**Abdul-Qadir Mohammed, EVALUATING THE IMPACT OF GENDER ON FACIAL EXPRESSION RECOGNITION**

**. (Under the direction of Dr. Mustafa Atay) Department of Computer Science, Winston-Salem State University, October 2024**

**ABSTRACT**

Facial expression recognition (FER) has emerged as a pivotal technology with significant applications across various industries, including healthcare, technology, security, and media . The ability to automatically detect and categorize facial expressions and emotions has been greatly enhanced by advancements in machine learning and deep learning models. However, the presence of biased datasets and variations in facial image poses challenges in achieving consistent and accurate predictions. This research aims to evaluate the impact of gender on facial expression recognition using various deep learning models, with an objective of identifying the existence of gender bias, determining which algorithms are most effective in mitigating such biases, and understanding which facial expressions are more susceptible to gender bias.

The study employs five deep learning algorithms which are Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3, selected for their benchmarking performance in facial expression recognition. These algorithms are trained and tested on male-only, female-only, and mixed datasets, with fine-tuned hyperparameters to optimize performance. The experimentation is conducted in three phases: training, testing, and performance assessment. During the training phase, each algorithm is trained using the specified datasets, while the testing phase involves evaluating the algorithms on separate validation sets to assess bias and accuracy. The final phase employs statistical measures such as accuracy, precision, recall, and F-1 score to assess performance, alongside bias metrics like accuracy difference, F1-score difference and disparate impact.

The research problem is centered around evaluating the impact of gender on facial expression recognition across various deep learning models. The study seeks to determine if gender bias exists, which algorithms mitigate gender bias the most, and which facial expressions are more susceptible to gender bias. The methodology involves extensive studying, training, and testing of multiple deep learning algorithms to make comparisons and evaluate the level of bias among various algorithms. By utilizing mixed, male-only, and female-only datasets, the research aims to provide insights into the demographic effects on facial emotion recognition and the impact of gender on fairness and recognition of expressions.

The significance of this research lies in its potential to address the challenges of gender bias in facial expression recognition and contribute to the development of fairer and more accurate FER systems. By exploring biases in facial expression analysis, the study aims to provide a comprehensive understanding of the demographic effects on facial emotion expression and the impact of gender on fairness and recognition of expressions. The findings of this research are expected to have implications for the design and implementation of facial expression recognition systems, particularly in terms of ensuring fairness and accuracy across different demographic groups.

**EVALUATING THE IMPACT OF GENDER ON FACIAL EXPRESSION RECOGNITION**

**A thesis presented to the graduate faculty of the Department of Computer Science Winston-Salem State University In partial fulfillment of the requirements for the degree of MASTERS OF SCIENCE**

**By:**

**Abdul-Qadir Mohammed**

**December 2024**

**EVALUATING THE IMPACT OF GENDER ON FACIAL EXPRESSION RECOGNITION**

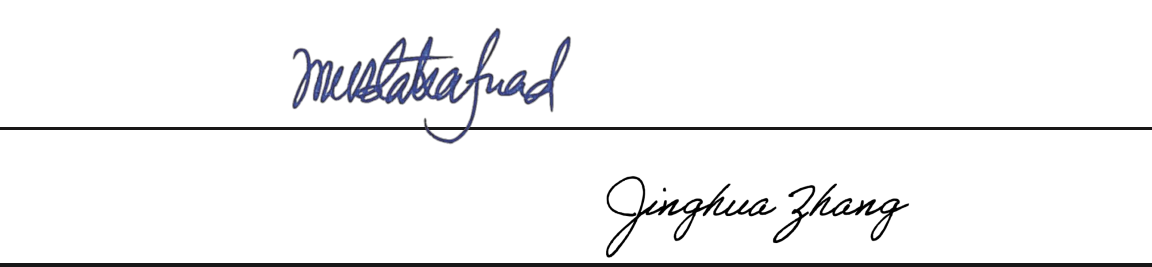
**By:**

**Abdul Mohammed**

**Computer Science and Information Technology**



**COMMITTEE APPROVALS:**



**Thesis Advisor:** Mustafa Atay, Ph.D.

**Committee Member:** Muztaba Fuad, Ph.D.



**Committee Member:** Jinghua Zhang, Ph.D.



**DEPARTMENT APPROVAL:**



**Department Chair:** Muztaba Fuad , Ph.D.

**Dedication**

I wish to dedicate this thesis to my family, whose unwavering support has been invaluable throughout this journey. I am especially grateful to my parents, Tina and Abdul Mohammed, who made the profound decision to leave their home country to provide my brothers and me with the opportunity to grow and learn in America. Their sacrifices have been the cornerstone of my achievement in realizing my goal of becoming a Software Engineer. Their courage, resilience, and dedication have been a constant source of inspiration, instilling in me the values of hard work and perseverance. I am forever indebted to them for their love and encouragement, which have guided me through every challenge and success.

In addition to my parents, I extend my heartfelt gratitude to my brothers, Adam and Hassan Mohammed, for their unwavering encouragement and guidance throughout this process. My elder brother, Hassan, has been a profound influence, consistently urging me to pursue a career in technology and inspiring my interest in computing from the age of ten. My younger brother, Adam, with his exceptionally inquisitive nature, continually challenges me to strive for excellence each day.

I would also like to dedicate this thesis to my fiancé, Kayla Dion, and her family. Kayla was a constant source of strength and encouragement, providing unwavering support and understanding throughout this journey. Her belief in my abilities has been instrumental in helping me overcome challenges and maintain focus on my goals. Additionally, I am deeply grateful to her parents, Chris and Christina Dion, for their kindness and encouragement. Their support has been invaluable, offering both practical

assistance and emotional reassurance when it was most needed. I would also like to extend my heartfelt thanks to her grandmother, Donna Tibbetts, for her warmth and inspiration, which have greatly enriched my journey. Together, their influence has been a vital component of my success.

Finally, I extend my deepest gratitude to my son, Malik Mohammed, whose boundless curiosity and joy have been a continuous source of inspiration for me. His enthusiasm and passion for life remind me daily of the importance of perseverance and dedication. Malik's love and laughter have been a comforting presence, providing me with the strength and motivation to pursue and achieve my goals

**Acknowledgements**

I would like to express my deepest gratitude to my graduate research advisor, Professor Mustafa Atay, whose expertise in AI and machine learning has been instrumental in guiding my research journey. His insightful advice and unwavering support have been pivotal in shaping my understanding of complex concepts and have greatly contributed to my development as a software engineer. Professor Atay's dedication to fostering a rigorous research environment has inspired me to pursue excellence and innovation in my work.

I am also profoundly thankful to my undergraduate research advisor, Professor Zhang, who introduced me to the world of developing educational games. Her mentorship during my undergraduate studies laid the foundation for my passion for software engineering. Professor Zhang's encouragement and guidance in exploring the intersection of technology and education have been invaluable in my academic and professional growth.

Together, the mentorship and support of Professors Atay and Zhang have been crucial in helping me achieve my goal of becoming a software engineer through research. Their influence has not only enriched my academic journey but also instilled in me a lifelong commitment to learning and discovery. I would also like to extend thanks to NSF for providing funding needed for this research under research grant NSF#1900087.

i

**Table of Contents**

[List of Figures](#page11) [iii](#page11)

[List of Tables](#page12) [iv](#page12)

[CHAPTER 1: INTRODUCTION](#page13) [1](#page13)

[1.1 Problem Statement](#page14) [2](#page14)

[1.2 CONTRIBUTIONS](#page15) [3](#page15)

[CHAPTER 2: BACKGROUND AND LITERATURE REVIEW](#page18) [6](#page18)

[2.1 Demographic Bias in Facial Expression Recognition](#page19) [7](#page19)

[2.2 Gender Bias in Facial Expression Recognition](#page20) [8](#page20)

[2.3 Selected Deep Learning (DL) Algorithms](#page21) [9](#page21)

[2.4 Facial Expression Image Database Used for Demographic Bias](#page26) [14](#page26)

[CHAPTER 3: SYSTEM DESIGN & DEVELOPMENT](#page29) [17](#page29)

[3.1 System Architecture](#page29) [17](#page29)

[3.2 Evaluation Plan](#page31) [19](#page31)

[3.3 System Software](#page33) [21](#page33)

[CHAPTER 4: EXPERIMENTAL STUDY](#page36) [24](#page36)

ii

[4.1 Experimental Design](#page37) [25](#page37)

[4.1.1 Methodology](#page37) [25](#page37)

[4.1.2 Preparing and Preprocessing Data](#page41) [29](#page41)

[4.1.3 Experiment Setup](#page44) [32](#page44)

[4.2 Experimentation](#page46) [34](#page46)

[4.2.1 Gender Bias Across DL Algorithms](#page46) [34](#page46)

[4.2.2 Gender Bias Across Balanced and Imbalanced Datasets](#page50) [38](#page50)

[4.2.3 Gender Bias Across Facial Emotions](#page57) [45](#page57)

[4.3 Discussion](#page68) [56](#page68)

[CHAPTER 5 CONCLUSION & FUTURE WORK](#page74) [62](#page74)

[Bibliography](#page78) [66](#page78)

iii

**List of Figures**

[Figure 1 Types of Deep Learning algorithms [18]](#page23) [11](#page23)

[Figure 2 Sequential Model Architechture [23]](#page24) [12](#page24)

[Figure 3 ResNet50 Model Architechture [24]](#page24) [12](#page24)

[Figure 4 DenseNet Model Architechture [25]](#page25) [13](#page25)

[Figure 5 MobileNetV2 Model Architecture [22]](#page25) [13](#page25)

[Figure 6 InceptionV3 Architechture [19]](#page26) [14](#page26)

[Figure 7 Images from KDEF dataset [16]](#page27) [15](#page27)

[Figure 8 Block diagram of the general FER model [26]](#page30) [18](#page30)

[Figure 9 Code snippet showing static training variables](#page37) [25](#page37)

[Figure 10 Anaconda Navigator User Interface [28]](#page38) [26](#page38)

[Figure 11 Code snippet of model training set up in Jupyter Lab](#page39) [27](#page39)

[Figure 12 Algorithmically splitting the datasets using Jupyter Notebook](#page43) [31](#page43)

[Figure 13 Data Configurations for training and validation scenarios](#page45) [33](#page45)

[Figure 14 Directories created for model training and validation](#page47) [35](#page47)

[Figure 15 Deep learning model F-1 delta evaluation](#page49) [37](#page49)

[Figure 16 KDEF dataset female subset being utilized for training and validation](#page51) [39](#page51)

[Figure 17 Fear F-1 Scores across validation datasets](#page59) [47](#page59)

[Figure 18 Surprise F-1 Scores across validation datasets](#page60) [48](#page60)

[Figure 19 Neutral F-1 Scores across validation datasets](#page62) [50](#page62)

[Figure 20 Happy F-1 Scores across validation datasets](#page63) [51](#page63)

[Figure 21 Sad F-1 Scores across validation datasets](#page65) [53](#page65)

