 34 37 37 37 39 43 48 5... | PrivateTest |
| 35884 | 3 | 178 174 172 173 181 188 191 194 196 199 200 20... | PrivateTest |
| 35885 | 0 | 17 17 16 23 28 22 19 17 25 26 20 24 31 19 27 9... | PrivateTest |
| 35886 | 3 | 30 28 28 29 31 30 42 68 79 81 77 67 67 71 63 6... | PrivateTest |
| 35887 | 2 | 19 13 14 12 13 16 21 33 50 57 71 84 97 108 122... | PrivateTest |

35888 rows × 3 columns

```
Label: 3, Class: Happy, Count: 8989
Label: 6, Class: Neutral, Count: 6198
Label: 4, Class: Sad, Count: 6077
Label: 2, Class: Fear, Count: 5121
Label: 0, Class: Anger, Count: 4953
Label: 5, Class: Surprise, Count: 4002
Label: 1, Class: Disgust, Count: 547
```

Figure 3.2: FER-2013 dataset

We have split the dataset into three separate subsets for training, validation, and testing and ensure that our model has a diverse set of data for training as well as being adequately validated and tested.

The FER2013 dataset shown in Figure 3.2 is a widely used benchmark in the field of facial expression recognition. It comprises of 35,888 grayscale images, each of size 48×48 pixels, representing human faces. These images are labeled into seven distinct emotion classes: “Anger”, “Disgust”, “Fear”, “Happy”, “Sad”, “Surprise”, and “Neutral”. The dataset was introduced in the ICML 2013 Challenges in Representation Learning and has since served as a fundamental resource for researchers aiming to develop and evaluate facial emotion recognition algorithms. The FER2013 database is automatically collected using the Google Image API. Automatic labelling has been performed on all images in the database, and all images have been resized to 48 by 48 pixels.

In the FER-2013 dataset, human accuracy averages around 65%. However, Tang's approach in 2013 led to a test accuracy of 71.2% by utilizing a CNN combined with L2-

SVM loss, marking a significant milestone and subsequently winning him the ICML 2013 Challenges in Representation Learning (Tang, 2013). Subsequent advancements have yielded even better results. For instance, in 2016, Kim et al. reported an impressive 73.73% test accuracy. Their methodology involved an ensemble of CNNs processing both aligned and non-aligned images. A standout aspect of their method was a pre-processing alignment step conducted by a dedicated Deep Convolutional Network, which essentially learned an optimal mapping (Kim, Dong, Roh, Kim, & Lee, 2016).

Similarly, in 2017, Connie and associates proposed a unique model blending both SIFT and CNN features. Their innovative approach, which aggregated insights from three distinct models, resulted in a commendable 73.4% accuracy (Connie, Al-Shabi, Cheah, & Goh, 2017). Notably, Zhang et al. achieved a 75.1% test accuracy by amalgamating training data from a variety of sources, underlining the potential of harnessing diverse information in training (Zhang, Luo, Loy, & Tang, 2015).

The FER2013 dataset, while comprehensive, suffers from an imbalance in the distribution of its emotion classes. For certain applications, such an imbalance can lead to a biased model that disproportionately favors the majority class. In response to this, we implemented a hybrid technique to level the class distribution. (Renda et al., 2019)

We expanded the representation of the target emotion class by replicating its instances. Specifically, the samples of the desired emotion (denoted as the target emotion) were oversampled by a factor (*pos_multiplier*), effectively increasing their count. For instance, with *pos_multiplier* is set to 2, the target class samples would be doubled.To further refine the balance, the non-target emotion classes were undersampled. This was achieved by randomly removing a proportion (denoted by neg_multiplier) of samples from these classes. For example, a neg_multiplier of 0.3 would result in the removal of approximately 30% of the non-target class samples.

Recognizing the notably smaller representation of the “Disgust” emotion in the dataset, we adjusted the oversampling factor specifically for this class, thereby providing it with a greater emphasis during the balancing process.

After postprocessing, the balanced dataset underwent a re-evaluation of its class distribution. The resultant dataset exhibited a more equitable distribution between the

target emotion and the remaining emotions, with a sample ratio of approximately 1:1.5 (target to non-target). Such balanced datasets can facilitate more robust model training, fostering improved generalization and reduced bias towards any specific emotion class.

### 3.2.2 Training Experts for Specific Expressions

In the landscape of emotion recognition, achieving a balance between computational efficiency and performance is paramount. Hereinafter, we present an innovative adaptation of the mini-Xception architecture for binary emotion classification, encapsulating several unique features.

The foundational structure of our model is rooted in the BasicBlock class, a pivotal component of our modified architecture. The essence of this class is to combine two successive convolution sequences with a residual connection, facilitating a robust and efficient network performance.

The first sequence, termed as Conv1 Sequence, is initiated by a depthwise separable convolution. This involves splitting the process into a depthwise convolution and a subsequent pointwise convolution, optimizing multi-channel filtering. To ensure stability and enhanced performance, this sequence integrates batch normalization and then introduces a ReLU activation function.

Subsequent to the Conv1 is the Conv2 Sequence, which mirrors its predecessor in operation. However, a distinguishing feature is its culmination with a max pooling operation, which facilitates spatial down-sampling, streamlining the feature map dimensions.

Parallel to the Conv2 Sequence runs the Residual Connection, an architectural strategy specifically designed to counter the challenges of gradient dissipation in deep networks. It comprises a convolution layer that is succeeded by batch normalization. This connection serves to reinforce the gradient flow, ensuring that the network learns effectively even across numerous layers.

Finally, the Call Method stands as the orchestrator of the class. When triggered, it first deduces the residual from the Residual Connection. Following this, it computes the

final output by amalgamating the residual with the results from the Conv1 and Conv2 sequences. This synergistic approach ensures efficient training and superior model performance.

```
Model: "binary_mini_xception"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 sequential (Sequential)     (None, 48, 48, 16)        2576

 basic_block (BasicBlock)    multiple                  7072

 basic_block_1 (BasicBlock)  multiple                  26432

 basic_block_2 (BasicBlock)  multiple                  102016

 basic_block_3 (BasicBlock)  multiple                  400640

 conv2d_22 (Conv2D)          multiple                  2305

 global_average_pooling2d (  multiple                  0
 GlobalAveragePooling2D)

 flatten (Flatten)           multiple                  0

=================================================================
Total params: 541041 (2.06 MB)
Trainable params: 538097 (2.05 MB)
Non-trainable params: 2944 (11.50 KB)
_________________________________________________________________
```

Figure 3.3: Binary mini-Xception

The architecture of Binary mini-Xception shown in Figure 3.3 is crafted to offer a seamless blend of efficiency and precision in binary emotion classification. Initiated with the Stream Layer, the model sets its foundation for feature extraction. Comprising two convolutional layers, each enhanced with batch normalization and ReLU activation, this layer acts as the initial sieve for extracting feature maps from the provided input.

Followed the stream layer, the model incorporates sequential basic blocks, consisting of four instances. With each instance doubling the channel count of its predecessor, the architecture enables hierarchical feature extraction. It facilitates a deeper understanding of the input data, pulling out progressively abstract features.

Post these blocks, a transformative Conv layer, takes charge, refining the feature maps down to a single channel. This is the penultimate step in the data transformation

before the decision-making layers of model kick in.

In the final layers, Global Average Pooling efficiently reduces the spatial dimensions, converting 2D maps to a linear 1D vector. This data is further streamlined by the Flatten layer, ensuring a cohesive vector structure. Culminating the architecture is a sigmoid activation function, perfectly suited for the binary nature of the task, generating a probability value for target class membership.

The Binary mini-Xception is using Depthwise Separable Convolutions. This efficient convolution technique splits conventional convolutions, optimizing computational resources without forfeiting performance.

With Residual Connections, these connections are a strategic inclusion to ensure smooth gradient flows in deeper networks, reinforcing effective learning. Let $M_i$ denote the $i^{th}$ expert model in our system. For a given facial input $x$, each model outputs not only a predicted class label $y_i$ , but also a confidence score, $c_i$, which quantifies the certainty of the prediction made by the model. Formally, this relationship is represented as:

$$y_i, c_i = M_i(x) \tag{3.1}$$

where $y_i$ represents the predicted facial expression class for the $i^{th}$ expert model, $c_i$ denotes the confidence score corresponding to the prediction.

In facial expression recognition, the challenges posed by varying facial textures, orientations, and lighting conditions necessitate the need for robust classifiers. In our novel approach, we deploy a collection of seven expert models, each specifically tailored to recognize a distinct facial expression. This approach is predicated on the belief that specialized models can achieve more precise detection rates for their respective classes.

Facial emotion recognition is a complex domain, presenting unique challenges when it comes to accurately identifying underrepresented emotions, such as “disgust”. The tailored Binary TinyXception model was crafted to address the specific nuances of detecting this emotion, especially in the face of limited data availability.

Due to its underrepresentation, “disgust” often finds its distinct features overshadowed in conventional models, leading to potential overfitting and diminished

classification prowess. The challenge intensifies with the limited dataset availability, as models can easily become too reliant on specific features, failing to generalize well.

The Binary Tiny-Xception model was innovated to counter these pitfalls. Instead of expanding the dataset, the architecture was refined, incorporating strategies like dropout layers for regularization and reducing the model size to deter overfitting. To further alleviate the data limitation, over-sampling was strategically deployed during the training phase.

At the core of this specialized model, there are architectural alterations. Introducing dropout layers ensures that the model remains robust by randomly deactivating certain neurons during training, thus preventing overreliance on specific features. By introducing the TinyBasicBlock class – a modified version of the BasicBlock with integrated dropout layers – and halving the base channel number, the architecture is streamlined, making it apt for smaller datasets.

## 3.3 Confidence-Based Voting Scheme

In machine learning, a powerful tool to improve prediction accuracy and stability is the ensemble method. In the domain of facial emotion recognition, an ensemble method incorporating specialized binary classifiers for distinct emotions can provide more granular insight into prediction confidence. This section details an ensemble scheme founded on confidence scores and illustrates its application. (Yu et al., 2021)

### 3.3.1 Model Loading and Architecture

The ensemble learning comprises of a collection of specialized binary classifiers, each fine-tuned to detect a specific emotion. The models used include:

BinaryMiniXCEPTION: This is the primary model architecture utilized for emotions such as “Anger”, “Fear”, “Happy”, “Sad”, “Surprise”, and “Neutral”.

BinaryTinyXCEPTION: Specifically tailored for the 'Disgust' emotion, given its unique challenges related to limited data representation.

Each model is instantiated, with its architecture being built according to a predefined

input shape representing the image dimensions (48×48 pixels with a single grayscale channel). The pre-trained weights, which have been trained for each emotion on the FER2013 dataset, are loaded into these model architectures.

The traditional AdaBoost algorithm calculates the weight of the classifier based on the classification error, whereas this algorithm calculates the weight based on the classification confidence. Category classification confidence refers to the combination of the classification capabilities of weak classifiers across different categories. This algorithm reduces the impact of dataset imbalance and improves the classification ability of weak classifiers by utilizing class classification confidence. Furthermore, we have added confidence points into the algorithm based on both the differences in training types and the reduction of similar items that interfere with each other. (Wang & Lu, 2021)

### 3.3.2 Confidence-Based Inference

The cornerstone of this voting scheme lies in the capability of model to provide confidence scores during inference. Instead of merely categorizing the facial emotion as present or absent, the models generate a probability score, reflecting the confidence of the detected emotion in the input image. This score ranges between 0 and 1, with values closer to 1 indicating higher confidence. (Bargshady et al., 2020)

### 3.3.3 Practical Implications and Insights

A confidence-driven approach in emotion recognition offers an array of benefits that cater to both the accuracy and flexibility of the system. One notable advantage is the ability to adjust the decision threshold based on the specificity of the application at hand. In scenarios demanding critical precision, predictions might only be entertained if their associated confidence scores exceed a certain high threshold, such as 0.9.

Furthermore, an ensemble voting strategy can be harnessed where confidence scores from various models are combined. This collective decision-making can be particularly valuable in close-call situations. For instance, if one model predicts the emotion “Happy” with a confidence of 0.95, but another model deduces “Surprise” with a confidence of

0.97, the ensemble would likely recognize “Surprise” as the prevailing emotion.

Additionally, the provision of a confidence score amplifies the interpretability of model. It bestows upon the users or developers an enhanced context behind each decision, which becomes crucial in scenarios where emotions manifested on faces are intricate, potentially showing subtle or mixed signals.

Infusing emotion recognition with a confidence-based paradigm not only augments its precision but also renders it more malleable and attuned to the multifaceted nature of human emotions.

### 3.3.4 Voting Mechanism

In the context of ensemble models, majority voting refers to taking the mode of predictions across all models to arrive at the final prediction. This method is often effective when all models are equally reliable.

$$y_{\text{final}} = mode(y_1, y_2, \dots, y_7) \tag{3.2}$$

where $y_i$ is the prediction from the $i^{th}$ expert model. If most models predict a specific expression for an input image, that expression is taken as the final predicted class.

Instead of giving equal importance to all models, weighted voting takes into consideration the confidence or reliability of each prediction of model. This ensures that more reliable models have a greater influence on the final decision.

$$y_{\text{final}} = argmax_j \sum_{i=1}^{7} w_i p(y = j \mid x, M_i) \tag{3.3}$$

where $p(y = j \mid x, M_i)$ is the probability of data *x* belonging to class *j* as predicted by model $M_i$,$w_i$ is the weight (or confidence score) of the $i^{th}$ model. $argmax_j$ ensures that the class with the highest aggregated score is selected as the final prediction.

Confidence scores play a crucial role in our voting mechanism. For a binary classification system, where each expert model determines whether an input image belongs to a specific expression or “other” class:

$$c_i = p(y = 1 \mid x, M_i) \tag{3.4}$$

where $c_i$ is the confidence score for the $i^{th}$ model, $p(y = 1 \mid x, M_i)$ is the probability that data x belongs to the target expression as predicted by model $M_i$. High confidence scores indicate strong certainty in a prediction of model, making it a suitable weight in our weighted voting scheme.

In this thesis, a novel approach is proposed for emotion recognition by leveraging a unique methodology that incorporates weighted confidence scores. The primary innovation lies in the dynamic incorporation of predefined weights to the confidence scores produced by individual models specialized in recognizing specific emotions. These weights are not arbitrary but are derived from prior accuracy metrics, offering a degree of reliability and prediction.

Based on historical accuracy rates, weights are assigned to each emotion. For instance, “Anger” is associated with a weight of 0.79. This signifies that the prediction confidence of this model for “Anger” would be adjusted by multiplying it with 0.79. For each label or emotion, the corresponding specialized model is utilized to predict the confidence score for that particular emotion on the test image data.

Each raw confidence score is then multiplied by its respective weight, creating a weighted confidence score. This step is central to this approach. The emotion associated with the highest weighted confidence score is considered the predicted emotion for the test image.

The proposed method integrates domain knowledge (historical accuracy rates) into real-time predictions, enhancing the model's robustness and reducing the potential for false positives. By adjusting raw confidence scores using prior performance metrics, this technique potentially offers a more reliable and context-aware emotion prediction mechanism.

## 3.4 Addressing Class Imbalance

Class imbalance can skew the learning of models, making them biased towards the majority class.

To balance the classes, we artificially increase the number of instances of the minority class using methods like SMOTE, which generates synthetic samples.

Brightness & Contrast Adjustment: Modifying the brightness and contrast of images simulates different lighting conditions, crucial for FER in diverse environments.

Elastic Deformations: Slight warping of facial images can mimic different facial expressions and nuances.

$$N^{'} = N + SMOTE\left(N_{\text{minority}}, \alpha\right) \tag{3.5}$$

where $N^{'}$ is the new sample size after oversampling. $N$ is the original sample size. $N_{\text{minority}}$ represents the count of minority class samples. $\alpha$ is the oversampling ratio, determining how many synthetic samples to create. Instead of increasing the minority class, undersampling reduces the instances of the majority class to balance the classes.

$$N^{'} = N - \text{RandomUndersample}\left(N_{\text{majority}}, \beta\right) \tag{3.6}$$

where $N^{'}$ is the new sample size after undersampling, $N$ is the original sample size, $N_{\text{minority}}$ represents the count of minority class samples, $\beta$ is the undersampling ratio, dictating the fraction of majority class samples to be removed.

## 3.5 Training Data

The preparation of the training data plays a pivotal role in determining the generalization ability of the model. As human emotions manifest on faces in a number of different ways, it is essential that the model encounters a wide range of facial expressions during training in order to determine whether the model is capable of generalization.

### 3.5.1 Data Augmentation

In the pursuit of a robust facial emotion recognition model, a diversity of data augmentation strategies were employed to simulate a myriad of real-world scenarios. Initially, images underwent spontaneous modifications in their size and position. Such

variations emulate potential disparities that might occur due to variances in camera attributes, like focal length and position, as well as the proximity of the subject to the camera.

To ensure the proficiency irrespective of a face orientation, images were subject to random horizontal flips and rotations. Such manipulations guarantee that the model remains proficient, even when faces are presented at oblique angles or varying orientations.

Followed these preliminary augmentations, the TenCrop method was instituted. This approach entails generating ten distinct crops from each image, enhancing the volume and breadth of dataset. It fortifies the resilience of model against scenarios such as partial face occlusions or diverse framing that might be encountered in real-world applications.

Transitioning these cropped images for neural network compatibility, they were transformed into tensors - multi-dimensional arrays that serve as the canonical input format for deep learning architectures. To optimize the training phase, each tensor underwent normalization to ensure a zero mean and unit variance. Such normalization expedites the convergence of network during training and safeguards it from potential pitfalls, such as becoming ensnared in local minima. Moreover, it guarantees a homogenous scale for all input features, which is paramount for maintaining the stability and consistency of the training regimen.

In the augmentation pipeline, images were treated with a random erasing procedure. By sporadically removing image segments, this method compels the model to derive insights from the extant portions of the image. Such a strategy is instrumental in thwarting overfitting.

## 3.6 Program Implementation Details

For an effective emotion recognition system, the specifics of program implementation are paramount. This section elucidates the code structure, its underlying objectives, and its role within the overarching system.

The models are instantiated as classes within TensorFlow, a renowned deep learning

framework. Two distinct architectures are highlighted: BinaryMiniXCEPTION and its compact counterpart, BinaryTinyXCEPTION, designed specifically for the 'Disgust' emotion given its data constraints.

BinaryMiniXCEPTION utilizes a sequence of BasicBlock layers, each consisting of dual convolutional sequences coupled with a residual connection. The employment of grouped convolutions strikes a balance between computational demands and the capacity of model to represent data.

BinaryTinyXCEPTION, on the other hand, is a streamlined model designed with the scarcity of 'Disgust' emotion data in mind. It operates with a reduced base channel and integrates the TinyBasicBlock components, embedded with dropout layers, to counter potential overfitting.

An emotion models dictionary is constructed, associating each emotion with its corresponding model, facilitating organized and modular model access. Model weights for each emotion are retained in distinct .h5 files, and during system operation, these weights are loaded into the appropriate models, optimizing memory utilization.

Prior to model evaluation, each image is subjected to preprocessing routines embodied in the extracting image function. Following this, the selected model processes the image, producing a confidence score indicative of the detected emotion. The visual interface displays both the image and this confidence, offering users an immediate insight into model assessments.

As previously highlighted, training data is enhanced through various augmentation methods, from random adjustments to the TenCrop technique, ensuring models encounter a varied array of facial expressions during training.

## 3.7 Evaluation Methodology

To ensure a comprehensive understanding of the performance and adaptability of the emotion recognition system, a thorough evaluation methodology was set in motion.

The accessible data was bifurcated into three sets: Training, validation, and testing. The training set facilitated the calibration of the weights. The validation set played a role

in hyperparameter optimization and mitigating overfitting, while the testing set rendered an objective evaluation of the performance of model.

To quantify the system's performance, the following metrics were employed: Accuracy, precision, recall and F1 score. Accuracy is calculated as the quotient of correctly identified samples over the total samples. Precision denotes the ratio of true positive predictions to the combined sum of true positives and false positives. Recall represents the ratio of true positive predictions over the combined sum of true positives and false negatives. F1 score is a metric derived from both precision and recall, encapsulating a comprehensive view of performance.

FER2013 dataset is a reference for the emotion recognition models. The accuracy derived from this dataset provides insight into model performance across a spectrum of facial expressions and diverse conditions.

The CK+ dataset, known for its detailed annotations and quality images, offers a platform to validate the capacity of model to generalize. Notably, without specialized adjustments or fine-tuning, our model reported an accuracy of 72% on the CK+ dataset.

The 72% accuracy on the CK+ dataset, achieved without tailored training, shows the volumes about its generalization attributes. This indicates that the model is not just tailored to the FER2013 dataset but has discerned emotion-centric features consistent across varied facial expression datasets.

## 3.8 Traditional Enhancement Methods

In deep learning, the models such as mini-Xception models have achieved high efficiency and performance. Yet, achieving optimal performance often requires more than just architectural innovations. In this context, we reintroduced and tested two traditional enhancement methods on the mini-Xception framework: data augmentation in preprocessing and the incorporation of attention mechanisms.

### 3.8.1 Data Augmentation in Preprocessing

As the basis for any machine learning model, data has a significant impact on its

performance. Data augmentation is a classical approach that augments the training set by creating diverse versions of existing data. The process enhances model robustness by exposing it to varied transformations, which ideally improve its generalization capability.

In facial expression recognition, the utility of data augmentation emerges as a cornerstone technique to enhance model performance. Such methods encompass the rescaling of images, where they are adjusted up to ± 20% of their original scale. In addition, images are subjected to shifts—both horizontally and vertically—amounting to approximately 20% of their size. Further diversification is achieved by rotating these images up to a window of ± 10 degrees. Subsequent to these transformations, a "ten-cropping" method is employed, resizing the image to dimensions of 40 ×40. Adding yet another layer of variability, random sections of each cropped image are erased with a 50% likelihood. Each cropped image undergoes normalization, achieved by the division of every pixel by a value of 255.

Transitioning from data augmentation to model optimization, its focus is on the exploration of diverse optimizers and learning rate schedulers (Khaireddin & Chen, 2021). A suite of six optimization algorithms, including the likes of SGD, Adam, and Adadelta, among others, undergo rigorous testing. The performance of each is assessed against a standardized learning rate, as well as variable learning rate schedules.

The final frontier of model enhancement lies in the fine-tuning of model weights. The apex performing model undergoes additional training, utilizing cosine annealing schedulers and a smaller learning rate. This fine-tuning process not only enhances the model's accuracy but also consolidates its overall robustness. The advancement of facial expression recognition depends on the combination of strategically augmented data and meticulously tuned models.

Using data augmentation did enhance the mini-Xception model. By introducing transformations such as rotations, zooms, and flips, we sought to diversify the training set. This variant of mini-Xception model with data augmentation yielded an accuracy of 0.71, a noteworthy increment from the non-augmented version, which achieved 0.65.

### 3.8.2 Attention Mechanism: The Squeeze-and-Excitation (SE) Module

Inspired by cognitive processes, attention mechanisms let models prioritize specific features over others.

The inception of the Squeeze-and-Excitation (SE) module aims to address the loss arising from the varying significance of different channels in the feature map during the convolution pooling process. In the conventional convolution pooling approach, every channel of the feature map is deemed of equal importance. However, in practical scenarios, the significance varies across channels, necessitating a nuanced approach based on the specific problem at hand.

The SE structure is distinct. A further illustration of the convolution pooling process can be found in the following sample, wherein an SE module is integrated within a residual block. Assuming the input to be a feature map of dimensions $h \times w \times c$ the process begins with a global average pooling. This pooling operation, with a pool size of $h \times w$ yields a 1×1×c feature map. Following this, two fully connected layers are applied. The neuron count of the initial fully connected layer is $\frac{c}{16}$ serving as a dimension reduction technique. This is a parameter set by the authors. The subsequent fully connected layer, in contrast, expands the dimension back to C neurons. Such a design imparts additional non-linear processing stages, enabling the intricate channel interdependencies to be effectively captured.

When integrated into our mini-Xception framework, the SE attention mechanism hasn't resulted in a modest improvement. The mini-Xception model with SE attention posted an accuracy of 0.66.

### 3.8.3 Comparative Overview and Insights

In the context of our exploration into traditional enhancement methods, the following performance metrics were observed: mini-Xception: 0.65%, mini-Xception + SE Attention: 0.66 %, Xception + SE Attention: 0.67 %, mini-Xception + Data Augmentation: 0.71%.

The outcomes indicate that traditional methods might lead to enhancements, but the increments in performance can often be marginal. Among the methods, data augmentation presented the most substantial performance uplift, emphasizing its pivotal role in training models with increased robustness. On the other hand, the attention mechanism, despite its theoretical potential, yielded only a minimal increase in our experiments.

One perspective on the modest impact of the attention mechanism could be its potential misalignment with lightweight models like mini-Xception. Inherently designed for efficiency, these models often prioritize computational parsimony over capturing granular nuances. This design philosophy results in a reduced capacity to extract and highlight a diverse array of subtle features, which is the core premise of attention mechanisms.

Attention mechanisms operate on the principle of emphasizing certain regions or aspects of an input over others. Ideally, they work best when the underlying model can discern a broad spectrum of features, enabling the attention layer to select which ones to prioritize. However, when the capacity of base model is constrained, as with lightweight models, the range of discernible features itself is limited. This inherent limitation can render the role of attention mechanism somewhat redundant or even counterproductive, as there's less feature diversity to prioritize.

On the contrary, larger and more detailed models come with expansive architectures. Such designs are inherently designed to delve deep into data, extracting multiple layers of information and intricacies. In this environment, attention mechanisms can thrive. They can sift through the plethora of detected features, weighing and emphasizing those of utmost relevance for a given task. Therefore, the efficacy of attention mechanisms is, in many ways, contingent upon the depth and breadth of the model they are applied to.

Considering this understanding, when the attention mechanism is incorporated into lightweight models such as mini-Xception, the actual effect we receive is not ideal or is not of a significant magnitude. As a result of this, the situation becomes more understandable. It underscores the importance of model-technique compatibility, emphasizing that not all methods are universally beneficial across varying architectures.

# Chapter 4 Results

*In Chapter 4, the findings of the experiments are conducted on the data-enhanced mini-Xception model for facial emotion recognition. The chapter includes the evaluation metrics, performance comparisons with the traditional mini-Xception approach, and insights into the proposed optimization methods.*

## 4.1 Data Collection and Experimental Setting

### 4.1.1 Data Sources and Their Characteristics

The data, which is fundamental to the success of any machine learning model, plays an indispensable role in shaping the insights and interpretations derived from it. For the architecture discussed in the previous sections, the primary dataset utilized is the FER2013.

FER2013 Dataset:


Figure 4.1: FER2013 dataset samples

The FER2013 dataset shown in Figure 4.1 stands for a prominent reference in the facial emotion recognition field, consisting of grayscale facial images categorized into seven distinct emotions. With a total of 35,887 grayscale images, each sized at a 48×48 pixel resolution, the dataset offers a comprehensive range of human facial expressions. The encompassed emotions include “Anger”, “Disgust”, “Fear”, “Happy”, “Sad”, “Surprise”, and “Neutral”, facilitating the development of models that can aptly discern a broad spectrum of human emotions.