[Figure 22 Anger F-1 Scores across validation datasets](#page66) [54](#page66)

iv

[Figure 23 Disgust F-1 Scores across validation datasets](#page67) [55](#page67)

[Figure 24 Gender bias on Emotion across DL models](#page68) [56](#page68)

[Figure 25 Male vs Female F-1 Score Bar Chart](#page72) [60](#page72)

**List of Tables**

[Table 1 Overall F-1 scores for each DL algorithm](#page48) [36](#page48)

[Table 2: Overall F-1 Score’s for trained and validated models](#page52) [40](#page52)

[Table 3 Accuracy Difference and Disparate impact chart [34][35]](#page56) [44](#page56)

[Table 4 : F-1 Scores for Fear emotion](#page59) [47](#page59)

[Table 5: F-1 Scores for Surprise emotion](#page60) [48](#page60)

[Table 6: F-1 Socres for Neutral Emotion](#page62) [50](#page62)

[Table 7 : F-1 Scores for Happy emotion](#page63) [51](#page63)

[Table 8 : F-1 Score Sad Emotion](#page64) [52](#page64)

[Table 9: F-1 Score Anger emotion](#page66) [54](#page66)

[Table 10: F-1 Score Disgust emotion](#page67) [55](#page67)

[Table 11 F-1 Score of Fear for Balanced Dataset](#page70) [58](#page70)

[Table 12 F-1 scores for Happiness using Mixed or balanced dataset](#page71) [59](#page71)

1

**CHAPTER 1: INTRODUCTION**

Facial Expression Recognition (FER) has emerged as a pivotal technology in various domains, including human-computer interaction, security, and psychological research. By interpreting human emotions through facial cues, FER systems enhance communication and decision-making processes, offering significant societal and technological benefits. However, the efficacy of these systems is often compromised by inherent biases, particularly gender bias, which can lead to skewed results and misinterpretations.

Gender bias in FER refers to the systematic discrepancies in the performance of recognition systems across different gender groups. This bias can manifest in various ways, such as higher error rates in emotion detection for one gender compared to another. These biases not only undermine the reliability of FER systems but also perpetuate stereotypes and inequalities, making it crucial to address these issues in the development and deployment of FER technologies.

Deep learning models have revolutionized FER by providing robust frameworks for emotion classification. Among the most prominent models are Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3. These models leverage convolutional neural networks (CNNs) to automatically extract and learn features from facial images, offering high accuracy and efficiency. However, the training datasets’ composition significantly

2

influences these models' performance, potentially introducing gender bias if the data is not representative.

**1.1 Problem Statement**

Facial expression recognition (FER) has become an essential technology with applications across various sectors, including healthcare, technology, security, and media. Despite the advancements in machine learning algorithms that can detect and categorize facial expressions and emotions, challenges persist due to biased datasets and variations in facial images. These biases can lead to inconsistent predictions, particularly concerning demographic factors such as gender. This research aims to evaluate the impact of gender on facial expression recognition across various deep learning models, focusing on identifying the existence of gender bias, determining which algorithms are most effective in mitigating such biases, and understanding which facial expressions are more susceptible to gender bias [**5]**.

The presence of gender bias in facial expression recognition systems can have significant implications for the fairness and accuracy of these technologies. Gender bias can result in unequal performance across different demographic groups, leading to potential discrimination and unfair treatment. This issue concerns applications where FER is used for decision-making processes, such as in security or healthcare settings. Therefore, it is crucial to understand the extent of gender bias in FER systems and to develop strategies for mitigating this bias [**5]**.

To address this problem, our research study will employ five deep learning algorithms: Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3. These algorithms have been selected for their benchmarking performance in facial expression

3

recognition. The study will involve training and testing these algorithms on male-only, female-only, imbalanced and balanced datasets, with fine-tuned hyperparameters to optimize performance. The experimentation will be conducted in three phases: training, testing, and performance assessment. Statistical measures such as accuracy, precision, recall, and F-1 score will be used to assess performance, alongside bias metrics like accuracy difference, F1-score difference and disparate impact [**6]**.

The findings of this research study are expected to provide valuable insights into the existence and extent of gender bias in facial expression recognition systems. By evaluating the impact of gender on FER across various deep learning models, the study aims to contribute to the development of fairer and more accurate FER systems. This research is crucial in addressing the challenges of gender bias in facial expression recognition and aims to provide a comprehensive understanding of the demographic effects on facial emotion expression and the impact of gender on fairness and recognition of expressions.

**1.2 CONTRIBUTIONS**

This research makes several significant contributions to the field of facial expression recognition (FER), particularly in understanding and addressing gender bias across various deep learning models.

* **Identification of Gender Bias**: The study systematically evaluates the presence of gender bias in facial expression recognition systems. By employing a range of deep learning algorithms, including Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3, the research identifies the extent to which gender bias exists in these models.

4

* **Algorithmic Evaluation**: The research provides a comprehensive evaluation of five deep learning algorithms to determine which are most effective in mitigating gender bias. By benchmarking these algorithms for overall facial expression recognition accuracy, the study offers insights into their strengths and weaknesses in handling gender-related biases.
* **Dataset Analysis**: The study explores the impact of different datasets on gender bias in FER systems. By using male-only, female-only, imbalanced and balanced datasets, the research assesses how dataset composition affects the performance and fairness of deep learning models. This analysis is vital for understanding the role of data in perpetuating or alleviating bias in machine learning systems.
* **Implications for Fairness and Ethics**: By addressing gender bias in FER systems, the research contributes to the broader discourse on fairness and ethics in artificial intelligence. The findings underscore the importance of developing equitable machine learning models that perform consistently across diverse demographic groups [**5]**. This contribution is essential for ensuring that FER technologies are used responsibly and do not perpetuate existing societal biases.
* **Simultaneous DL model testing**: This research is unique in its comprehensive evaluation of five deep learning algorithms—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—specifically for facial expression recognition (FER). These models were selected for their benchmarking performance and are systematically tested to assess their effectiveness in mitigating gender bias.

5

* **Identified a unique bias metric**: In addition to traditional bias metrics like Accuracy Difference (AD) and Disparate Impact (DI), we introduced a unique bias metric—F1-Score Difference. This metric provides additional insights into the performance disparities across different demographic groups and was integral to our evaluation process.
* **Identified unique dataset configurations**: Our study stands out as it extensively experiments with five distinct datasets, including male-only, female-only, balanced, male-dominant, and female-dominant datasets. This approach ensures a thorough evaluation of the impact of dataset composition on model performance and gender bias.

Overall, this research advances the understanding of gender bias in facial expression recognition and provides actionable insights for developing fairer and more accurate FER systems. The contributions made by this study are expected to have significant implications for the design and implementation of FER technologies across various industries.

6

**CHAPTER 2: BACKGROUND AND**

**LITERATURE REVIEW**

The literature on facial expression recognition highlights the challenges posed by biased datasets and the resulting inconsistencies in prediction accuracy [**6]**. Studies have shown that demographic bias can significantly impact the performance of FER systems, leading to unequal treatment of different demographic groups. This issue is particularly concerning in contexts where FER is used for decision-making processes, such as in security or healthcare settings [**5]**.

Recent research has focused on understanding and mitigating these biases by employing various deep learning algorithms [**6]**. The selected algorithms for this study—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—have been benchmarked for their performance in facial expression recognition. These models are trained and tested on male-only, female-only, and mixed datasets to evaluate their effectiveness in addressing gender bias [**4][38]**.

The exploration of demographic bias in facial expression recognition is crucial for developing fairer and more accurate systems. By examining the impact of gender on FER across different deep learning models, this research aims to contribute to the broader discourse on fairness and ethics in artificial intelligence. The findings are expected to provide valuable insights into the design and implementation of FER technologies, ensuring they perform consistently across diverse demographic groups [**5]**.

This chapter will delve into the existing literature on facial expression recognition, focusing on the challenges of gender bias and the methodologies employed to address these issues.

7

It will provide a comprehensive overview of the current state of research in this field, setting the stage for the experimental studies conducted in subsequent chapters.

**2.1 Demographic Bias in Facial Expression Recognition**

Facial expression recognition (FER) systems have become increasingly prevalent in various sectors, including healthcare, technology, security, and media, due to their ability to automatically detect and categorize facial expressions and emotions. However, the effectiveness of these systems is often compromised by demographic biases present in the datasets used for training [**4]**. Demographic bias refers to the unequal performance of machine learning models across different demographic groups, which can lead to unfair treatment and discrimination [**5][39]**.

One of the primary sources of demographic bias in FER systems is the composition of the training datasets. Many datasets used in training these models are not representative of the diverse population they are intended to serve. For instance, datasets may have an overrepresentation of certain demographic groups, such as a particular gender or ethnicity, leading to models that perform well for those groups but poorly for others [**4]**. This imbalance can result in models that are less accurate and reliable when applied to underrepresented groups [**5][38]**.

Gender bias is a specific type of demographic bias that has been studied in the context of FER [**4]**. Researchers have pointed out that gender bias can impact the performance of FER systems, leading to unequal treatment of male and female subjects. This issue concerns applications where FER is used for decision-making processes, such as in security or healthcare settings, where biased outcomes can have serious implications [**5]**.

8

To address demographic bias in FER systems, it is crucial to develop strategies that ensure fair and equitable performance across all demographic groups [**5]**. This includes using diverse and representative datasets for training, as well as employing algorithms that are robust to variations in demographic characteristics. Additionally, evaluating models using bias metrics, such as accuracy difference, difference in positive proportions in predicted labels, and disparate impact ratio, can help identify and mitigate biases in FER systems [**6]**.

**2.2 Gender Bias in Facial Expression Recognition**

The root of gender bias in FER systems often lies in the datasets used for training these models. Many datasets are not adequately balanced in terms of gender representation, leading to models that are more accurate for the overrepresented gender. This imbalance can result in skewed learning, where the model becomes more adept at recognizing expressions from the dominant gender in the dataset, while its performance degrades for the underrepresented gender [**6]**.

Research has shown that gender bias can impact the fairness and reliability of FER systems. For instance, studies have indicated that certain facial expressions may be more susceptible to gender bias, with models showing varying levels of accuracy in recognizing emotions such as happiness or anger across genders [**6]**. This discrepancy can have serious implications, particularly in applications where FER is used for decision-making processes, such as in security screenings or psychological assessments [**5]**.

To address gender bias in FER, it is crucial to employ strategies that ensure equitable performance across genders. This includes using balanced datasets that represent both male and female subjects equally, as well as implementing algorithms that are robust to demographic variations [**6]**. Additionally, evaluating models using bias metrics, such as

9

equal opportunity difference and disparate impact, can help identify and mitigate biases in FER systems [**6]**.

The research outlined in this document aims to explore the existence and extent of gender bias in FER systems by employing a range of deep learning algorithms, including Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3 [**6]**. By training and testing these models on male-only, female-only, and mixed datasets, the study seeks to determine which algorithms are most effective in mitigating gender bias as well as which facial expressions are more prone to such biases.