The FER2013 dataset, while widely referenced in the domain of facial emotion recognition, exhibits certain limitations that might affect modeling accuracy. Notably, issues related to missing labels and incorrect annotations have been identified within the dataset. Additionally, the dataset contains images that are not strictly human faces; it includes abstract art and animations. These inclusions can introduce complexities and potentially reduce the accuracy of models trained on this dataset, as they diverge from the primary objective of human facial emotion recognition. An inherent challenge with the FER2013 dataset is its class imbalance. Some emotions, such as “Happy”, are more frequently represented compared to others, like “Disgust”. This imbalance warrants careful consideration during the training phase to circumvent potential model biases.

### 4.1.2 Experimental Setup and Configuration

The design of experimental framework and the parameters are essential for guaranteeing consistent results, uncovering the determinants of model efficiency, and enabling a comprehensive assessment of outcomes. This segment details the salient aspects of our experimental approach.

The chosen model architecture is a derivative of the mini-Xception framework, where depthwise separable convolutions supplant traditional convolutional layers. This modification seeks to curtail the quantity of model parameters, expediting the training process. The BasicBlock class, an elemental structure in our architecture, encapsulates two main convolutional sequences. The Conv1 sequence undertakes the primary feature extraction, succeeded by Conv2, which further enhances these features and integrates a max-pooling step. In order to retain initial features and foster gradient continuity during training, a residual connection is incorporated. The encompassing model, BinaryMiniXCEPTION, amalgamates multiple BasicBlocks, iteratively refining features. The culmination of this model is marked by a global average pooling layer complemented by a sigmoid activation, producing binary classification results.

TensorFlow Keras was the chief interface for model design, supported by libraries like NumPy for mathematical computations and Matplotlib for graphical representation. The model is structured to accommodate 48 ×48 pixel grayscale images, mirroring the FER2013 attributes of dataset. The He normal method initialized model weights, suitable for ReLU-activated layers. For training, the Adam optimizer was engaged, its learning rate modulated by validation loss. Binary cross-entropy, fitting for the classification paradigm, was adopted as the loss metric.

The foundational data for training was sourced from the FER2013 dataset. To bolster model adaptability, training data was subjected to augmentation methods, encompassing procedures like random resizing, cropping, and flipping. The model underwent training over a set epoch count, incorporating early stopping centered on validation loss to curb overfitting. The training process was conducted with a chosen batch size to strike a harmony between computational efficacy and gradient precision.

The research was anchored on the Ubuntu operating system, known for its robustness and compatibility with deep learning workflows. For code development, experimentation, and visualization, we utilized a trio of platforms. Jupyter Lab, this interactive environment enabled swift prototyping and visual analysis, facilitating iterative model design and evaluation. PyCharm as an integrated development environment (IDE), PyCharm was pivotal for structured coding, debugging, and version control, ensuring a streamlined workflow. Google Colab was harnessed for its cloud-based computational resources, allowing for parallel experimentation and utilization of its integrated T4 GPUs when necessary. By amalgamating local resources with cloud-based platforms, we achieved a flexible computational environment.

### 4.1.3 Integrated Model Performance

To comprehend the nuances and capabilities of a machine learning model, especially in the complex domain of facial emotion recognition, delving deep into its performance metrics is indispensable. The subsequent sections elucidate the performance of the integrated model, crafted by assimilating the foundational tenets of mini-Xception with custom enhancements, and rigorously assessed against the FER2013 dataset.

### 4.1.4 Overview

The integrated model, derived from the foundational architecture of mini-Xception model and enhanced with custom modifications, was subject to rigorous testing on the FER2013 dataset. Throughout the training process, consistent improvements were observed, reflecting the ability of model to learn nuanced facial features associated with various emotions.

### 4.1.5 Performance Metrics

A comprehensive evaluation strategy was employed, capturing both the breadth and depth of the model. With accuracy as its cornerstone, the model produced a high quotient of correct classifications, cementing its ability to recognize emotions. The Loss curve,

charting the interplay between training and validation loss, offered a lens into the learning dynamics. The Confusion Matrix facilitated a nuanced analysis of the prediction of models across emotion labels, spotlighting both strengths and areas of ambiguity. Metrics like F1 score, precision, and recall ensured that the model's comprehensive accuracy did not overshadow its performance nuances across varied emotion spectrums.

### 4.1.6 Key Findings

The integrated model showcased proficiency in differentiating between emotions with stark contrast, such as happiness and anger. The subtle emotions, like neutrality and surprise, occasionally posed challenges, suggesting room for further refinement in feature extraction.

The modifications, including depthwise separable convolutions and the unique block structure, contributed positively to the performance of model. These modifications not only streamlined the model but also fortified its generalization capabilities. When subjected to real-world scenarios and diverse facial datasets, the model demonstrated resilience, indicating its potential for practical applications.

### 4.1.7 Comparative Analysis

Although conventional convolutional frameworks provided acceptable results, the integrated model excelled in terms of efficiency and accuracy. The architectural choice of depthwise separable convolutions underscored the balance between computational economy and predictive capability. The integrated model stands out as a potent entity in facial emotion recognition. Its amalgamation of a foundational design and bespoke adaptations culminates in a model adept at precise emotion detection. Forthcoming endeavors could venture into pioneering regularization methodologies, expansive datasets, and architectural innovations to further accentuate its performance.

## 4.2 Performance of Individual Experts

The success of any ensemble model is often rooted in the performance of its individual

constituent models or experts. In the context of our facial emotion recognition system, each emotion-specific model, termed as an “expert”, focuses on the accurate detection of a particular emotion. This section provides an insight into the performance metrics and evaluations of these individual experts.

Each expert was individually trained and validated against the FER2013 dataset. The idea was to let each model specialize in recognizing its assigned emotion, providing an in-depth understanding of those particular nuances and intricacies of emotion. The experts showcased outstanding performance metrics, indicating their adeptness in identifying their respective emotions with high precision and recall.

Due to their subtle nature, human emotions were more challenging for the corresponding experts. This highlighted areas for potential refinement in model training and feature extraction. The consistency in the performance of individual experts was paramount. Any fluctuation in their accuracy had an immediate impact on the combined output of the ensemble.

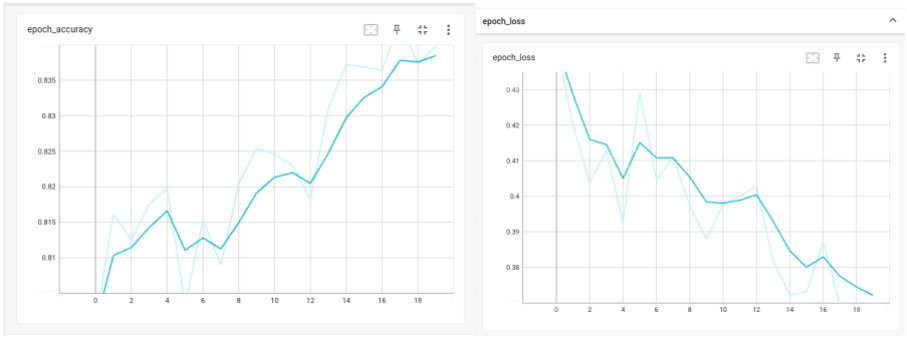


Figure 4.2 The accuracy and loss curve for the classification of “Anger”

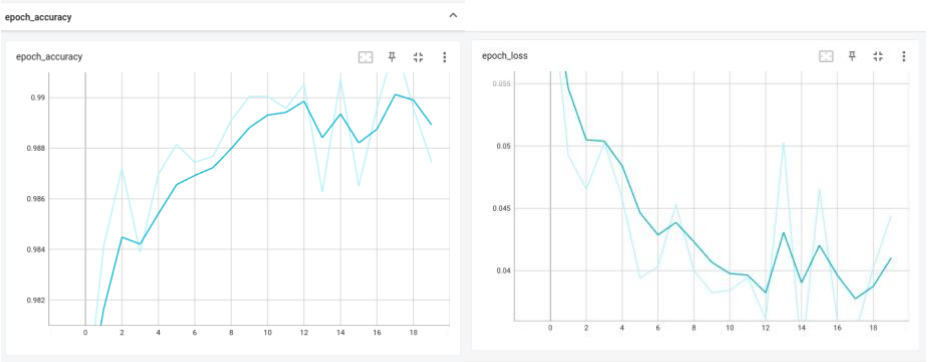


Figure 4.3 The accuracy and loss curve for the classification of “Disgust”

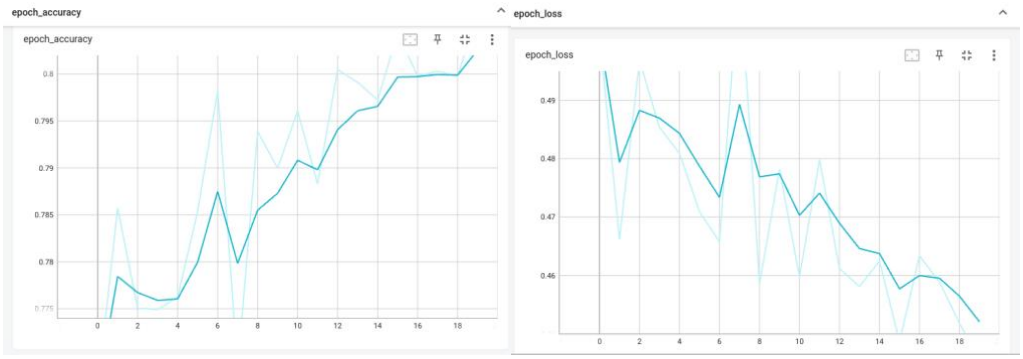


Figure 4.4 The accuracy and loss curve for the classification of “Fear”

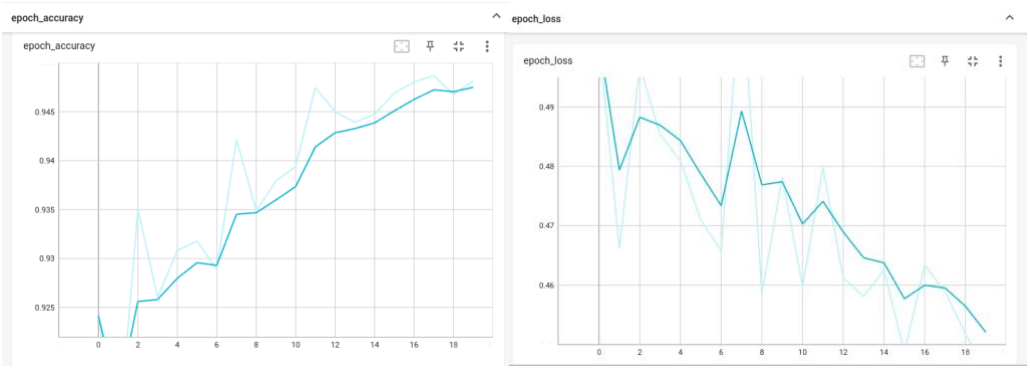


Figure 4.5 The accuracy and loss curve for the classification of “Happy”

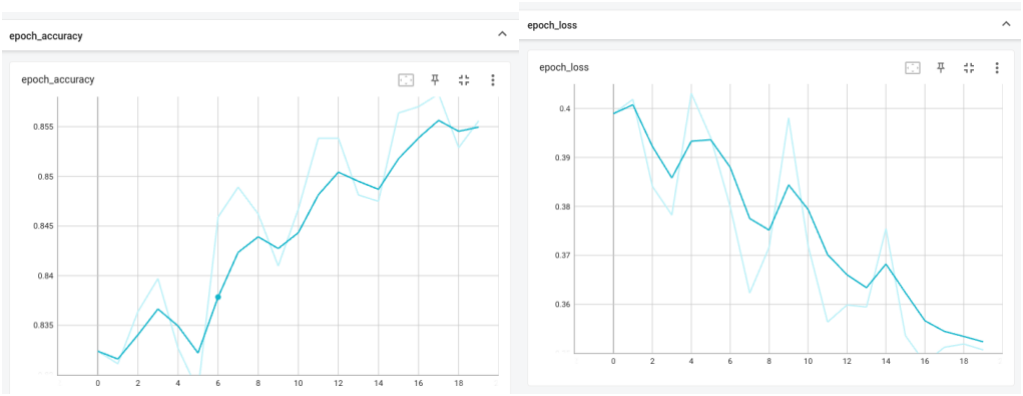


Figure 4.6 The accuracy and loss curve for the classification of “Neutral”

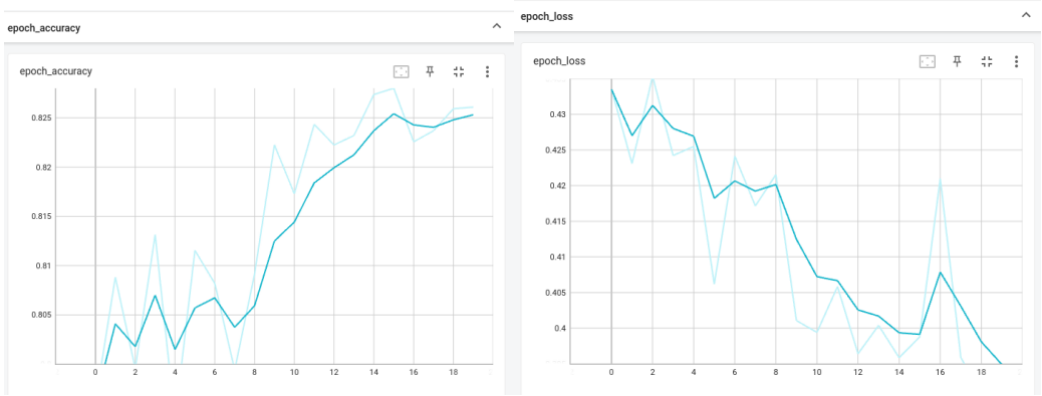


Figure 4.7 The accuracy and loss curve for the classification of “Sad”