Addressing gender bias in facial expression recognition is essential for developing fair and accurate systems that perform consistently across genders. By understanding and mitigating these biases, researchers and practitioners can ensure that FER technologies are used responsibly and do not perpetuate existing societal inequalities [**5]**.

**2.3 Selected Deep Learning (DL) Algorithms**

This study not only contributes to the technical advancement of FER systems but also addresses the ethical implications of deploying biased technologies in society. Facial

Expression Recognition (FER) is a sophisticated process that involves identifying and categorizing human emotions based on facial expressions. This technology is a subset of computer vision and leverages machine learning and deep learning techniques to interpret facial cues.

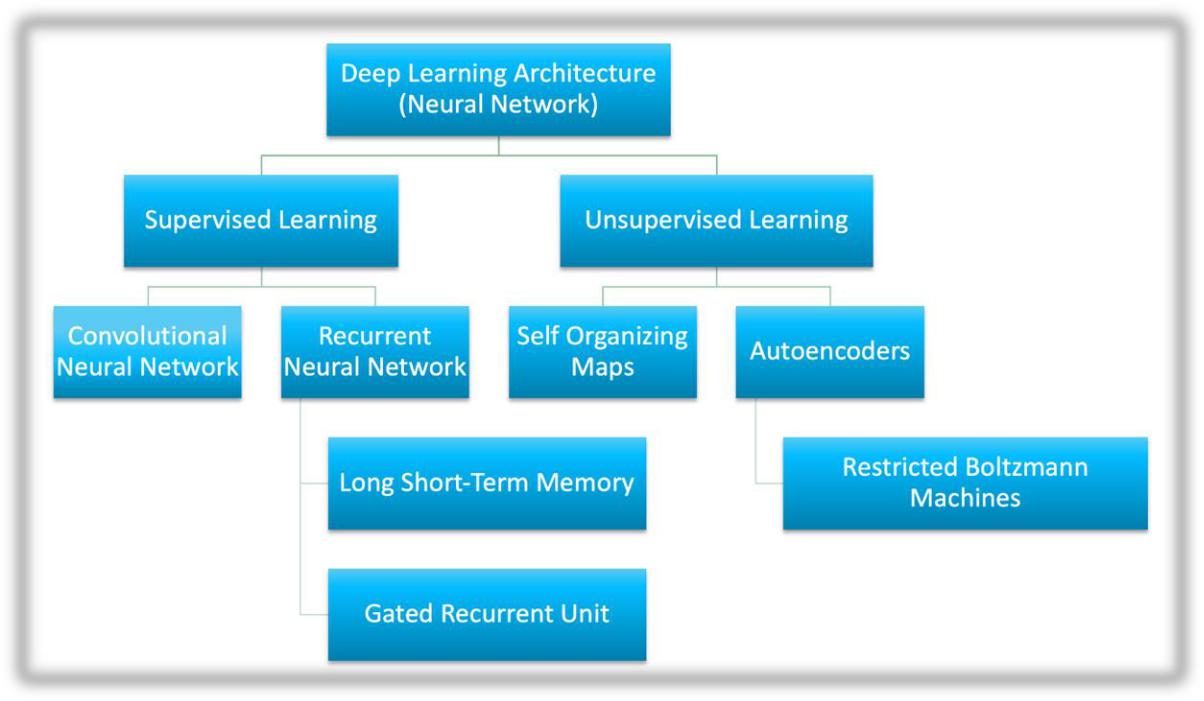
Deep learning architectures have come to represent a pivotal advancement in Facial Expression recognition, deep learning architectures are mainly comprised of significant key components [18].

10

* **Neurons:** These are the basic building blocks of neural networks, functioning similarly to brain neurons by receiving inputs, applying weights, and generating outputs.
* **Activation Functions:** These functions decide if a neuron should activate. For instance, the ReLU (Rectified Linear Unit) function passes through positive values while zeroing out negative ones, influencing the neuron's activity.
* **Weights and Biases:** These are crucial adjustable parameters that determine the importance of input features and set activation thresholds. Weights reflect feature significance, while biases ensure a minimum activation level is met. Training adjusts these to enhance model accuracy.
* **Loss Functions:** These functions measure the gap between predicted and actual outcomes, guiding the learning process to minimize this gap. Popular loss functions include Mean Squared Error (MSE) and Cross-Entropy Loss [30].
* **Optimizers:** Essential for updating weights and biases based on the loss, optimizers refine model parameters to reduce loss and boost performance. Techniques like Stochastic Gradient Descent (SGD), Adam, and RMSprop are widely used [30].

Armed with these fundamental concepts, we can explore more sophisticated deep learning architectures, where layers of neurons collaborate to tackle complex challenges [18].

11

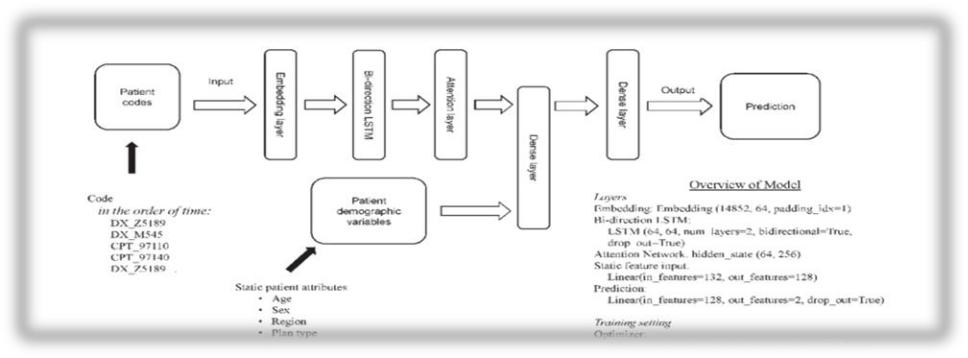


*Figure 1 Types of Deep Learning algorithms [18]*

The study utilizes five deep learning models for facial expression recognition: Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3. These models were selected based on their benchmarking performance for overall facial expression recognition accuracy.

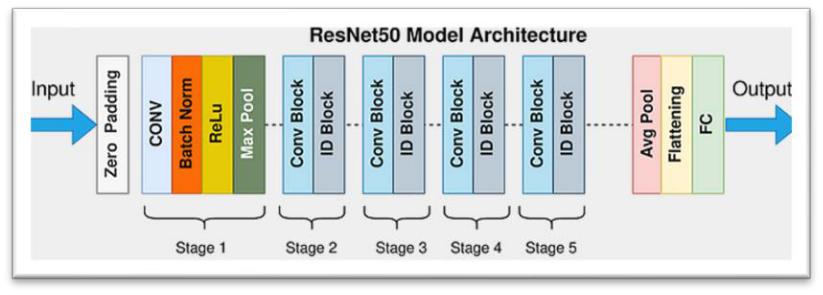
* **Sequential Model**: This is a simple, linear stack of layers in a neural network, where each layer has exactly one input tensor and one output tensor. It is often used for straightforward neural network architectures [1].

12



*Figure 2 Sequential Model Architechture [23]*

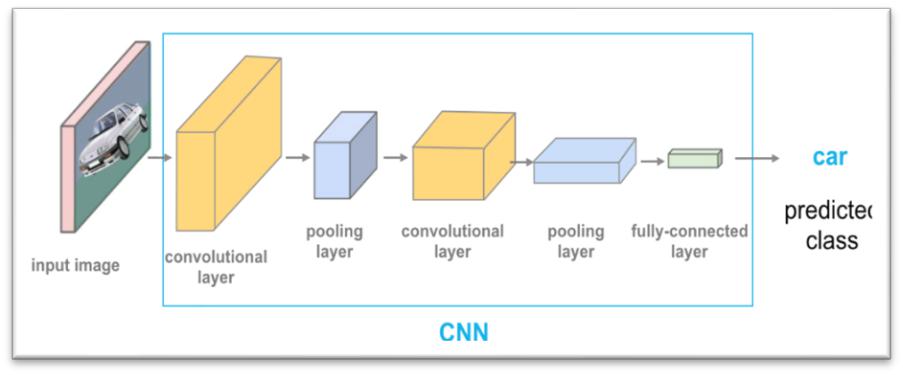
* **ResNet-50**: ResNet-50 is a deep residual network that consists of 50 layers. It is known for its ability to train very deep networks by using residual learning, which helps in mitigating the vanishing gradient problem [**5]**.



*Figure 3 ResNet50 Model Architechture [24]*

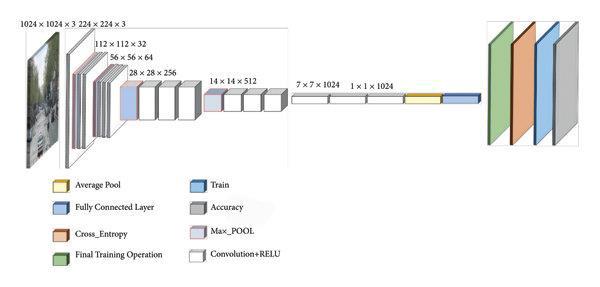
* **DenseNet**: DenseNet, or Dense Convolutional Network, connects each layer to every other layer in a feed-forward fashion. This model is efficient in terms of parameter usage and is known for its ability to strengthen feature propagation and encourage feature reuse [**1]**.

13



*Figure 4 DenseNet Model Architechture [25]*

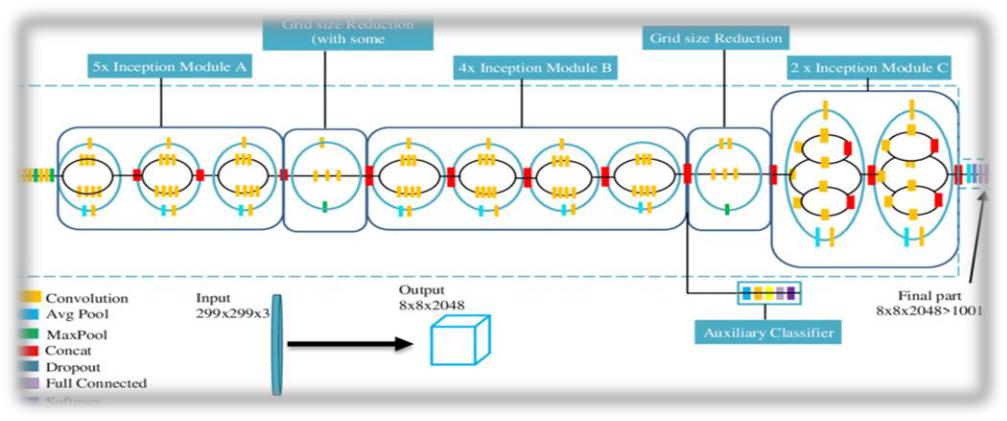
* **MobileNetV2**: MobileNetV2 is a lightweight deep learning model designed for mobile and edge devices. It uses depth wise separable convolutions to reduce the number of parameters and computational cost, making it efficient for real-time applications [**5]**.