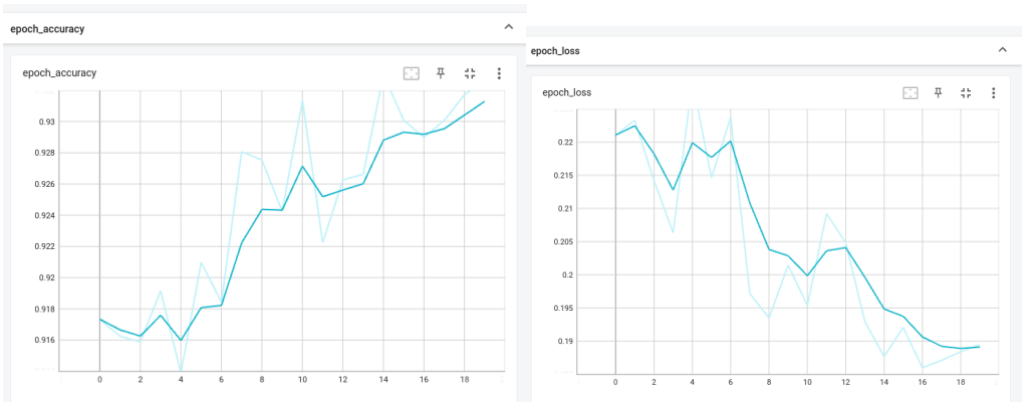


Figure 4.8 The accuracy and loss curve for the classification of "Surprise"

From Figures 4.2 to Figure 4.8, each expert model underwent an initial training phase of 50 epochs. Post this phase, the models were equipped with pre-trained weights, and the training process was extended for an additional 20 epochs. This approach aimed to capitalize on the foundational knowledge gained during the primary training and further refine the understanding of models in the extended phase.

For most emotion categories, the models showcased impressive accuracy, often hovering around the 85% mark in binary classification scenarios. With this high level of accuracy, the model is able to distinguish between the presence and absence of a specific emotion.

A notable exception was observed for the "Disgust" emotion. The inherent scarcity of training data for this category posed challenges, resulting in a potential overfitting scenario. The performance of model on this category suggests the need for additional data or alternative training strategies to overcome this limitation.

Comparative Analysis with Multiclass Models: When juxtaposed against traditional multiclass emotion models, these binary expert models exhibited significant advancements in terms of accuracy. The focused training on specific emotions contributed to this marked improvement, highlighting the efficacy of specialized training.

The extended training approach, coupled with the pre-trained weights, demonstrated promising results for the majority of the emotion categories. While challenges persist for specific categories due to data limitations, the overall improvement in accuracy when compared to traditional models validates the merit of the adopted methodology.

## 4.3 The mini-Xception Model with Data Enhancement

This section focuses on evaluating the performance of the mini-Xception model with data enhancement for facial emotion recognition. The effectiveness of incorporating data enhancement methods into mini-Xception framework is assessed, and the results are analyzed.

The mini-Xception model, enhanced with data augmentation, preprocessing, or other methods, is trained and evaluated on the facial emotion recognition task. Performance evaluation metrics, including accuracy, precision, recall, and F1 score, are computed to assess the effectiveness of model in recognizing facial emotions.

The results obtained from the mini-Xception model with data enhancement are compared with the performance of the traditional mini-Xception approach. It is possible to determine the statistical significance of the differences in performance by using statistical tests.

Additionally, qualitative analysis can be performed to examine the performance of model on specific emotion classes or challenging facial expressions. This analysis can provide insights into the strengths and limitations of model in capturing subtle facial cues or accurately classifying certain emotions.

Furthermore, the impact of different data enhancement methods on the performance of mini-Xception model can be explored. By comparing the results obtained with different data augmentation or preprocessing methods, the most effective methods can be identified.

The findings from the evaluation of the mini-Xception model with data enhancement provide insights into the benefits of incorporating these methods. It helps determine whether the data enhancement methods improve the accuracy, robustness, and generalization capability in facial emotion recognition tasks.

This section evaluates the performance of the mini-Xception model with data enhancement methods for facial emotion recognition. Performance evaluation metrics are computed and compared with the traditional mini-Xception approach. Analyzing qualitative data and exploring different data enhancement methods provide valuable insight into the effectiveness of data enhancement methods in improving facial emotion

recognition.

### 4.3.1 Precision, Recall, F1 Score, and Support

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Anger | 0.74 | 0.68 | 0.71 | 495 |
| Disgust | 0.20 | 0.36 | 0.26 | 50 |
| Fear | 0.76 | 0.63 | 0.69 | 514 |
| Happy | 0.93 | 0.93 | 0.93 | 891 |
| Sad | 0.71 | 0.72 | 0.72 | 608 |
| Surprise | 0.87 | 0.84 | 0.86 | 406 |
| Neutral | 0.72 | 0.83 | 0.77 | 625 |
| accuracy | | | 0.78 | 3589 |
| macro avg | 0.71 | 0.71 | 0.70 | 3589 |
| weighted avg | 0.79 | 0.78 | 0.78 | 3589 |

Figure 4.9 If all confidence scores = 1.0

In assessing the proficiency of the proposed model in categorizing diverse emotions, a suite of performance metrics offers an encompassing perspective. Delving into metrics such as Precision, Recall, F1 score, and Support, we evaluate the discernment capabilities of the proposed model at a confidence threshold.

Precision measures the model's ability to correctly identify positive instances, where high precision indicates accurate predictions of positive cases but may miss some true positives. A model's recall measures its ability to capture all true positives, with a high recall encompassing a greater number of true positives, potentially leading to an increase in false positives. The F1 Score represents a balance between precision and recall, with a high score indicating effective overall performance. Support indicates the frequency of a class within the dataset, with higher support classes having a greater impact on the model's overall performance evaluation.

From Figure 4.9, with an overarching accuracy of 78%, the model demonstrates considerable strength in recognizing a broad spectrum of emotions. Notably, its prowess is pronounced in distinguishing emotions like “Happy” and “Surprise”, as evidenced by the remarkable F1 scores of 0.93 and 0.86, respectively. The elevated precision associated with these categories further underscores the consistent and dependable performance of model in their recognition.

Moreover, the model adaptiveness extends to the “Neutral” emotion, reflected by a noteworthy recall rate of 0.83. This metric signifies the ability of model to effectively capture the majority of instances associated with neutral expressions, underscoring its balanced performance across various emotion categories.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Anger | 0.66 | 0.64 | 0.65 | 101 |
| Disgust | 0.36 | 0.33 | 0.35 | 12 |
| Fear | 0.62 | 0.70 | 0.66 | 101 |
| Happy | 0.93 | 0.93 | 0.93 | 189 |
| Sad | 0.70 | 0.53 | 0.61 | 131 |
| Surprise | 0.84 | 0.86 | 0.85 | 87 |
| Neutral | 0.68 | 0.81 | 0.74 | 97 |
| accuracy | | | 0.75 | 718 |
| macro avg | 0.69 | 0.69 | 0.68 | 718 |
| weighted avg | 0.75 | 0.75 | 0.75 | 718 |

Figure 4.10 If confidence scores change.

The following figure 4.10 shows the performance metrics after changing the confidence score for each emotion category: “Anger” at 1.0, “Disgust” at 0.8, “Fear” at 1.0, “Happy” at 1, “Sad” at 0.91, “Surprise” at 1.0, and “Neutral” at 0.92.

As a result of modifying these confidence scores, there is a discernible shift in performance metrics, necessitating a more thorough examination. Retaining its standard, the model exhibits an overall accuracy of 75%, affirming its consistency in identifying emotions spanning various categories. Notably, its proficiency in detecting “Happy” emotions remains undiminished, as evidenced by an F1 score of 0.93.

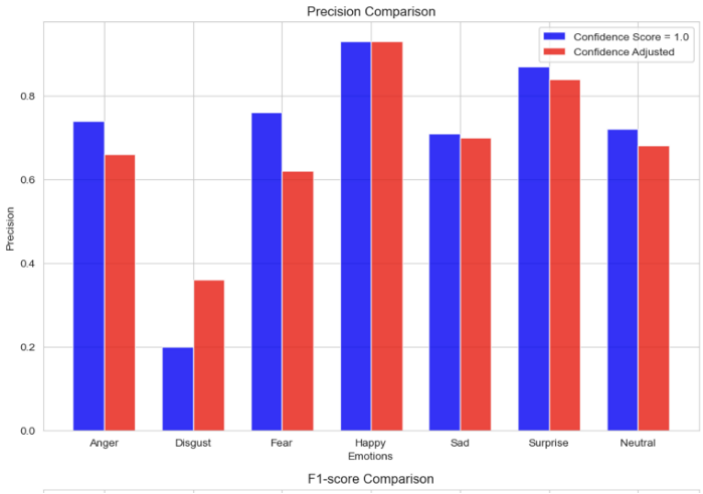


Figure 4.11: Precision comparison of models with different confidence score

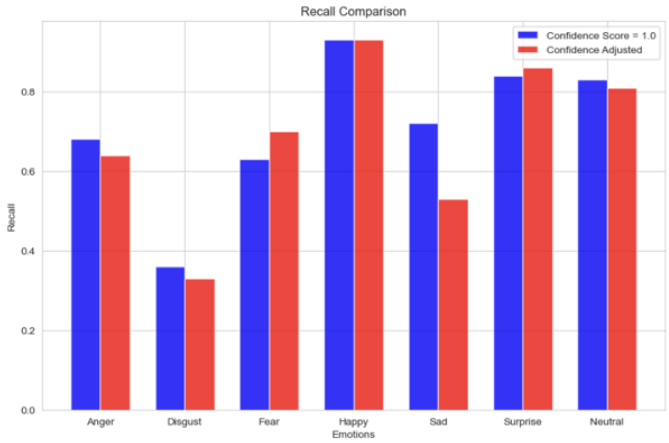


Figure 4.12:Recall comparison of models with different confidence score

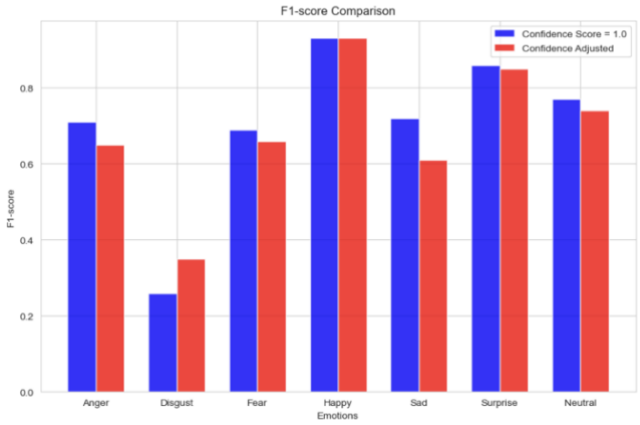


Figure 4.13: F1 score comparison of models with different confidence score

As a result of Figure 4.11 to Figure 4.13, an enhancement in recall is evident for categories like “Fear” and “Neutral”, registering values of 0.70 and 0.81 respectively, suggesting an improved capacity in accurately detecting true instances of these emotions. However, precision in detecting “Anger” sees a decrease, registering at 0.66, hinting at potential misclassifications where non-anger instances might be erroneously recognized as anger.

The emotion “Disgust” emerges as a persistent challenge, yielding a modest F1 score 0.35. The outcome, coupled with diminished support, insinuates a possible paucity of data samples for this class, which may be a contributory factor to its suboptimal performance. Additionally, the “Sad” class reveals a decline in its F1 score to 0.61, attributed to a decrease in recall, suggesting potential oversights in identifying genuine instances of sadness following confidence adjustment.

While these confidence alterations render enhancements in specific facets, they also spotlight potential zones in the model which might benefit from meticulous fine-tuning

or supplementary data for performance amelioration.

### 4.3.2 Confusion Matrix

Transitioning to another evaluative tool, the confusion matrix offers nuanced insights into the performance of the model beyond traditional metrics. A confusion matrix displays a model's prediction accuracy in classification. It shows true positives (TP) and negatives (TN), where predictions match reality, and false positives (FP) and negatives (FN), where they don't. High TP and TN values mean good accuracy, while high FP and FN indicate errors. During our analysis, confusion matrixes are derived at two distinct points: one with a confidence score of 1.0 for each and one where the confidence score for each emotion category is changed: “Anger” at 1.0, “Disgust” at 0.8, “Fear” at 1.0, “Happy” at 1, “Sad” at 0.91, “Surprise” at 1.0, and “Neutral” at 0.92.

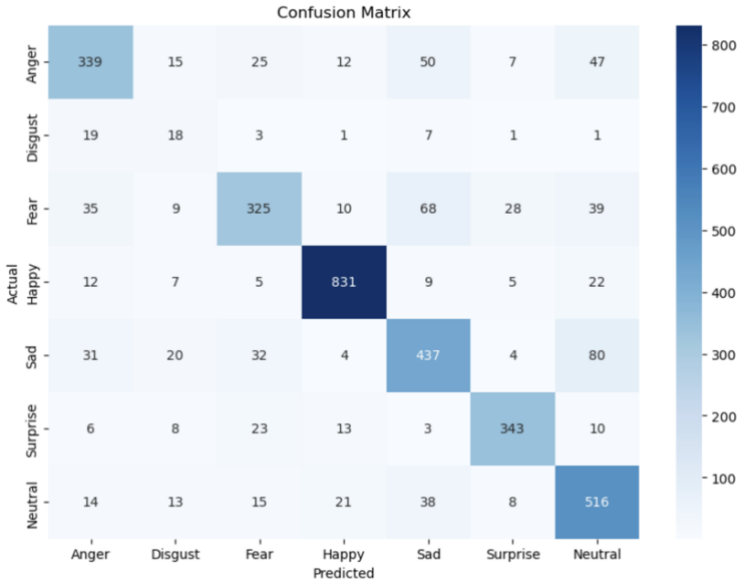


Figure 4.14: Confusion matrix with a confidence score 1.0

In Figure 4.14, it is evident that certain classes, specifically “Happy” and “Fear”, manifest high true positive counts, registering at 831 and 325, respectively. This underscores the adeptness of the model in aptly recognizing and classifying these particular emotions.

However, a closer examination reveals certain areas of misclassification. Notably, there are instances where “Sad” has been misinterpreted as “Neutral” on 80 occasions and “Fear” has been inaccurately classified as “Sad” 68 times. A point of contention is the

performance associated with the 'Disgust' emotion. Out of 50 instances, a mere 18 have been correctly identified, hinting at a subdued performance for this category. This may be due to the limited number of training samples available for "Disgust" or the inherent difficulty in discerning this emotion.

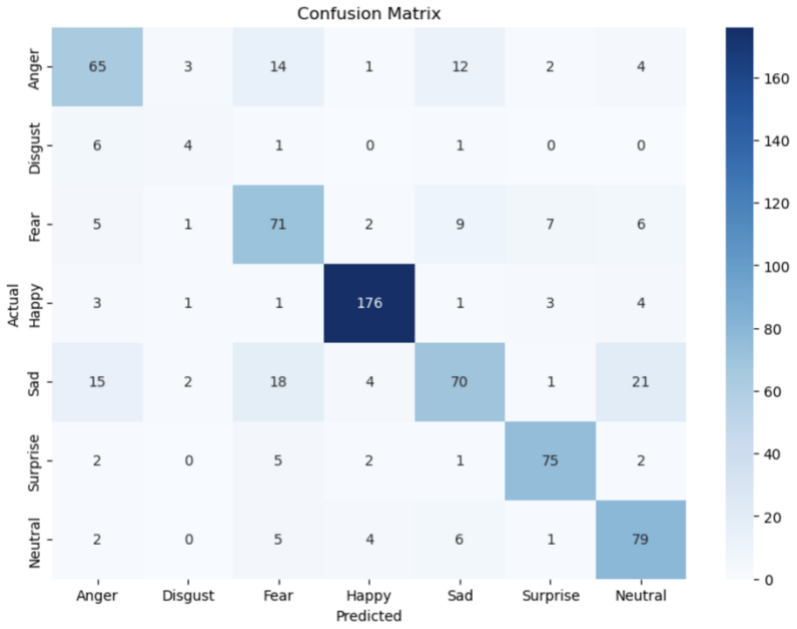


Figure 4.15: Confusion matrix after confidence score adjustment

Upon the application of adjusted confidence scores, notable improvements were observed in the accuracy of several emotion categories. For instance, the “Fear” class witnessed 71 accurate classifications, while “Neutral” achieved 79 true positives.

Furthermore, there was a commendable decrease in the misclassification rate of “Sad” instances as “Neutral”, plummeting from a substantial 80 to a mere 21. Yet, a persisting challenge remained in the “Disgust” class, which, even after the confidence adjustment, could only correctly classify 4 out of the 12 instances.

Post-adjustment, the model exhibited a discernible decline in misclassifications across multiple categories. This observation reinforces the notion that the confidence adjustments have indeed augmented the predictive accuracy of model.

A recurring theme in both matrices was the struggle of model with accurately classifying “Disgust”. This consistent challenge suggests a potential demand for enriching the dataset pertaining to this emotion or re-evaluating the approach employed for its identification.

The revised confidence scores appear to have rendered the model more prudent in its classifications. This led to a reduction in the number of instances classified. However, this circumspection resulted in heightened precision.

The modifications in confidence scores seemingly bolstered the competence of model in myriad aspects. Nonetheless, the consistent difficulty in identifying emotions such as “Disgust” signifies potential avenues for future enhancements.

### 4.3.3 Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve is a vital performance metric for classification problems, showcasing the trade-off between the true positive rate and false positive rate. The Area Under the ROC Curve (AUC) indicates the ability of model to discriminate between the positive and negative classes—A higher AUC suggests a better model.

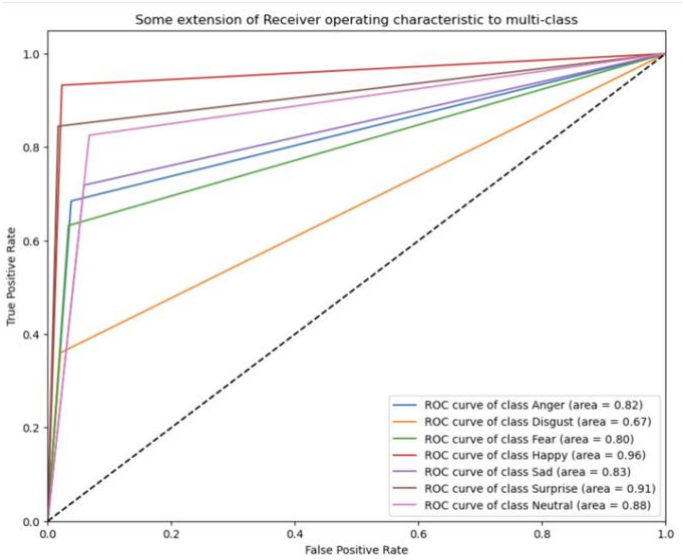


Figure 4.16: ROC Curve with a confidence score 1.0

In Figure 4.16, “Happy” emotion stands out with an impressive AUC (Area Under the Curve) of 0.96, underscoring the exemplary capability to discern true positives from negatives within this category. Meanwhile, emotions such as “Fear”, “Sad”, “Surprise”, and “Neutral” exhibit AUC values surpassing 0.80, reflecting their substantial discriminatory efficacy within the model. However, the “Disgust” class presents a challenge, registering an AUC of 0.67. While this score is within the acceptable range, it

remains notably inferior compared to the other classes, resonating with the insights gleaned from our previous confusion matrix evaluation.

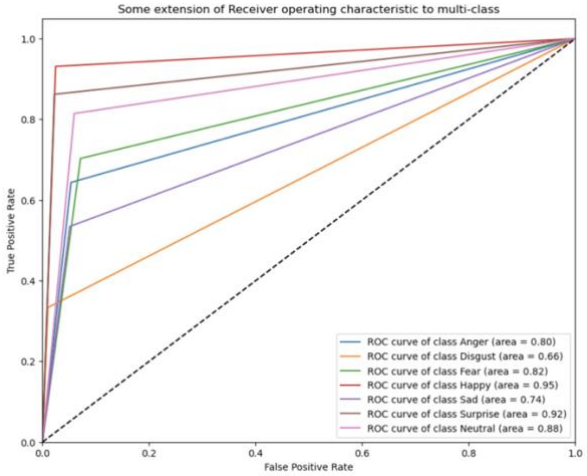


Figure 4.17: ROC curve after confidence score adjustment

Followed the adjustment in confidence scores as shown in Figure 4.17, there were discernible, albeit minor, shifts in the AUC (Area Under the Curve) values of most emotions. For instance, the AUC for “Fear” witnessed an enhancement, climbing from 0.80 to 0.82. Contrarily, “Sad” experienced a decline, receding from 0.83 to 0.74.

Remarkably, the emotions displayed a robust performance: “Happy” managed to sustain its commendable AUC, registering only a slight dip to 0.95. Meanwhile, “Surprise” showed a promising upward trend, elevating from 0.91 to 0.92.

However, challenges persisted with the “Disgust” class. Its AUC value witnessed a marginal decrement, settling at 0.66, which reiterated the inherent complexities associated with effectively classifying this particular emotion.

One key observation from these modifications is the overall stability in the discriminatory capacity of model. The subtle oscillations in AUC values insinuate that, despite adjustments, the model largely retained its competence in differentiating between various emotion classes.

Moreover, evident improvements in certain emotion categories, notably 'Surprise', illuminate the potential advantages of tweaking confidence scores, at least for specific emotions.

As an analysis grounded on the ROC curve offers profound insights into the prowess

of model in emotion classification. While it exhibits admirable discriminatory capabilities for a majority of emotions, there are undeniable challenges, especially with categories like 'Disgust', signaling areas ripe for further investigation and enhancement.

Model convergence and the subsequent stability of its accuracy is crucial for understanding its learning capability and generalization power. Here, we analyze the accuracy trends of our model under two different setups: with a fixed confidence score of one and after confidence score adjustments.

### 4.3.4 Accuracy

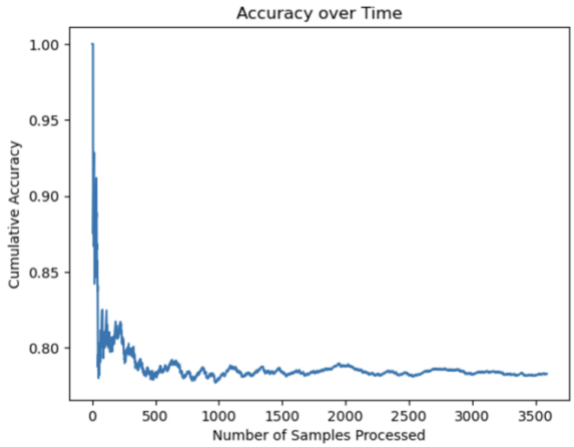


Figure 4.18: Accuracy with a confidence score 1.0

In Figure 4.18, initially, the model exhibited an great performance, achieved an accuracy of 0.85 to 1.00, which is indicative of flawless predictions.. However, followed this peak, the accuracy experienced a pronounced decline, eventually plateauing at 0.78. Rather than stabilizing at this juncture, the accuracy exhibited periodic fluctuations. This indicates that the model has entered a stable range of prediction probabilities. Based on the amplitude of the oscillation, the model is very stable in terms of prediction.

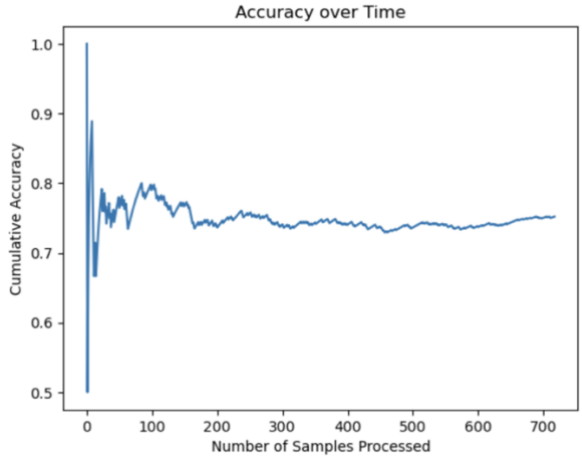


Figure 4.19: Accuracy after the confidence score adjustment

Figure 4.19 shows that in the adjusted setup, the behavior mirrored that of its predecessor: launching with an impeccable accuracy of 1.00. However, this zenith was followed by a descent, culminating in a slightly more conservative accuracy of 0.75—a modest decrement relative to the 0.78 observed in the fixed confidence setup. The undulating nature of accuracy persisted even after this descent, reaffirming the oscillatory patterns seen in the prior configuration. Despite adjustments in confidence scores, the inherent oscillatory trait remained largely unaffected. This similarity between the two configurations suggests that the confidence score modifications don't radically reshape the overarching behavior of model. A notable difference, albeit slight, was the settling accuracy post-adjustment, which could hint at the increased caution or ambiguity of model in its predictions.

In a broader perspective, understanding a prowess of model in deciphering facial emotions from videos is indispensable for a spectrum of applications, including sentiment analytics and enhanced human-machine interfacing.

### 4.3.5 Video Comparison

For ease of comparison, the predictions of original mini-Xception model are denoted by using a red bounding box and red font on the right side. The ensemble prediction of models, post integration, are indicated using a white bounding box with red font above. The white box depicts the probability of detecting a face, and the red box depicts the probability of recognizing emotions using ensemble learning with mini-Xception.

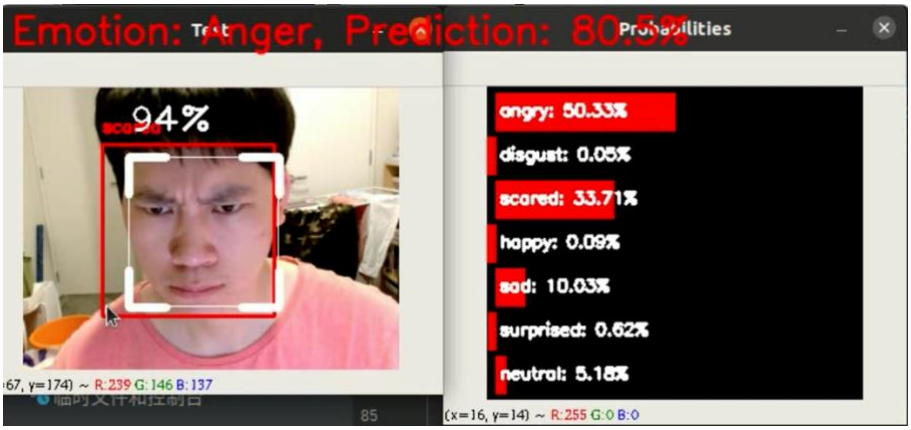


Figure 4.20: Sample frame – "Anger" expression

In a sample frame, Figure 4.20 shows a human expression of emotion "Anger", the original mini-Xception model yielded a prediction confidence of 50.33% for "Anger" and 33.71% for "Scared/Fear". In contrast, the ensemble version of Mini Xception exhibited a heightened confidence in its prediction, attributing an 80.5% likelihood to the "Anger" expression. This comparison underscores the ensemble learning model enhanced capability to discern the depicted emotion with greater precision.