*Figure 5 MobileNetV2 Model Architecture [22]*

* **InceptionV3**: InceptionV3 is a convolutional neural network that is part of the Inception family. It is known for its complex architecture that includes multiple types of convolutions and pooling operations, allowing it to capture a wide range of features [**5]**.

14



*Figure 6 InceptionV3 Architechture [19]*

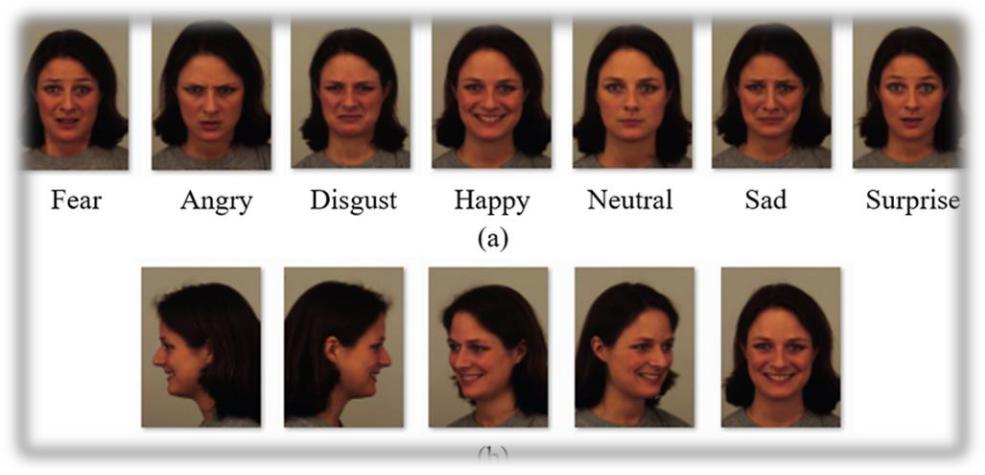
These models are trained and tested on male-only, female-only, and mixed datasets, with fine-tuned hyperparameters to optimize performance [**1]**. The experimentation involves three phases: training, testing, and performance assessment, using statistical measures such as accuracy, precision, recall, and F-1 score, alongside bias metrics like accuracy difference and disparate impact [**5]**.

**2.4 Facial Expression Image Database Used for Demographic Bias**

The Karolinska Directed Emotional Faces (KDEF) dataset is a widely used resource in FER research, providing a comprehensive collection of facial expressions across various emotions. KDEF dataset contains 4900 images with 7 facial expressions (happy, sad, surprised, angry, disgust, afraid and neutral). Participants are 35 males and 35 females with images taken at 5 different angles. Its structured and diverse nature makes it an invaluable tool for training and evaluating FER models. However, the dataset's demographic composition, particularly in terms of gender representation, can impact the models' ability to generalize across different gender groups. This thesis aims to evaluate the impact of

15

gender bias on FER by employing the deep learning models and the KDEF dataset. By systematically analyzing the performance of these models across gender-diverse datasets, the research seeks to uncover the extent of gender bias [1].



*Figure 7 Images from KDEF dataset [16]*

Along with the widely used KDEF dataset, initial benchmarking was conducted using other widely used datasets FER2013 and CK+. The FER2013 dataset is a large-scale dataset used for facial expression recognition tasks. It consists of approximately 30,000 grayscale images of faces, each with a resolution of 48x48 pixels. The images are labeled with seven different emotions: angry, disgust, fear, happy, sad, surprise, and neutral. This dataset was introduced as part of the ICML 2013 Challenges in Representation Learning

* The CK+ (Extended Cohn-Kanade) dataset is also a well-known dataset in the field of facial expression recognition. It includes both posed and spontaneous expressions collected from 123 participants. The dataset contains 593 sequences across 123 subjects, with each sequence capturing the onset to peak of an expression. It covers seven facial expression categories: anger, contempt, disgust, fear, happiness, sadness, and surprise. The CK+

16

dataset is often used for developing and evaluating facial expression recognition algorithms

[17].

17

**CHAPTER 3: SYSTEM DESIGN &**

**DEVELOPMENT**

**3.1 System Architecture**

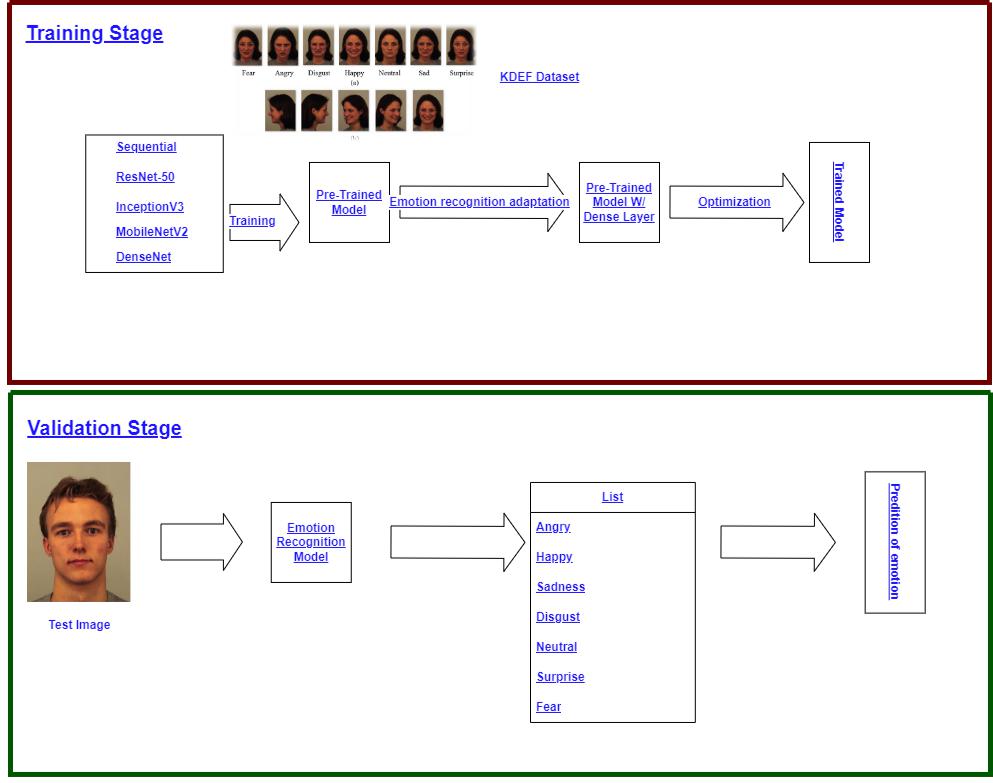
The system architecture for facial expression recognition (FER) is designed to evaluate gender bias and improve fair performance across diverse demographic groups. This architecture leverages advanced deep learning models and robust data handling techniques to enhance the accuracy and fairness of FER systems [12].

1. **Data Collection and Preprocessing**: Using datasets with male-only, female-only, and mixed-gender samples to mitigate gender bias. Then applying data augmentation techniques like rotation and scaling to improve model robustness.
2. **Face Detection and Landmark Identification**: Implement algorithms such as Haar cascades or CNNs for face detection. Use Active Shape Models (ASM) or Active Appearance Models (AAM) for identifying facial landmarks.
3. **Feature Extraction**: Analyze geometric relationships between landmarks and use texture-based methods to capture facial expression features.
4. **Model Selection and Training**: Employ deep learning models like Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3. Conduct training, testing, and performance assessment using statistical measures and bias metrics.
5. **Expression Classification**: Use classification models such as SVMs or CNNs to recognize expressions like happiness, sadness, and anger.

18

1. **Post-Processing and Interpretation**: Integrate classified expressions into applications for insights or actions, ensuring ethical and fair use of FER technologies.
2. **Bias Evaluation:** Evaluate models using metrics like accuracy difference and disparate impact to identify and mitigate gender bias.

This architecture aims to advance understanding of gender bias in FER systems and provide insights for developing fairer and more accurate technologies [1].



*Figure 8 Block diagram of the general FER model [26]*

19

Figure 8 Block diagram gives an example of the same workflow visualization utilized to conduct experimentation for gender bias. Within the model extra measures are conducted to calculate accuracy differences and disparate impacts across male-trained, female-trained, and balanced-trained deep learning models.

**3.2 Evaluation Plan**

The evaluation plan for this research on facial expression recognition bias aims to systematically assess the impact of gender bias across various deep learning models. The plan is to ensure comprehensive analysis and validation of the study’s objective, determining whether a bias exists across various deep learning models. We adopted the below methodology for conducting bias experimentation in facial expression recognition following the similar research studies in the literature [27].

Dataset Preparation

* Balanced and Diverse Datasets: Utilize male-only, female-only, and mixed-gender datasets to ensure diverse representation. This includes both balanced and imbalanced datasets to evaluate the impact of dataset composition on model performance and bias.
* Data Augmentation: Apply techniques such as rotation, scaling, and flipping to enhance dataset diversity and improve model robustness [17].

Model Training and Testing

20

* Algorithm Selection: Employ five deep learning algorithms—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—selected for their benchmarking performance in FER.
* Benchmarking: The benchmarking process involves algorithmic evaluation, where the strengths and weaknesses of each model are analyzed in the context of gender bias, by systematically comparing the performance of different algorithms across various datasets ( Male-Only, Female-Only, Balanced ) [4].
* Training Phases: Conduct training in three phases: training, testing, and performance assessment. Each algorithm will be trained using the specified datasets, with fine-tuned hyperparameters to optimize performance.

Performance Metrics

* Statistical Measures: Evaluate model performance using F-1 score to provide a comprehensive assessment of each algorithm's effectiveness.
* Bias Metrics: Use bias metrics such as accuracy difference, F1-score difference and disparate impact to identify and quantify gender bias in the models [24].

Bias Mitigation Evaluation

* Algorithmic Evaluation: Assess which algorithms are most effective in mitigating gender bias by comparing their performance across different datasets and demographic groups.
* Fairness Metrics: Implement fairness metrics like equal opportunity difference to ensure equitable performance across genders [24].

Post-Processing and Interpretation

21

* Result Integration: Analyze the classified facial expressions to provide insights or take actions based on the recognized emotions. This step involves integrating results into applications like adaptive learning systems or real-time audience analysis.
* Ethical Considerations: Address fairness and ethics by ensuring that FER technologies do not perpetuate societal biases, contributing to responsible AI development [24].

Reporting and Analysis

* Comprehensive Reporting: Document the findings, highlighting the existence and extent of gender bias in FER systems. Provide actionable insights for developing fairer and more accurate FER technologies.
* Implications for Future Research: Discuss the implications of the findings for the design and implementation of FER systems, particularly in terms of ensuring fairness and accuracy across different demographic groups.

The plan involves employing five deep learning algorithms—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—chosen for their benchmarking performance in FER tasks. These models are trained and tested on diverse datasets, including male-only, female-only, and mixed datasets, to ensure a comprehensive evaluation of their performance and bias [27].

**3.3 System Software**

Traditionally facial expression recognition experimentation has been conducted using one of the most versatile and easy to use programming languages, Python, following the

22

success of many previous FER experiments, our experiments were conducted using Python as well. While Python is very common in many development environments, the development environment Jupyter Notebook specifically allowed for more flexibility in fine-tuning model parameters, dataset preprocessing, and model accuracy evaluation. Due to the fine-tuned nature of this experimentation Anaconda, a python distribution specifically designed for data science and machine learning [28], was utilized to create a specific instance of python with Jupyter Notebook

Jupyter Notebook enhances flexibility in model parameter tuning, dataset preprocessing, and evaluation. Our experiments utilized a modular approach, the Data Preprocessing Module handled data cleaning and augmentation, the Model Training Module implemented algorithms such as ResNet50, DenseNet, and deep learning frameworks like TensorFlow and Keras. The Model Evaluation Module assessed model’s bias and performance by using metrics including disparate impact and accuracy difference. The User Interface Module facilitated interaction through Jupyter widgets for parameter adjustment and result visualization. These modules interacted through a seamless pipeline, supported by Anaconda, which ensured the integration of essential libraries like:

* Numpy 1.15.4
* Pandas 1.1.5
* Keras 2.2.2 / Matplotlib 3.3.4
* Anaconda Navagator 2.5.2
* Python 3.6.2
* Jupyter Notebook 6.4.3
* Scikit-learn 0.24.2

23

* TensorFlow 1.1.0

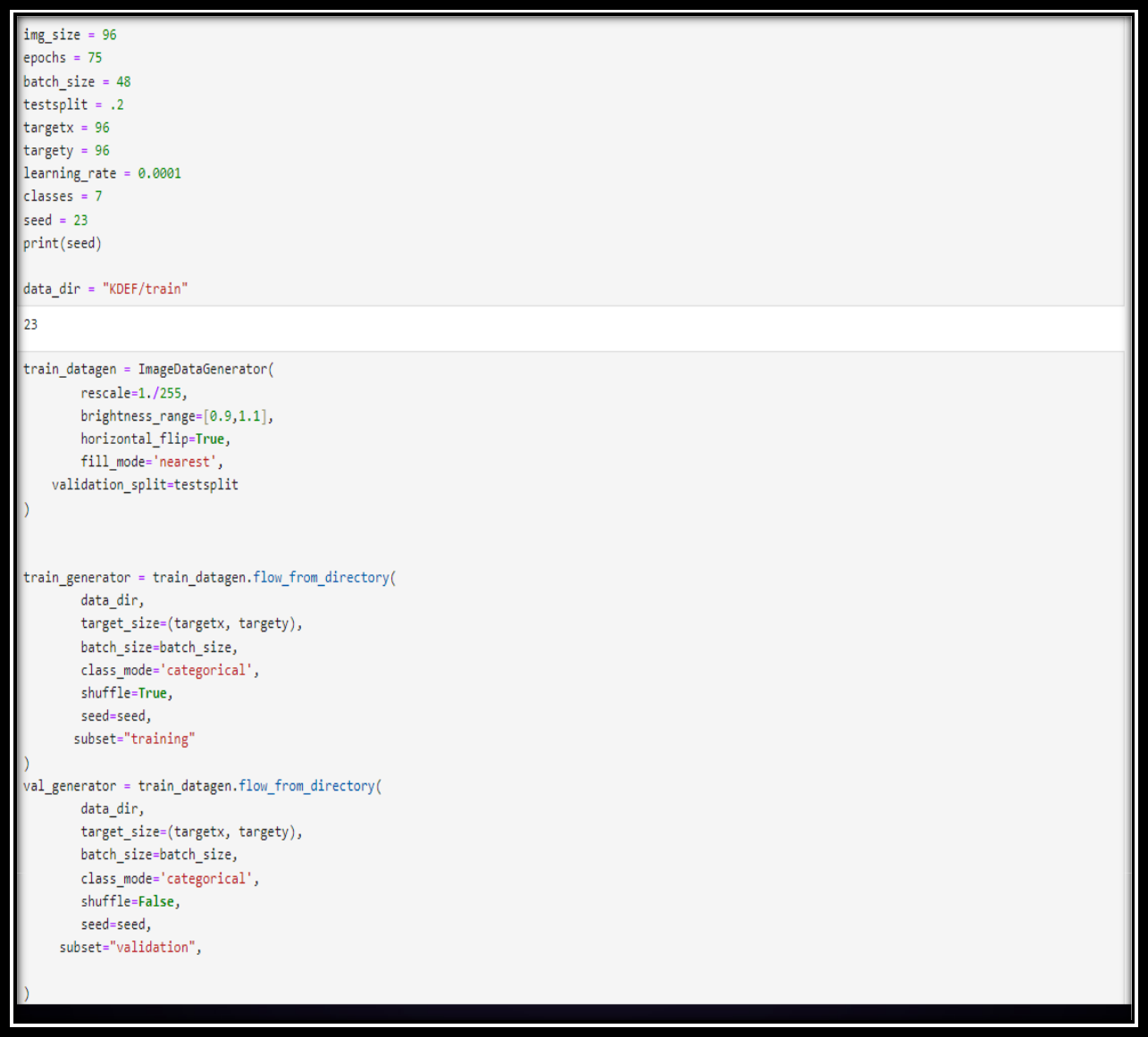
The experimentation was conducted using a Dell G16 7620 laptop, the system can conduct FER experimentation equipped with a 12th Gen Intel® Core ™ i7 processor, the system is also equipped with a NVIDIA GeForce RTX 3050 Ti GPU in order to handle the heavy processing requirements for FER experimentation.

24

**CHAPTER 4: EXPERIMENTAL STUDY**

The experimental study involves employing five deep learning algorithms— Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—selected for their benchmarking performance in FER tasks. These models are trained and tested on diverse datasets, including male-only, female-only, and mixed datasets, to ensure a comprehensive evaluation of their performance and bias. The experimentation is conducted in three phases: training, testing, and performance assessment. Statistical measures such as accuracy, precision, recall, and F-1 score are used to assess performance, alongside bias metrics like accuracy difference and disparate impact [16].

25



*Figure 9 Code snippet showing static training variables*

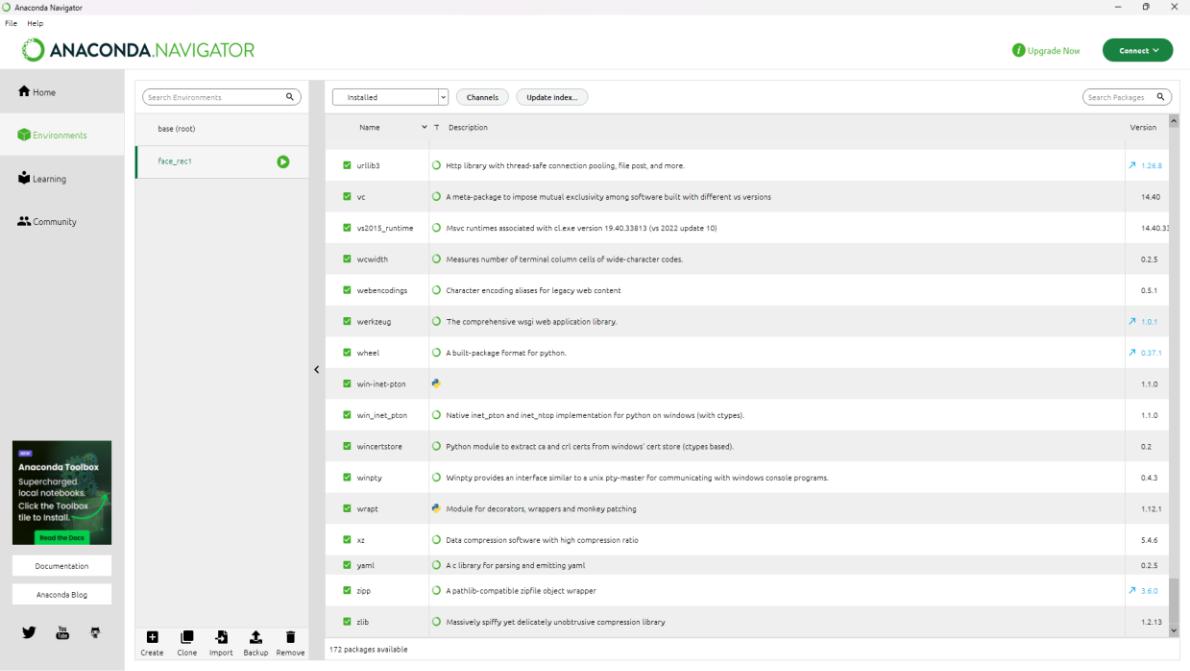
**4.1 Experimental Design**

**4.1.1 Methodology**

The following steps were taken to conduct this experiment:

26

* System Setup: Use a high-performance computer with a powerful processor and GPU to handle the processing requirements for facial expression recognition (FER) experimentation.
* Development Environment: Set up a Python environment using Jupyter Notebook for flexibility in model parameter tuning, dataset preprocessing, and evaluation. Use Anaconda to manage the environment and integrate essential libraries. Figure 10 illustrates a development environment created for facial expression recognition; this environment allowed for multiple instances of Jupyter notebook to utilize the same libraries during experimentation.