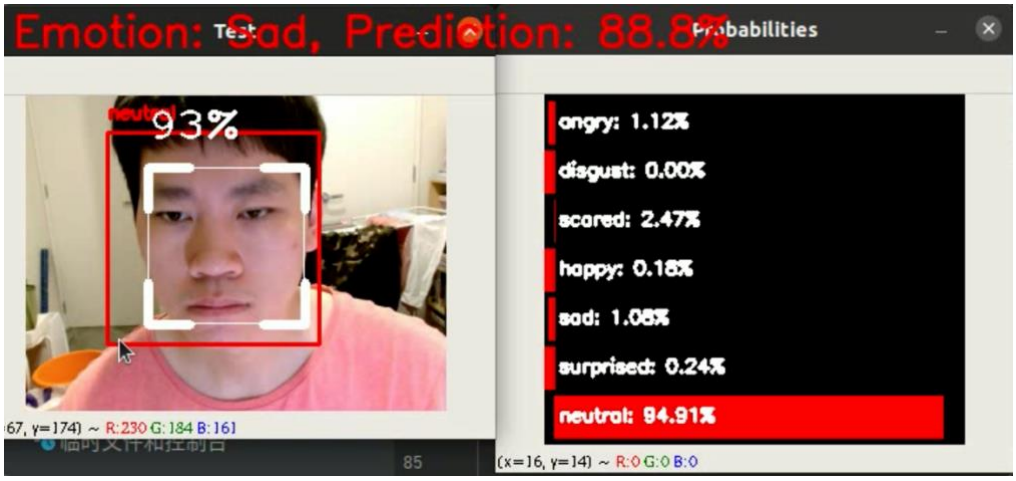


Figure 4.21: Sample Frame – "Sad" Expression

For a sample frame capturing a human expression of emotion "Sad" as shown in Figure 4.21, the predictions from the two models diverged notably. The original mini-Xception model assigned a 94.91% probability for "Neutral" and a 2.47% likelihood for "Scared". On the other hand, the ensemble variant of Mini Xception delivered a markedly elevated confidence, attributing an 88.8% probability solely to the 'Sad' emotion.

The enhanced confidence observed in the ensemble learning with mini-Xception model can be pivotal, especially in applications necessitating unequivocal and definitive

outcomes. Notably, the ensemble approach mitigates the ambiguity inherent in the predictions of the original model. For instance, whereas the original model demonstrated split probabilities between “Anger” and “Scared/Fear” for the “Anger” frame, the ensemble learning predictions almost exclusively favored the human emotion “Anger”.

Furthermore, the primary predictions of ensemble learning models consistently align more accurately with the ground truth, reflecting its superior emotion recognition capabilities. The disparity in the two models' predictions suggests that, by leveraging ensemble strategies, the mini-Xception model might achieve improved generalization, curtailing false positives and heightening true recognition rates.

The integration of ensemble learning methodologies into the mini-Xception framework seems to yield advancements in video-based emotion recognition, based on the increased prediction confidence and alignment of ensemble learning with ground truth. Extending its utility beyond static images, the system handles video data. Each frame in the video undergoes face detection, emotion prediction, and a resultant annotation that showcases the details of the prediction.

### 4.3.6 Multiclass Models Augmented With Data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Anger | 0.62 | 0.65 | 0.63 | 491 |
| Disgust | 0.95 | 0.65 | 0.77 | 55 |
| Fear | 0.59 | 0.45 | 0.51 | 528 |
| Happy | 0.87 | 0.90 | 0.88 | 879 |
| Sad | 0.56 | 0.60 | 0.58 | 594 |
| Surprise | 0.81 | 0.78 | 0.80 | 416 |
| Neutral | 0.66 | 0.73 | 0.69 | 626 |
| accuracy | | | 0.70 | 3589 |
| macro avg | 0.72 | 0.68 | 0.70 | 3589 |
| weighted avg | 0.70 | 0.70 | 0.70 | 3589 |

Figure 4.22: Result of mini-Xception with data augmentation

In this part, we contrast the efficacies of two modelling paradigms for emotion classification: ensemble learning and data augmentation. Delving into their respective performance metrics uncovers subtle variations in their proficiency across distinct

emotional categories.

In terms of performance, Figure 4.9 indicates that the ensemble learning method registered an overall accuracy rate of 78%. According to Figure 4.22, the data augmentation technique yielded a slightly reduced accuracy of 70%. Delving deeper into category-specific performances, the ensemble accuracy of model displayed considerable variance, ranging from as low as 20% to as high as 93%. In contrast, the data augmentation method maintained a tighter performance range, with accuracies oscillating between 56% and 95%. From an F1 score perspective, the ensemble values of model spanned from 0.26 to 0.93, whereas the data augmentation model exhibited scores ranging from 0.51 to 0.88.

Our discussion pivots on the nuances of these findings. The ensemble model, while showcasing superior overall accuracy and impressive metrics in most emotion categories, faltered significantly when classifying the 'Disgust' emotion. This dip may be attributable to inadequate representation of this emotion in the dataset or perhaps some inherent shortcomings in feature extraction. In comparison, the data augmentation model, although trailing in overall accuracy, demonstrated a more even-handed performance across categories, eschewing any severe performance drops.

While the ensemble learning exhibits a definitive edge in broader accuracy metrics, its struggles in accurately classifying human emotions, notably “Disgust”, necessitate further investigation. Conversely, the consistent performance of the data augmentation method across varying emotions suggests it might be better suited for specific applications demanding uniform accuracy.

### 4.3.7 Accuracy Rates of Various Emotion Recognition Models

Human emotion recognition, over the years, has witnessed significant advancements with a myriad of models being proposed. In this section, we systematically compare our proposed method, the “mini-Xception + ensemble learning,” with prior works to provide a comprehensive understanding of its performance standing within the domain.

In Table 4.1, it's evident that our approach, the mini-Xception + Ensemble learning,

achieves an accuracy of 78.20%, surpassing all other listed methods. It indicates a progression in the domain, where each subsequent model, in general, has been pushing the accuracy benchmarks further. Notably, earlier models like CNN proposed by Liu et al. in 2016 reported an accuracy of just over 62%, while more recent endeavors, like the VGG-19 model by Vignesh et al. in 2023, achieved an impressive 75.97%. Our method builds upon these previous endeavors, refining and incorporating ensemble learning methods to achieve the higher performance.

Table 4.1: Accuracy rates of various emotion recognition models

| Method | Accuracy Rate |
|---|---|
| CNN (Liu et al., 2016) | 62.44% |
| GoogleNet (Giannopoulos et al., 2017) | 65.20% |
| VGG+SVM (Georgescu et al., 2019) | 66.31% |
| Conv + Inception layer (Mollahosseini et al., 2016) | 66.40% |
| Attentional ConvNet (Minaee & Abdolrashidi, 2019) | 70.02% |
| ARM (ResNet-18)(Shi & Zhu, 2021) | 71.38% |
| Inception (Pramerdorfer & Kampel, 2016) | 71.60% |
| ResNet (Pramerdorfer & Kampel, 2016) | 72.40% |
| VGG (Pramerdorfer & Kampel, 2016) | 72.70% |
| LHC-Net(Pecoraro et al., 2022) | 74.42% |
| VGG-19(Vignesh et al., 2023) | 75.97% |
| Mini-Xception-Ensemble learning（This work） | 78.20% |

However, while accuracy is an essential metric, it's crucial to understand that each model might have its unique strengths, advantages, and limitations. The factors like computational efficiency, model complexity, and generalizability also play pivotal roles in determining the practicality and applicability of these models in real-world scenarios.

## 4.4 Limitations of this Thesis

In advancing emotion recognition through the integration of ensemble learning into the mini-Xception framework, we observed marked enhancements in terms of both accuracy and prediction confidence. Yet, like all empirical investigations, our research is not without its limitations. Here we outline the principal constraints of our study:

Our primary data source was the FER2013 dataset. While it offers a comprehensive

collection of facial expressions, potential biases, noise, or imbalances in specific emotional categories may exist. Such shortcomings can influence the ability of model to generalize effectively across a myriad of real-world situations.

While the mini-Xception model offers computational efficiency, its relatively simpler architecture might not encapsulate all subtleties of facial expressions as proficiently as more intricate models. Furthermore, our ensemble approach, which amalgamates a number of “expert” models, might compromise on real-time processing capabilities due to increased computational demands.

Our ensemble framework fundamentally rests on binary classifiers, dedicated to distinguishing between two emotions. This binary emphasis might not encapsulate the multifaceted nature of human emotions, especially in scenarios where emotions blur boundaries.

Intrinsic subjectivity characterizes emotion recognition, and a plethora of external elements—ranging from lighting conditions to cultural nuances—can sway it. Relying predominantly on a single dataset means our model may not be fully equipped to address such diverse variances.

The FER2013 dataset, much like its contemporaries, may fall short in capturing global variations in facial expressions. Consequently, our model might harbor biases towards specific demographics, affecting its dependability across varied cultural or regional backdrops.

For assessment, we leaned on traditional metrics such as accuracy, F1 scores, and ROC curves. However, these metrics might not holistically represent the efficacy of model, especially under dynamic real-world conditions.

Lastly, while the adaptations to the mini-Xception model have proven beneficial within the confines of our dataset, their effectiveness might wane when applied to divergent datasets or distinct applications.

# Chapter 5 Analysis and Discussions

*In Chapter 5, we delve into a comprehensive analysis of our findings, reflecting on the broader implications of the employed models. Through informed discussions, we critically assess the significance of our results, the challenges faced, and the potential pathways for future research in emotion recognition.*

## 5.1 Analysis

In this section, we delve deep into the intricacies and nuances of the models listed in the comparative table, especially focusing on the top-performing models and our proposed mini-Xception + Ensemble learning approach. Through a detailed analysis, we aim to understand the factors that contributed to the performance improvements and the distinct characteristics that differentiate these models.

### 5.1.1 Model Complexity and Computational Efficiency

Early models like the CNN proposed by Liu et al., 2016 were characterized by relatively simpler architectures. As the field advanced, more intricate designs like GoogleNet and VGG were introduced. While these models often boasted improved accuracy, they also came with increased computational demands.

The ensemble learning approach we utilized in mini-Xception combines multiple models or experts to make final decisions. While this generally improves accuracy, it may also introduce additional computational overhead. However, the advantage of ensemble learning lies in its ability to harness diverse decision boundaries and mitigate individual model weaknesses.

### 5.1.2 Feature Learning and Representations

Deep neural networks, especially ones like ResNet and VGG, are known for their capacity to learn hierarchical features. The depth of these models often correlates with their ability to discern complex patterns in the data. Our mini-Xception + Ensemble learning model, by combining the strengths of multiple experts, potentially benefits from diverse feature representations, leading to its superior performance.

### 5.1.3 Generalizability and Robustness

A critical aspect of any model is its ability to generalize to unseen data. While accuracy on a benchmark dataset is vital, the performance of this proposed model is based on real-

world, diverse, and possibly noisy data is crucial. The ensemble approach, by virtue of its design, tends to offer better generalizability. By combining decisions from multiple models, it can often avoid overfitting to specific data.

### 5.1.4 Evolution of Architectural Choices

From the simple CNNs to more recent designs like the Attentional ConvNet, the evolution in architectural choices is evident. Innovations like attention mechanisms have allowed models to focus on more salient features. Our approach, though rooted in the Xception architecture, leverages ensemble methods to further refine predictions.

## 5.2 Discussion

In this section, we reflect upon the broader implications of our study, the nuances of the models we evaluated, and the potential pathways for future research in the domain of emotion recognition.

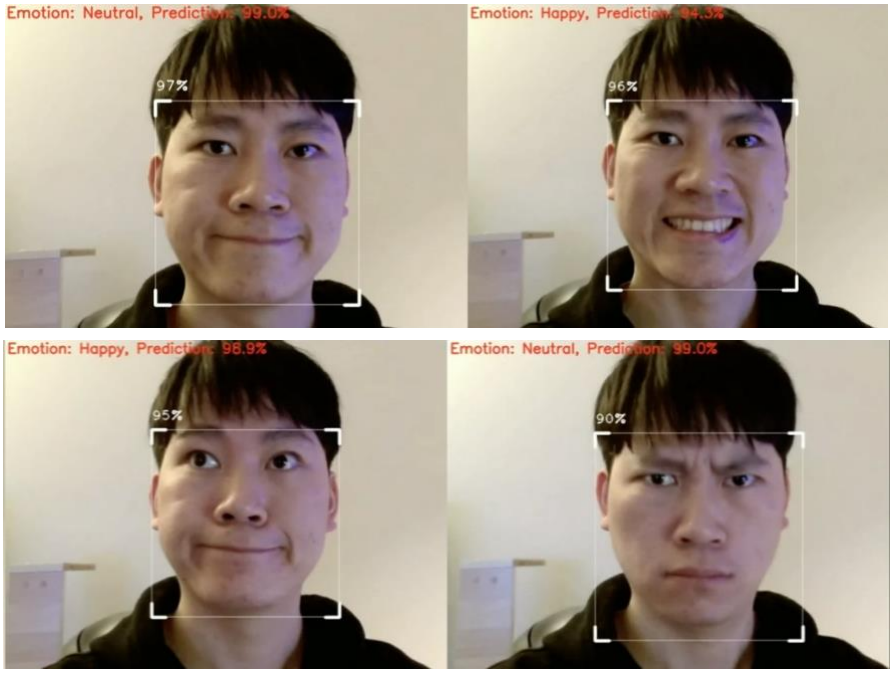


Figure 5.1: Subtle changes in facial features

We identified some limitations in our study methodology during our research. Models trained on databases with limited features tend to underperform in facial recognition, particularly in discerning fine details. Due to this disparity, it is difficult to perceive subtle changes in facial features. As shown in Figure 5.1, the model yields different results for

a pursed-lip smile compared to a toothy smile. However, enlarging the eyes in the video enables the model to recognize the expression as happiness. Sometimes, an angry expression is mistakenly identified as neutral. It demonstrates the importance of using more detailed datasets. Indirectly, it demonstrates the importance of confidence score tuning in practical applications, suggesting adjustments are needed to enhance real-world performance.