*Figure 10 Anaconda Navigator User Interface [28]*

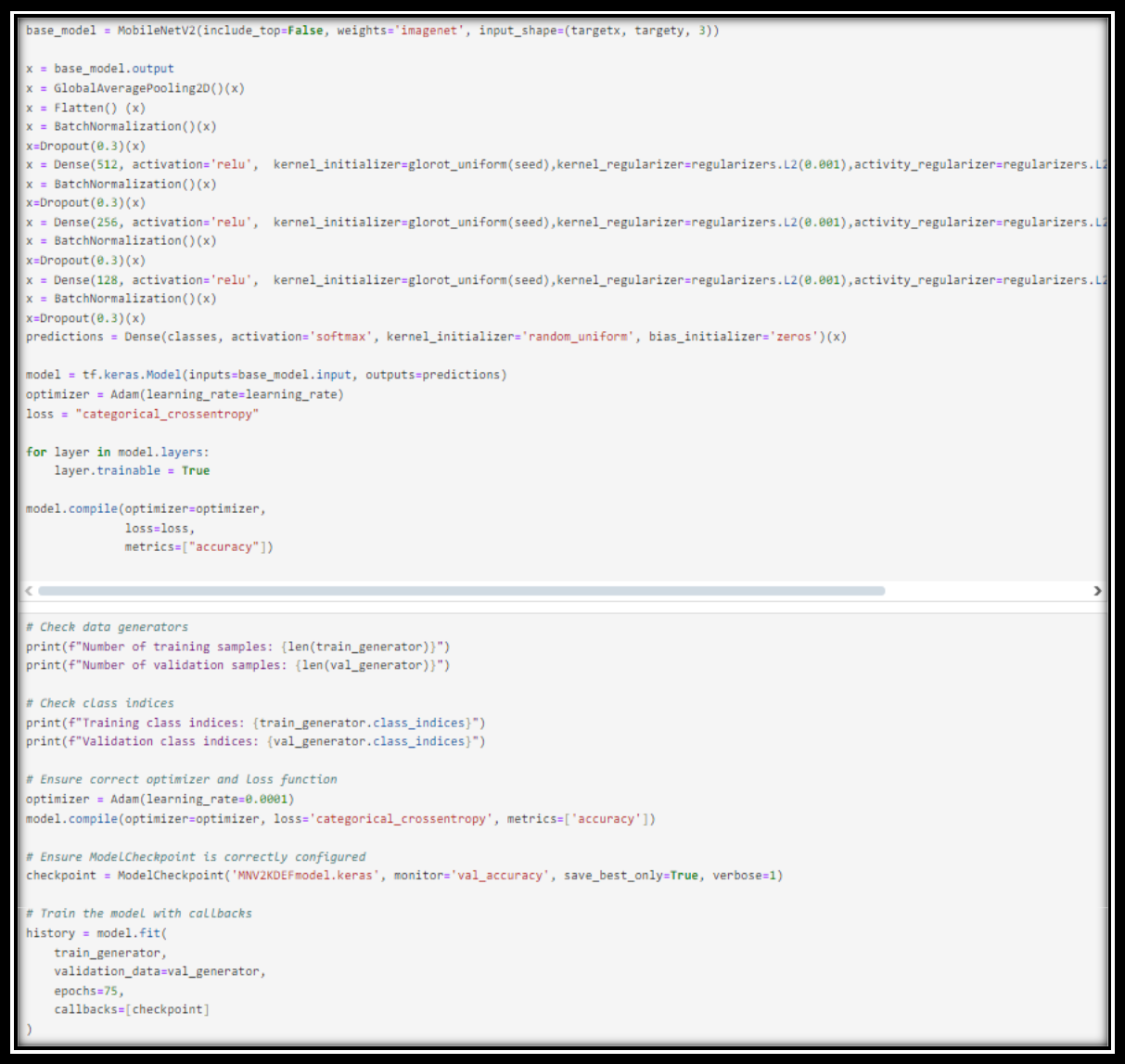
* Algorithm Selection: Choose deep learning algorithms like Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3 for their performance in FER tasks.

o Prior to final experimentation the deep learning model VGG-16 had also been considered. After initial benchmarking using unmodified FER

27

datasets, we found the performance of VGG-16 to be not comparable to the performance of the chosen deep learning algorithms.

* Training Phases: Conduct experimentation in three phases: training, testing, and performance assessment. Train each algorithm using the specified datasets, with fine-tuned hyperparameters to optimize performance.



*Figure 11 Code snippet of model training set up in Jupyter Lab*

28

* Figure 11 gives a code snippet of how each of the models were trained during experimentation with the following steps:

o Base model setup: Each experimental trial began with initializing the base model with pre-trained weights from ImageNet to reduce the training workload. Along with the pre-trained weights the input shape is set with *targetx, targety, and* 3 to indicate RGB channels.

o Model Customization: The model’s output is passed through a

GlobalAveragePooling2D layer to reduce the feature maps dimensions, several fully connected (Dense) layers are added with ReLU activation [31], each followed by Batch Normalization and Dropout layers, all here to prevent overfitting of the model.

* + The final layer is a dense layer with a ‘softmax’ activation, which

outputs class probabilities for multi-class classification task

1. Model Compilation: A tensorflow keras model is created using the base model input and defined custom layers. The model is compiled with an Adam optimizer, a categorical cross-entropy loss function [32], and

accuracy as the metric with all layer set to trainable.

o Model Checkpoint Configuration: A model checkpoint callback is configured to save the best model based on the validation accuracy.

1. Model Training: A model is trained using training and validation data generators for 75 epochs, a complete pass of a training dataset through a machine learning algorithm [1], the model checkpoint callback is used during training to save the best-performing model.

29

* Performance Metrics: Evaluate model performance using statistical measures such as accuracy, precision, recall, and F-1 score.
* Algorithmic Evaluation: Assess which algorithms are most effective in mitigating gender bias by comparing their performance across different datasets and gender groups.
* Bias Metrics: Use bias metrics like accuracy difference, F1-score difference and disparate impact to identify and quantify gender bias in the models.
* Post-Processing and Interpretation: Analyze the classified facial expressions to provide insights or take actions based on the recognized emotions. Integrate results into applications like adaptive learning systems or real-time audience analysis.
* Reporting and Analysis: Document the findings, highlighting the existence and extent of gender bias in FER systems. Provide actionable insights for developing fairer and more accurate FER technologies.

**4.1.2 Preparing and Preprocessing Data**

Prior to selecting KDEF as the primary experimentation dataset, benchmarking was conducted using both the CK+ and FER2013 datasets which showed poor results overall for specific reasons. The process and results for the benchmarking conducted with CK+ and FER2013 are as follows.

The benchmarking process using deep learning models against the CK+ dataset encountered challenges, particularly reflected in the low F-1 scores achieved by the models. Despite the CK+ dataset's comprehensive coverage of facial expressions, the deep learning algorithms—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—struggled to maintain high performance levels. This underperformance can

30

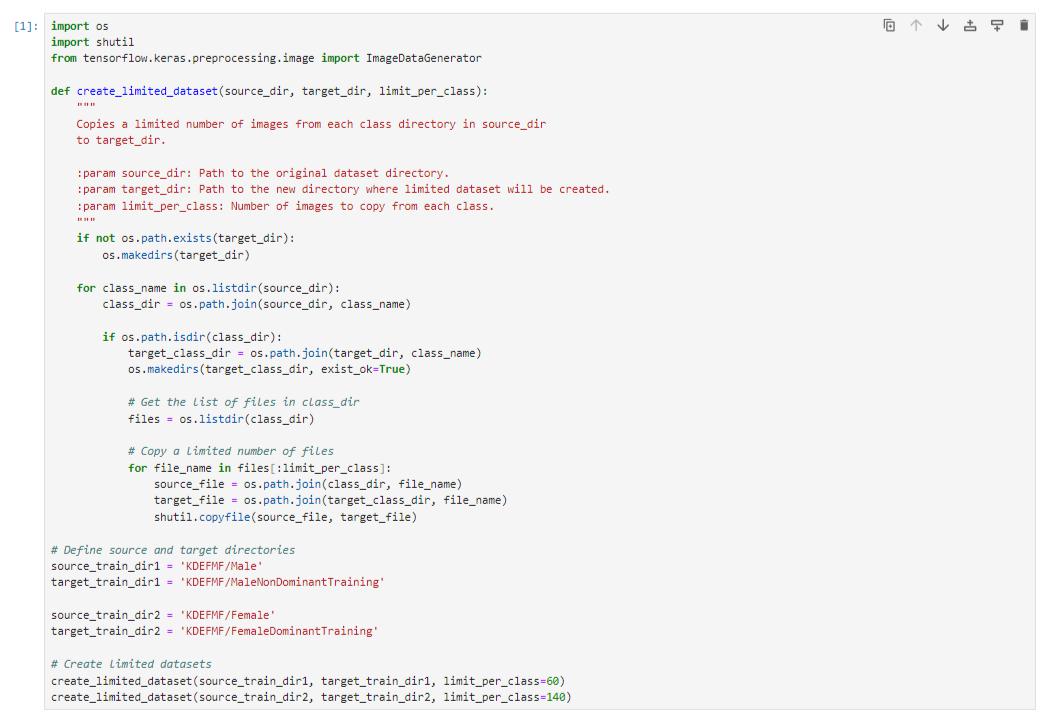
be attributed to several factors, including the dataset's demographic composition and the inherent complexity of accurately classifying subtle facial expressions. The low F-1 scores indicate a significant discrepancy between precision and recall, suggesting that the models were either generating a high number of false positives or failing to detect true positives effectively. This outcome highlights the need for further refinement in model training and dataset preparation to enhance the robustness and accuracy of facial expression recognition systems when applied to datasets like CK+.

The benchmarking process using deep learning models against FER2013 dataset faced significant challenges, primarily due to the dataset’s format [17], which consists of grayscale images represented in pixel values. This format posed difficulties in sorting images by gender, as the dataset lacks explicit gender labels or metadata that could facilitate such categorization. Consequently, the inability to segregate the images by gender hindered the evaluation of gender bias in facial expression recognition (FER) systems. The deep learning models—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—were unable to effectively assess and address gender-specific performance discrepancies, as the dataset's structure did not support a straightforward analysis of gender-based biases. This limitation underscores the importance of having well-annotated datasets that include demographic information, which is crucial for conducting comprehensive bias evaluations and ensuring the development of fair and accurate FER technologies.

Following the benchmarking conducted using all three datasets KDEF was found to produce the most consistent results across all models and in both balanced and diverse subset evaluations. With benchmarking concluded, dataset preparation was able to

31

commence for experimentation. For the purposes of experimentation for gender bias, a total of five balanced and diverse datasets were prepared including, male-only, female-only, balanced, male-dominant, and female-dominant, datasets of the KDEF database. This is to ensure a diverse representation to evaluate the impact of dataset composition on model’s gender bias. Each dataset was prepared using both manual and automated methods, for male and female-only datasets images where manually selected from the KDEF database and placed into subsequent male and female subfolders. Balanced, male and female-dominant datasets were created algorithmically using os and shutil libraries in python.



*Figure 12 Algorithmically splitting the datasets using Jupyter Notebook*

Figure 12 depicts a direct code snippet of the algorithm used to create limited datasets for balanced, male and female-only training datasets. In this code snippet, the os

32

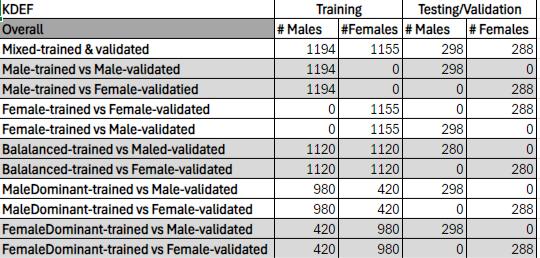
and shutil libraries are imported to handle file and directory operations, while ImageDataGenerator is imported from TensorFlow, although it is not utilized in this snippet. The function create\_limited\_dataset is defined to take three parameters: source\_dir, which is the path to the original dataset; target\_dir, the path for the new, limited dataset; and limit\_per\_class, which specifies the number of images to copy per class. The function begins by checking if the target\_dir exists, creating it with os.makedirs if it does not. It then iterates over each class directory within the source\_dir, constructing the path for each class directory (class\_dir). If the current path is a directory, a corresponding directory is created in the target\_dir. The function lists all files in the current class directory and copies up to limit\_per\_class files from the source class directory to the target class directory using shutil.copyfile. The directories source\_train\_dir1 and source\_train\_dir2 specify the source directories for male and female datasets, respectively, while target\_train\_dir1 and target\_train\_dir2 specify the target directories for the limited datasets. A total of 160 images per emotion are used for training with 40 images per emotion being used for validation. This algorithmic method ensures that images from the testing dataset are not used for validation to ensure accurate bias measures.

**4.1.3 Experiment Setup**

During the experimentation phase, each deep learning model—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—was trained over 75 epochs with a batch size of 32. The dataset was split with a test split of 0.25, and the models were designed to classify 7 different facial expression classes. This setup ensured that the models had sufficient time to learn and optimize their parameters for facial expression recognition tasks. Figure 9 illustrates how static variables were established during experimentation; the

33

number of epochs would remain the same for all five deep learning models. The training process involved iterating over the entire dataset 75 times, allowing the models to progressively adjust their weights and biases to minimize the loss function. This iterative process aimed to enhance the models' ability to accurately classify facial expressions while also evaluating their performance in terms of gender bias. By using 75 epochs, the study sought to balance the need for thorough training to reduce the risk of overfitting [6], ensuring that the models could generalize well to new, unseen data [6].



*Figure 13 Data Configurations for training and validation scenarios*

The dataset configurations in Figure 13 illustrate a strategic approach to understanding gender dynamics in facial expression recognition. In the mixed-trained and validated scenario, both the training and testing phases include a nearly equal number of males (1194 in training and 298 in testing) and females (1155 in training and 288 in testing), ensuring gender balance. In the male-trained scenarios, all training samples (1194) are male, with testing divided between male (298) and female (288) validation sets, allowing for a direct comparison of model performance across genders. Similarly, the female-trained scenarios exclusively use female samples for training (1155), with

34

validation occurring on female samples (288), and cross-gender testing on male samples (298). The balanced-trained approaches maintain equal male and female representation (1120 each) in training, testing separately on male (280) and female (288) samples. The male-dominant and female-dominant scenarios introduce a deliberate gender imbalance, with 980 males and 420 females or vice versa, providing insight into the effects of gender dominance in training data on model validation. Through these varied configurations, the study seeks to uncover potential biases and improve the fairness and accuracy of emotion recognition systems.

**4.2 Experimentation**

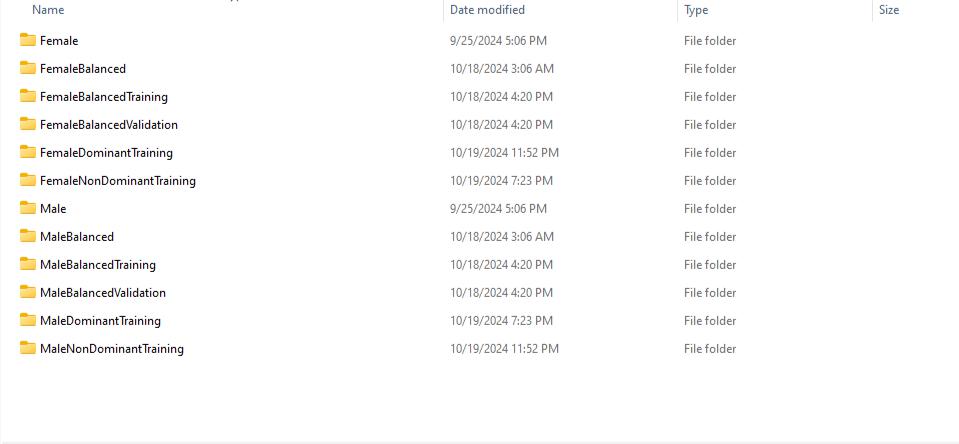
Utilizing a range of deep learning models, including Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3, the study aims to evaluate the performance and fairness of these algorithms across diverse datasets. The experimentation is structured into three phases: training, testing, and performance assessment, with a focus on optimizing model parameters and assessing bias through statistical measures such as accuracy, precision, recall, and F-1 score. Additionally, bias metrics like accuracy difference and disparate impact are employed to quantify gender bias. By employing male-only, female-only, and mixed-gender datasets, the research seeks to provide comprehensive insights into the demographic effects on FER and the efficacy of different models in mitigating gender bias. This section outlines the methodologies and tools used, including Python and Jupyter Notebook, to ensure a robust and flexible experimental framework.

**4.2.1 Gender Bias Across DL Algorithms**

Experimentation for gender bias was conducted in 6 separate stages, as baseline results which were collected using the KDEF dataset without any modifications to the

35

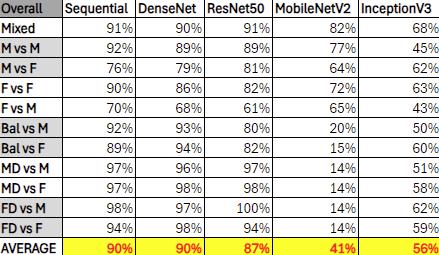
dataset. In the second stage, a subset of the KDEF dataset with manually selected images of only males, similarly to the second stage another subset of the KDEF dataset was manually created with only female images, with both subsets being used for training and validation of deep learning models. Following these experiments algorithmic methods were used to produce subsets for balanced, male-dominant, and female-dominant datasets. With all datasets compiled, training and validation was conducted across five separate deep learning algorithms.



*Figure 14 Directories created for model training and validation*

Figure 13 depicts the tree of directories created from manual and algorithmic methods; each directory contains seven sub-directories for each of the seven emotions being tested which are happy, sad, anger, fear, surprise, and neutral. This file structure is necessary in order for each deep learning algorithm to access training images labeled per each emotion, adding this extra layer provides metadata to determine the gender of each facial image.

36



*Table 1 Overall F-1 scores for each DL algorithm*

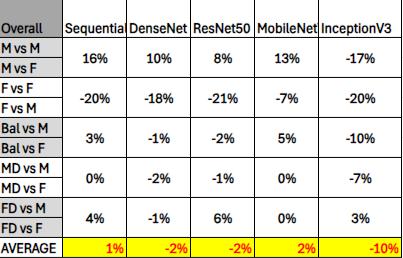
The results of the overall F-1 scores across different deep learning models in Table 1 reveal intriguing insights into the performance variations between imbalanced trained models and those trained on male or female datasets. Sequential and DenseNet models demonstrate consistently high performance, with both achieving an average F-1 score of 90%. These models appear to handle gender imbalances effectively, maintaining robust performance across mixed and gender-specific datasets.

In contrast, MobileNetV2 shows significant challenges, with an average F-1 score of only 41%. This model struggles particularly with female dominant datasets and imbalanced scenarios, suggesting a sensitivity to dataset composition that impacts its ability to generalize. Similarly, InceptionV3, with an average F-1 score of 56%, shows moderate performance but faces difficulties in scenarios involving imbalanced gender datasets.

ResNet50, with an average F-1 score of 87%, performs better than MobileNetV2 and InceptionV3 but still shows some sensitivity to gender imbalances, particularly in female vs. male and balanced vs. female scenarios. These results highlight the importance of dataset balance and model architecture in achieving high F-1 scores and suggest that

37

some models may require additional tuning or data augmentation to mitigate gender bias effectively. Overall, the observed disparities underscore the need for careful consideration of model selection and dataset preparation in FER systems to ensure equitable performance across demographic groups.



*Figure 15 Deep learning model F-1 delta evaluation*

The study evaluates the performance and gender bias of five deep learning models—Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3—used in facial expression recognition (FER). The Sequential model shows a strong ability to mitigate bias, with an average F-1 score delta of 1%, indicating balanced predictions across gender datasets. However, it slightly favors male data in direct male vs. female comparisons, countered by a preference for female data in female vs. male scenarios. InceptionV3 exhibits significant gender bias, favoring female data with an average F-1 score delta of -10%, and struggles with male data, showing large discrepancies in male vs. male and female vs. male comparisons. DenseNet and ResNet50 display minor biases favoring female data, each with an average F-1 score delta of -2%, while MobileNet shows a slight bias towards male data with a 2% delta. These findings highlight the varying

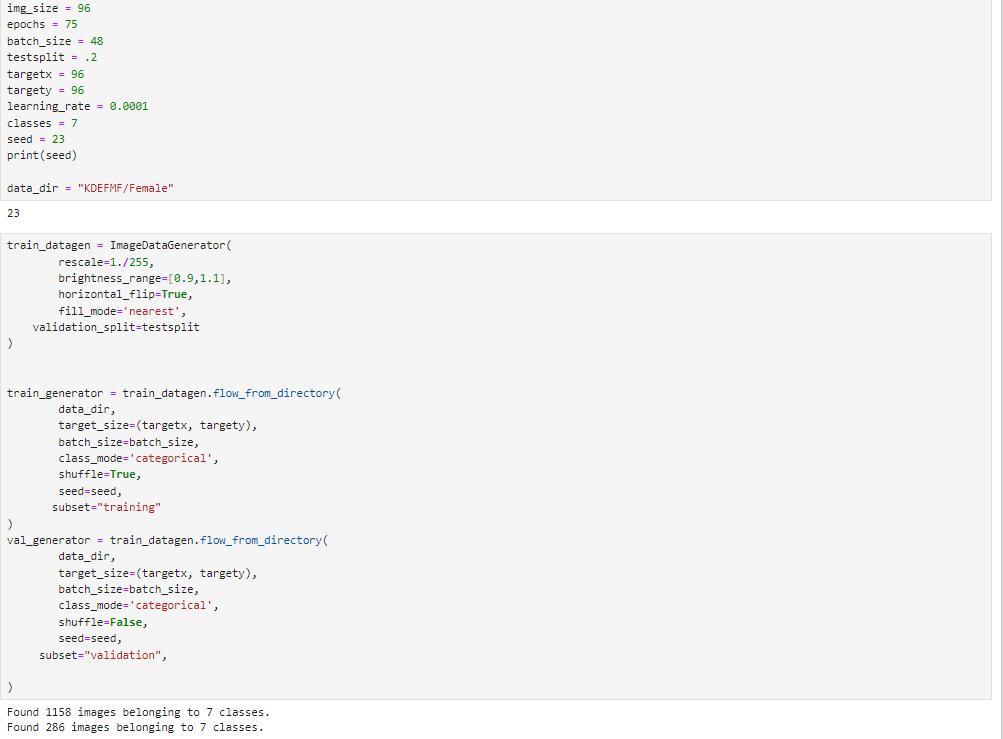
38

degrees of gender bias across models, emphasizing the need for careful model selection and dataset balancing to ensure fair and accurate FER. The study underscores the importance of continued research to address these biases and improve the fairness of FER systems. (Fig. 15)

**4.2.2 Gender Bias Across Balanced and Imbalanced Datasets**

Facial expression recognition (FER) systems have become integral in various sectors, including healthcare, security, and media, due to their ability to automatically detect and categorize human emotions. However, the effectiveness of these systems is often compromised by demographic biases, particularly gender bias, which can lead to unequal performance across different demographic groups. This bias is primarily rooted in the datasets used for training these models. Many FER datasets are not adequately balanced in terms of gender representation, resulting in models that perform better for the overrepresented gender. This imbalance can lead to skewed learning, where models become more adept at recognizing expressions from the dominant gender, while their performance degrades for the underrepresented gender. The research outlined in this document aims to explore the existence and extent of gender bias in FER systems by employing a range of deep learning algorithms trained on male-only, female-only, and mixed datasets. By systematically analyzing the performance of these models across gender-diverse datasets, the study seeks to uncover the extent of gender bias and provide insights into developing fairer and more accurate FER systems. Addressing gender bias is crucial for ensuring that FER technologies are used responsibly and do not perpetuate existing societal inequalities [14].