### 5.2.1 Relevance of Ensemble Learning

Our study affirms the potency of ensemble methods in the domain of deep learning. By combining the strengths of various experts or models, ensemble methods often transcend the performance of individual models. This is particularly relevant for applications like emotion recognition, where the difference in a few percentage points can drastically affect real-world applications, from human-computer interaction to healthcare diagnostics.

### 5.2.2 Model Interpretability

Though our mini-Xception-Ensemble learning model achieved higher accuracy, it is important to balance the trade-off between model complexity and interpretability. Deep learning models, due to their inherent complexity, often act as black boxes, making it challenging to discern their decision-making processes. Combining multiple models further complicates this. Our future work could focus on introducing more transparent ensemble methods or tools that aid in understanding model decisions better.

### 5.2.3 Real-World Applicability

While benchmark datasets provide a standardized measure of model performance, the true test of an efficacy of model lies in its performance in real-world scenarios. The factors such as lighting conditions, diverse facial expressions across cultures, and subtle emotion cues play a significant role in such environments. In spite of the fact that our model has not yet been extensively tested for its ability to adapt to these conditions, this is a challenge we will need to overcome in the future.

### 5.2.4 Future Directions

The continuous development of emotion recognition models suggests a wide range of possibilities for innovation. The development of hybrid models that combine traditional machine learning with deep learning, attention mechanisms that focus on critical emotion cues, and domain adaptation methods for cross-cultural emotion recognition are some of the avenues worth exploring. At the same time, we are also aware of the fact that hybrid models that are both lightweight and accurate merit further examination and investigation.

# Chapter 6 Conclusion and Future Work

*Chapter 6 provides a concise summary of the key findings and contributions of the thesis on facial emotion recognition. It concludes by highlighting the benefits of the integrated model and data enhancement methods, and suggests avenues for future research, including the exploration of novel models, the investigation of multimodal emotion recognition, and the consideration of long-term model adaptability in evolving societal contexts.*

## 6.1 Conclusion

Over the course of this study, we explored ensemble learning methods to achieve improved emotion recognition. Our ensemble approach using mini-Xception model clearly outperformed several state-of-the-art methods, underscoring the potential of combining multiple expert models for enhanced performance.

Our experimental results and comprehensive analysis highlighted the strengths and limitations of the approach, providing insights for the broader research community. Although ensemble models show promise, individual model selection and data quality play an important role in determining the final results, according to rigorous evaluations.

## 6.2 Future Work

In this thesis, we identify potential areas for future research and development in the field of facial emotion recognition, aiming to further enhance the proposed model and explore alternative integration methods. In our approach, we use specialized binary models for facial emotion recognition and combine this with parallel learning strategies. During the learning of new tasks, we 'freeze' certain parts of the model. This freezing keeps the knowledge the model already has while allowing it to learn new information. It helps the model to not forget old information while it learns something new. This method is efficient because the model builds on what it already knows without starting from scratch every time. This way, the model stays robust and can adapt to new tasks without losing its previous knowledge.

In our future work, we plan to extend the integrated model by adding more expert models. These will help us capture a wider variety of facial cues. We also want to explore combining other modalities like voice and body language. This would give us a fuller understanding of emotions. By using these extensions and strategic learning methods, we can improve the model's performance and its ability to detect subtle emotional expressions. As a result of this multifaceted approach, catastrophic forgetting will be overcome and human emotions will be more accurately understood.

# References


Agnihotri, D., Verma, K., Tripathi, P., & Singh, B. K. (2018). Soft voting technique to improve the performance of global filter based feature selection in text corpus. Applied Intelligence, 49(4), 1597–1619. https://doi.org/10.1007/s10489-018-1349-1

Ahonen, T., Hadid, A., & Pietikäinen, M. (2006). Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12), 2037–2041. https://doi.org/10.1109/tpami.2006.244

Ai, X., Sheng, V. S., Fang, W., Ling, C. X., & Li, C. (2020). Ensemble learning with Attention-Integrated Convolutional Recurrent Neural Network for imbalanced speech emotion recognition. IEEE Access, 8, 199909–199919. https://doi.org/10.1109/access.2020.3035910

Ali, G., Ali, A., Ali, F., Draz, U., Majeed, F., Yasin, S., Ali, T., & Haider, N. (2020). Artificial neural network based ensemble approach for multicultural facial expressions analysis. IEEE Access, 8, 134950–134963. https://doi.org/10.1109/access.2020.3009908

Alrefai, N., Ibrahim, O., Shehzad, H. M. F., Altigani, A., Abu-Ulbeh, W., Alzaqebah, M., & Alsmadi, M. K. (2022). An integrated framework based deep learning for cancer classification using microarray datasets. Journal of Ambient Intelligence and Humanized Computing, 14(3), 2249–2260. https://doi.org/10.1007/s12652-022-04482-9

Annappa, B. (2022). A comprehensive review of facial expression recognition methods. Multimedia Systems, 29(1), 73–103. https://doi.org/10.1007/s00530-022-00984-w

Arriaga, O., Valdenegro-Toro, M., & Plöger, P. G. (2017). Real-time convolutional neural networks for emotion and gender classification. The European Symposium on Artificial Neural Networks.

Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., & Wang, H. (2020). Ensemble neural network approach detecting pain intensity from facial expressions. Artificial Intelligence in Medicine, 109, 101954. https://doi.org/10.1016/j.artmed.2020.101954

Bellamkonda, S., Gopalan, N. P., Mala, C., & Settipalli, L. (2022). Facial expression recognition on partially occluded faces using component-based ensemble stacked CNN. Cognitive Neurodynamics.

https://doi.org/10.1007/s11571-022-09879-y

Ben, X., Jia, X., Yan, R., Zhang, X., & Meng, W. (2018). Learning effective binary descriptors for micro-expression recognition transferred by macro-information. Pattern Recognition Letters, 107, 50–58. https://doi.org/10.1016/j.patrec.2017.07.010

Chang, W., Schmelzer, M., Kopp, F., Hsu, C., Su, J., Chen, L., & Chen, M. (2019). A deep learning facial expression recognition based scoring system for restaurants. International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (2019).

https://doi.org/10.1109/icaiic.2019.8668998

Chirra, V. R. R., Reddy, U. S., & Kolli, V. K. K. (2021). Virtual facial expression recognition using deep CNN with ensemble learning. Journal of Ambient Intelligence and Humanized Computing, 12(12), 10581–10599. https://doi.org/10.1007/s12652-020-02866-3

Connie, T., Al-Shabi, M., Cheah, W. P., & Goh, M. K. O. (2017). Facial expression recognition using a hybrid CNN–SIFT aggregator. In Lecture Notes in Computer

Science (pp. 139–149).

https://doi.org/10.1007/978-3-319-69456-6_12

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. *International Journal of Digital Crime and Forensics* (IJDCF) 8 (1), 26-36.

DOI: 10.4018/IJDCF.2016010102

Cui, W. (2015) A Scheme of Human Face Recognition in Complex Environments. Master's Thesis, Auckland University of Technology, New Zealand.

https://hdl.handle.net/10292/7798

Dapogny, A., Bailly, K., & Dubuisson, S. (2017). Confidence-Weighted local expression predictions for occlusion handling in expression recognition and action unit detection. International Journal of Computer Vision, 126(2–4), 255–271. https://doi.org/10.1007/s11263-017-1010-1

Dhar, S., Vishwakarma, A., Ghanti, D., & Jana, N. D. (2022). Ensemble learning based plant leaf disease classification considering deep convolutional features from pre-trained CNN. IEEE Conference on Information and Communication Technology (CICT). https://doi.org/10.1109/cict56698.2022.9997819

Fan, Y., Lam, J. C., & Li, V. O. K. (2018). Multi-region ensemble convolutional neural network for facial expression recognition. In Lecture Notes in Computer Science (pp. 84–94). https://doi.org/10.1007/978-3-030-01418-6_9

Fayek, H. M., Lech, M., & Cavedon, L. (2016). Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. International Joint Conference on Neural Networks (IJCNN).

https://doi.org/10.1109/ijcnn.2016.7727250

Feng, Y., Pang, T., Li, M., & Guan, Y. (2020). Small sample face recognition based on ensemble deep learning. Chinese Control and Decision Conference (CCDC).

https://doi.org/10.1109/ccdc49329.2020.9163968

Gao, X., Nguyen, M., Yan, W. (2023) Enhancement of human face mask detection performance by using ensemble learning models. Pacific-Rim Conference on Image and Video Technology, New Zealand.

García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary Computation, 17(3), 275–306. https://doi.org/10.1162/evco.2009.17.3.275

Georgescu, M., Ionescu, R. T., & Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. IEEE Access, 7, 64827–64836. https://doi.org/10.1109/access.2019.2917266

Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2017). Deep learning approaches for facial emotion recognition: A case study on FER-2013. In Smart Innovation, Systems and Technologies (pp. 1–16).

https://doi.org/10.1007/978-3-319-66790-4_1

Gu, D., Nguyen, M., Yan, W. (2016) Cross models for twin recognition. *International Journal of Digital Crime and Forensics* 8 (4), 26-36.

DOI: 10.4018/IJDCF.2016100103

Habib, A. S. B., & Tasnim, T. (2020). An ensemble hard voting model for cardiovascular disease prediction. International Conference on Sustainable Technologies for Industry 4.0 (STI). https://doi.org/10.1109/sti50764.2020.9350514

Hans, A. S. A., & Rao, S. (2021). a CNN-LSTM based deep neural networks for facial emotion detection in videos. International Journal of Advances in Signal and Image Sciences, 7(1), 11–20. https://doi.org/10.29284/ijasis.7.1.2021.11-20

Hayashi, T., & Fujita, H. (2021). One-class ensemble classifier for data imbalance problems. Applied Intelligence, 52(15), 17073–17089. https://doi.org/10.1007/s10489-021-02671-1

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. http://export.arxiv.org/pdf/1704.04861

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8), 2011–2023. https://doi.org/10.1109/tpami.2019.2913372

Hussain, M., Qazi, E., Aboalsamh, H., & Ullah, I. (2023). Emotion recognition system based on two-level ensemble of deep-convolutional neural network models. IEEE Access, 11, 16875–16895.

https://doi.org/10.1109/access.2023.3245830

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. http://export.arxiv.org/pdf/1502.03167

Janet, B., Selvan, A. K., & Sivakumaran, N. (2022). Pre-trained ensemble model for identification of emotion during COVID-19 based on emergency response support system dataset. Applied Soft Computing, 122, 108842. https://doi.org/10.1016/j.asoc.2022.108842

Jia, C., Li, C. L., & Zhou, Y. (2020). Facial expression recognition based on the ensemble learning of CNNs. IEEE International Conference on Signal Processing. https://doi.org/10.1109/icspcc50002.2020.9259543

Jiang, H., Kim, B., Guan, M. Y., & Gupta, M. (2018). To trust or not to trust a classifier. https://arxiv.org/pdf/1805.11783.pdf

Ju, C., Bibaut, A. F., & Van Der Laan, M. J. (2018). The relative performance of ensemble methods with deep convolutional neural networks for image classification. Journal of Applied Statistics, 45(15), 2800–2818.

https://doi.org/10.1080/02664763.2018.1441383

Kang, J., & Gwak, J. (2021). Ensemble of multi-task deep convolutional neural networks using transfer learning for fruit freshness classification. Multimedia Tools and Applications, 81(16), 22355–22377. https://doi.org/10.1007/s11042-021-11282-4

Kang, P., & Cho, S. (2006). EUS SVMS: Ensemble of under-sampled SVMs for data imbalance problems. In Lecture Notes in Computer Science (pp. 837–846). https://doi.org/10.1007/11893028_93

Kanna, R. K., Surendhar, P. A., Rubi, J., Jyothi, G., Ambikapathy, A., & Vasuki, R. (2022). Human computer interface application for emotion detection using facial recognition. IEEE International Conference on Current Development in Engineering and Technology (CCET). https://doi.org/10.1109/ccet56606.2022.10080678

Khaireddin, & Chen. (2021). Facial emotion recognition: State of the art performance on FER2013. Computer Vision and Pattern Recognition (cs.CV), https://doi.org/10.48550/arXiv.2105.03588.

Kim, B., Dong, S., Roh, J., Kim, G., & Lee, S. Y. (2016). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/cvprw.2016.187

Kim, T., Yu, C., & Lee, S. (2018). Facial expression recognition using feature additive pooling and progressive fine‐tuning of CNN. Electronics Letters, 54(23), 1326‐1328. https://doi.org/10.1049/el.2018.6932

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report, Toronto University, Canada.

https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Krumhuber, E., Küster, D., Namba, S., Shah, D., & Calvo, M. G. (2021). Emotion recognition from posed and spontaneous dynamic expressions: Human observers versus machine analysis. Emotion, 21(2), 447–451. https://doi.org/10.1037/emo0000712

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., & Duin, R. P. W. (2003). Limits on the majority vote accuracy in classifier fusion. Pattern Analysis and Applications, 6(1), 22–31. https://doi.org/10.1007/s10044-002-0173-7

Kyerematen-Boateng, H., Josyula, D. P., & Conn, M. (2023). Computing confidence score for neural network predictions from latent features. International Conference on Control, Communication and Computing (ICCC). https://doi.org/10.1109/iccc57789.2023.10165294

Lajevardi, S. M., & Hussain, Z. M. (2010). Automatic facial expression recognition: feature extraction and selection. Signal, Image and Video Processing, 6(1), 159–169. https://doi.org/10.1007/s11760-010-0177-5

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, Y. S., & Bang, C. C. (2021). Framework for the classification of imbalanced structured data using under-sampling and convolutional neural network. Information Systems Frontiers, 24(6), 1795–1809. https://doi.org/10.1007/s10796-021-10195-9

Li, C., Li, D., Zhao, M., & Li, H. (2022). A light-weight convolutional neural network for facial expression recognition using Mini-Xception neural networks. IEEE International Conference on Current Development in Engineering and Technology (CCET). https://doi.org/10.1109/qrs-c57518.2022.00104

Li, D., Wen, G., Xu, L., & Cai, X. (2019). Graph-based dynamic ensemble pruning for facial expression recognition. Applied Intelligence, 49(9), 3188–3206.

https://doi.org/10.1007/s10489-019-01435-2

Li, H., Sun, J., Xu, Z., & Chen, L. (2017). Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network. IEEE Transactions on Multimedia, 19(12), 2816–2831. https://doi.org/10.1109/tmm.2017.2713408

Li, Y., Zeng, J., Shan, S., & Chen, X. (2019). Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 28(5), 2439–2450. https://doi.org/10.1109/tip.2018.2886767

Liang, X., Jiang, A., Li, T., Xue, Y., & Wang, G. (2020). LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM. Knowledge Based Systems, 196, 105845. https://doi.org/10.1016/j.knosys.2020.105845

Lin, P., & Luo, X. (2020). A survey of sentiment analysis based on machine learning. In Lecture Notes in Computer Science (pp. 372–387). https://doi.org/10.1007/978-3-030-60450-9_30

Liu, K., Zhang, M., & Pan, Z. (2016). Facial Expression Recognition with CNN Ensemble. International Conference on Cyberworlds. https://doi.org/10.1109/cw.2016.34

Liu, R., & Cocea, M. (2018). Nature-inspired framework of ensemble learning for collaborative classification in granular computing context. Granular Computing, 4(4), 715–724. https://doi.org/10.1007/s41066-018-0122-5

Lopes, A. T., De Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognition, 61, 610–628. https://doi.org/10.1016/j.patcog.2016.07.026

Lopez-Gil, J. M., & Garay-Vitoria, N. (2021). Photogram classification-based emotion recognition. IEEE Access, 9, 136974–136984. https://doi.org/10.1109/access.2021.3117253

Mehanović, D., Mašetić, Z., & Kečo, D. (2019). Prediction of heart diseases using majority voting ensemble method. In IFMBE proceedings (pp. 491–498). https://doi.org/10.1007/978-3-030-17971-7_73

Merghani, W., Davison, A. K., & Yap, M. H. (2018). Facial micro-expressions grand challenge 2018: Evaluating spatio-temporal features for classification of objective classes. IEEE International Conference on Automatic Face &Amp. https://doi.org/10.1109/fg.2018.00104

Minaee, S., & Abdolrashidi, A. (2019). Deep-Emotion: Facial expression recognition using attentional convolutional network. arXiv (Cornell University). https://arxiv.org/pdf/1902.01019

Mollahosseini, A., Chan, D. M., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. IEEE Winter Conference on Applications of Computer Vision. https://doi.org/10.1109/wacv.2016.7477450

Nguyen, M., Yan, W. (2022) Temporal color-coded facial-expression recognition using convolutional neural network. International Summit Smart City 360°: Science and Technologies for Smart Cities, pp 41–54.

https://doi.org/10.1007/978-3-031-06371-8_4

Nguyen, M., Yan, W. (2023) From faces to traffic lights: A multi-scale approach for emotional state representation. IEEE International Conference on Smart City.

Nugrahaeni, R. A., & Mutijarsa, K. (2016). Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification. International Seminar on Application for Technology of Information and Communication (ISemantic). https://doi.org/10.1109/isemantic.2016.7873831

Pantic, M., & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: the state of the art. IEEE Transactions on Pattern Analysis and Machine

Intelligence, 22(12), 1424–1445. https://doi.org/10.1109/34.895976

Pecoraro, R., Basile, V., & Bono, V. (2022). Local multi-head channel self-attention for facial expression recognition. Information, 13(9), 419. https://doi.org/10.3390/info13090419

Powers, D. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2010.16061

Prabha, K., Nataraj, B., Ajaydevan, R., Kabilan, S., & Muthuselvam, V. (2022). Real time facial emotion recognition methods using different machine learning methods . International Conference on Smart Electronics and Communication (ICOSEC). https://doi.org/10.1109/icosec54921.2022.9952133

Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the Art. arXiv (Cornell University). https://arxiv.org/pdf/1612.02903.pdf

Radhakrishna, A., Yan, W., Kankanhalli, M. (2006) Modeling intent for home video repurposing. IEEE MultiMedia 13 (1), 46-55. DOI: 10.1109/MMUL.2006.12

Rathour, N., Khanam, Z., Gehlot, A., Singh, R., Rashid, M., AlGhamdi, A. S., & Alshamrani, S. S. (2021). Real-time facial emotion recognition framework for employees of organizations using Raspberry-PI. Applied Sciences, 11(22), 10540. https://doi.org/10.3390/app112210540

Rivenski, A. (2022) Human Facial Emotion Recognition From Digital Images Using Deep Learning. Master's Thesis, Auckland University of Technology, New Zealand. https://hdl.handle.net/10292/15679

Renda, A., Barsacchi, M., Bechini, A., & Marcelloni, F. (2019). Comparing ensemble strategies for deep learning: An application to facial expression recognition. Expert Systems with Applications, 136, 1–11.

https://doi.org/10.1016/j.eswa.2019.06.025

Salama, E. S., El-Khoribi, R. A., Shoman, M., & Shalaby, M. (2021). A 3D-convolutional neural network framework with ensemble learning methods for multi-modal emotion recognition. Egyptian Informatics Journal, 22(2), 167–176. https://doi.org/10.1016/j.eij.2020.07.005

Savargiv, M., Masoumi, B., & Keyvanpour, M. R. (2020). A new ensemble learning method based on learning automata. Journal of Ambient Intelligence and Humanized Computing, 13(7), 3467–3482. https://doi.org/10.1007/s12652-020-01882-7

Shehu, H. A., Browne, W. N., & Eisenbarth, H. (2020). Emotion categorization from video-frame images using a novel sequential voting technique. In Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-64559-5_49

Shi, J., & Zhu, S. (2021). Learning to amend facial expression representation via de-albino and affinity. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2103.10189

Sikkandar, H., & Thiyagarajan, R. (2020). Deep learning based facial expression recognition using improved cat swarm optimization. Journal of Ambient Intelligence and Humanized Computing, 12(2), 3037–3053. https://doi.org/10.1007/s12652-020-02463-4

Soma, S., & Waddenkery, N. (2022). Machine-learning object detection and recognition for Surveillance system using YOLOv3. International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). https://doi.org/10.1109/iceeict53079.2022.9768558

Song, C., He, L., Yan, W., Nand, P. (2019) An improved selective facial extraction model for age estimation. International Conference on Image and Vision Computing New Zealand. DOI: 10.1109/IVCNZ48456.2019.8960965

Sridhar, K., Lin, W., & Busso, C. (2021). Generative approach using soft-labels to learn uncertainty in predicting emotional attributes. International Conference on Affective Computing and Intelligent Interaction (ACII). https://doi.org/10.1109/acii52823.2021.9597461

Sun, L., Ge, C., & Zhong, Y. (2021). Design and implementation of face emotion recognition system based on CNN Mini-Xception Frameworks. Journal of Physics, 2010(1), 012123. https://doi.org/10.1088/1742-6596/2010/1/012123

Sun, P (2017) Facial Expression Classification Using R-CNN Based Methods. Master's Thesis. Auckland University of Technology, New Zealand.

https://hdl.handle.net/10292/12161

Tang, J., Su, Q., Su, B., Fong, S., Cao, W., & Gong, X. (2020). Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition. Computer Methods and Programs in Biomedicine, 197, 105622. https://doi.org/10.1016/j.cmpb.2020.105622

Tang, Y. (2013). Deep learning using linear support vector machines. arXiv (Cornell University). https://arxiv.org/pdf/1306.0239.pdf

Tümen, V., Soylemez, O. F., & Ergen, B. (2017). Facial emotion recognition on a dataset using convolutional neural network. International Artificial Intelligence and Data Processing Symposium (IDAP). https://doi.org/10.1109/idap.2017.8090281

Verma, S., Kumar, P., & Singh, J. (2023). A unified lightweight CNN-based model for disease detection and identification in corn, rice, and wheat. IETE Journal of Research, 1–12. https://doi.org/10.1080/03772063.2023.2181229

Vignesh, S., Savithadevi, M., Sridevi, M., & Ramaswamy, S. (2023). A novel facial emotion recognition model using segmentation VGG-19 architecture. International Journal of Information Technology, 15(4), 1777–1787. https://doi.org/10.1007/s41870-023-01184-z

Wang, H., Yan, W. (2022) Face detection and recognition from distance based on deep learning. Aiding Forensic Investigation Through Deep Learning and Machine Learning Framework. IGI Global.

DOI: 10.4018/978-1-6684-4558-7.ch006

Wang, X., Yan, W. (2020) Cross-view gait recognition through ensemble learning. Neural Computing and Applications, 32, 7275–7287

https://doi.org/10.1007/s00521-019-04256-z

Wang, Y., Li, M., Wan, X., Zhang, C., & Wang, Y. (2020). Multiparameter space decision voting and fusion features for facial expression recognition. Computational Intelligence and Neuroscience, 2020, 1–17. https://doi.org/10.1155/2020/8886872

Wang, Y., & Lu, F. (2021). An adaptive boosting algorithm based on weighted feature selection and category classification confidence. Applied Intelligence, 51(10), 6837–6858. https://doi.org/10.1007/s10489-020-02184-3

Webb, G. I., & Zheng, Z. (2004). Multistrategy ensemble learning: Reducing error by combining ensemble learning methods. IEEE Transactions on Knowledge and Data Engineering, 16(8), 980–991. https://doi.org/10.1109/tkde.2004.29

Wu, M., & Li, X. (2021). Unbalanced data classification algorithm based on hybrid sampling and ensemble learning. International Conference on Intelligent Systems and Knowledge Engineering (ISKE). https://doi.org/10.1109/iske54062.2021.9755369

Xu, G., Yan, W. (2024) Facial emotion recognition using ensemble learning. Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems. IGI Global.

Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N., & Han, X. (2021). A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data. Information Sciences, 572, 574–589.

https://doi.org/10.1016/j.ins.2021.02.056

Yadav, A. (2023). COVID-LiteNet: A lightweight CNN based network for COVID-19 detection using X-ray images. International Conference on Developments in eSystems Engineering (DeSE). https://doi.org/10.1109/dese58274.2023.10099799

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer Nature.

https://doi.org/10.1007/978-3-030-10713-0

Yan, W. (2023) Computational Methods for Deep Learning: Theory, Algorithms, and Implementations. Springer Nature.

https://doi.org/10.1007/978-981-99-4823-9

Ye, G., Xiong, X., Liu, Y., Li, X., & Li, Q. (2022). A novel speech emotion recognition method based on feature construction and ensemble learning. PLOS ONE, 17(8), e0267132. https://doi.org/10.1371/journal.pone.0267132

Yin, Z., Liu, L., Liu, L., Zhang, J., & Wang, Y. (2017). Dynamical recursive feature elimination technique for neurophysiological signal-based emotion recognition. Cognition, Technology & Work, 19(4), 667–685. https://doi.org/10.1007/s10111-017-0450-2

Younas, F., Usman, A., Yan, W. (2023) A deep ensemble learning method for colorectal polyp classification with optimized network parameters. Applied Intelligence, 53, pages 2410–2433. https://doi.org/10.1007/s10489-022-03689-9

Younas, F., Usman, M., Yan, W. (2023) A deep neural network ensemble framework for colorectal polyp classification. Multimedia Tools and Applications, 82 (12) pp 18925–18946. https://doi.org/10.1007/s11042-022-14177-0

Younis, E. M. G., Zaki, S. M., Kanjo, E., & Houssein, E. H. (2022). Evaluating ensemble learning methods for multi-modal emotion recognition using sensor data fusion. Sensors, 22(15), 5611. https://doi.org/10.3390/s22155611

Yu, S., Li, X., Wang, H., Zhang, X., & Chen, S. (2021). C_CART: An instance confidence-based decision tree algorithm for classification. Intelligent Data Analysis, 25(4), 929–948. https://doi.org/10.3233/ida-205361

Zehra, N., Azeem, S. H., & Farhan, M. (2021). Human activity recognition through ensemble learning of multiple convolutional neural networks. Annual Conference on Information Sciences and Systems (CISS). https://doi.org/10.1109/ciss50987.2021.9400290

Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning methods : A tutorial and review. Information Fusion, 59, 103–126. https://doi.org/10.1016/j.inffus.2020.01.011

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015). Learning social relation traits from face images. IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2015.414