39

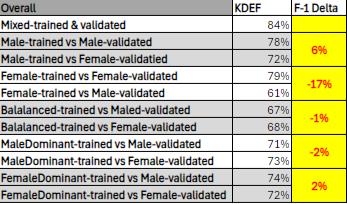


*Figure 16 KDEF dataset female subset being utilized for training and validation*

Figure 13 provides insights into training deep learning models for facial expression recognition (FER) using a female subset of data. It employs ImageDataGenerator for data augmentation, incorporating rescaling, brightness adjustments, and horizontal flipping to enhance data diversity and model generalization, which is particularly beneficial for limited datasets. A 20% validation split is used to assess the model's performance on unseen data, crucial for evaluating its generalization capabilities. By focusing on a female-only dataset, the model is trained to recognize expressions specific to this demographic, allowing for a detailed analysis of performance compared to models trained on male or mixed datasets, thus identifying gender-specific

40

biases or performance gaps. The dataset's class balance, with 1,158 training images and 286 validation images across seven classes, ensures effective learning without bias towards more frequent expressions. The training settings, including an image size of 96x96, a batch size of 48, and 75 epochs, are optimized for the learning process, with a small learning rate of 0.0001 to fine-tune model weights gradually. This setup provides a focused environment to explore how well a model can learn from female-specific data, helping to identify gender-related nuances in FER.



*Table 2: Overall F-1 Score’s for trained and validated models*

Table 2 provides insights into the performance of deep learning models in facial expression recognition (FER) across balanced and imbalanced datasets, highlighting the F-1 score differences (deltas) between various scenarios. Challenges arise with imbalanced datasets, as evidenced by the largest negative F-1 delta (-17%) in the female-trained vs. male-validated scenario, suggesting that models trained on female-only datasets perform significantly worse on male data. This highlights a potential gender-specific bias or insufficient representation within the training data. In contrast, the male-trained vs. male-validated scenario shows a 6% positive delta, implying that male-only

41

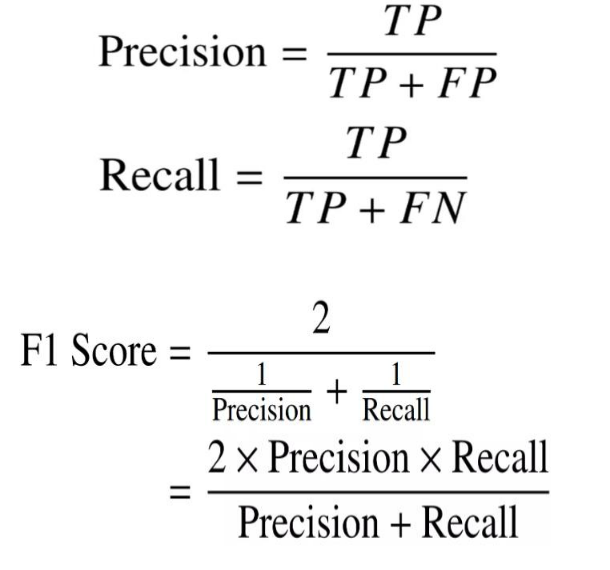
datasets may benefit from a slight performance boost, potentially due to better representation or model tuning. Balanced datasets, as seen in the balanced-trained vs. female-validated and balanced-trained vs. male-validated scenarios, shows a minimal delta (-1%), suggesting that balanced datasets help mitigate gender bias, providing more equitable performance across genders. The male-dominant scenarios (MD vs M and MD vs F) exhibits a small delta (-2%), indicating that training on diverse datasets with varied models can slightly impact performance but generally supports balanced outcomes. Overall, these insights underscore the importance of using balanced datasets to reduce gender bias and enhance model performance across demographic groups. They also highlight the need for careful dataset composition and model tuning to achieve equitable results in FER systems.

Along with measuring F-1 score, gain further insights into bias between balanced and imbalanced datasets the following metrics were used to further measure bias:

* **F-1 Score**: The F-1 score is a crucial metric used to evaluate the performance of facial expression recognition (FER) systems, particularly in assessing bias across different facial emotions. It is the harmonic mean of precision and recall, providing a balanced measure of a model's accuracy, especially in cases where the class distribution is imbalanced. Precision is calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP). Recall is calculated as the ratio of true positives (TP) to the sum of true positives and false negatives (FN). The F-1 score helps assess a model's accuracy by considering

42

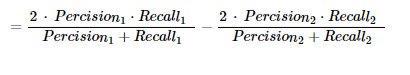
both false positives and false negatives, offering a comprehensive performance metric [35].



*Equation 1 Precision and recall Equations [35]*

*Equation 2 F-1 Score Equation [35]*

* **F-1 Score Difference/Delta:** F-1 Score delta is used in this experiment as a bias metric that allows comparison of performance across deep learning models and emotions for each gender. We calculate this as the difference in F-1 score averages across each validation scenario.



*Equation 3 F-1 Score Difference/Delta*

* **Accuracy Difference**: This metric measures the performance gap between different gender groups. A positive accuracy difference indicates a potential bias favoring slice 1 over slice 2, while a negative value indicates a potential bias in

43

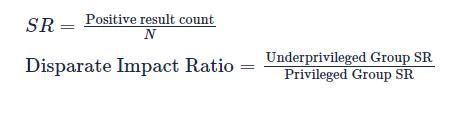
favoring slice 2 over slice 1. Zero value indicates no bias in between slice 1 and slice 2 [34].



*Equation 4 Accuracy Difference*

* **Disparate Impact**: This metric assesses fairness by comparing the success rate (SR) across different groups. Disparate impact became solidified from Title VII of the 1964 Civil Rights Act. The “80 percent” test was created by the State of

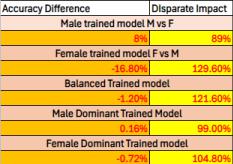
California Fair Employment Practice Commission (FEPC) in 1971 to determine if disparate impact exists. A value between 80% to 100% suggests no significant disparity, whereas values below 80% indicate potential bias against underprivileged group [35].



*Equation 5 Disparate Impact Ratio*

The analysis of gender bias in facial expression recognition (FER) systems reveals significant insights into the performance disparities across different training models in Table 3.

44



*Table 3 Accuracy Difference and Disparate impact chart [34][35]*

The Male Trained Model shows an 8% accuracy difference, indicating better performance on male data, with a disparate impact of 89%, suggesting no significant bias against females. Conversely, the Female Trained Model exhibits a substantial negative accuracy difference of -16.8%, reflecting poorer performance on male data, and a disparate impact of 129.6%, indicating a potential bias in favor of females. The Balanced Trained Model presents a small negative accuracy difference of -1.2%, suggesting more equitable performance across genders, yet the disparate impact of 121.6% indicates some residual bias. The Male Dominant Trained Model shows minimal accuracy difference (0.16%) and a disparate impact of 99%, suggesting near parity and fairness across genders. Meanwhile, the Female Dominant Trained Model displays a slight negative accuracy difference of -0.72%, with a disparate impact of 104.8%, indicating no significant bias. These findings underscore the challenges in achieving gender fairness in FER systems, highlighting that while balanced models tend to offer more equitable outcomes, some bias persists. Addressing these disparities is crucial for developing fair and effective FER technologies.

45

**4.2.3 Gender Bias Across Facial Emotions**

F-1 scores are a crucial metric in evaluating the performance of facial expression recognition (FER) systems, particularly when assessing bias across different facial emotions. The F-1 score, which is the harmonic mean of precision and recall, provides a balanced measure of a model's accuracy, especially in cases where the class distribution is imbalanced. In the context of FER, where gender bias is a significant concern, F-1 scores can be instrumental in identifying discrepancies in model performance across different demographic groups [36].

To measure bias using F-1 scores and F-1 score differences across gender groups, the study involves training and testing deep learning models on datasets that include male-only, female-only, and mixed-gender samples. By calculating the F-1 scores for each emotion category (such as happiness, sadness, anger, etc.) across these datasets, researchers can determine how well the models perform for each gender. A consistent F-1 score across genders would indicate that the model is fair and unbiased, while significant differences in F-1 scores would suggest the presence of gender bias [14][36].

Furthermore, by comparing the F-1 scores of models trained on balanced datasets versus those trained on imbalanced datasets, the study can assess the impact of dataset composition on model fairness. This approach allows for a comprehensive evaluation of how gender bias affects the recognition of specific emotions and helps identify which emotions are more susceptible to bias. Ultimately, using F-1 scores in this manner provides valuable insights into the fairness and accuracy of FER systems, guiding the development of more equitable and effective technologies. The following insights were made per each emotion with overall results from all deep learning models.

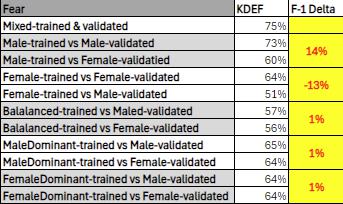
46

* Fear: Table 4 provides insights into the performance of deep learning models in recognizing the emotion of fear across different datasets, highlighting F-1 score differences (deltas). The highest F-1 score (75%) is observed when models are trained on both male and female data, indicating better overall performance for recognizing fear. In terms of imbalanced performance, the Male vs. Female (M vs
  1. scenario shows a significant positive F-1 delta (14%), suggesting that models trained on male data perform much better on male expressions of fear compared to female expressions. Conversely, the Female vs. Female (F vs F) scenario exhibits a significant negative F-1 delta (-13%), indicating that models trained on female-only data perform better on female expressions compared to male counterparts. Gender-specific challenges are evident in the Female vs. Male (F vs M) scenario, where the lowest F-1 score (51%) highlights poor performance in recognizing male expressions of fear using a model trained on female data. Models trained on balanced datasets show small positive F-1 deltas (1%) in both balanced scenarios (Bal vs M and Bal vs F), suggesting that balanced datasets help reduce bias, though some disparity remains. Additionally, both Male Dominant (MD) and Female Dominant (FD) effects show minimal positive deltas (1%) in the MD vs M and FD vs M scenarios, indicating that diverse and female-dominant models provide slight improvements in performance equity. Overall, these insights emphasize the importance of using diverse datasets to enhance the recognition of fear across genders. The table highlights the challenges in achieving consistent model

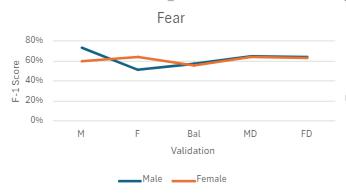
47

performance, particularly when models are trained on gender-imbalanced datasets.

Addressing these disparities is crucial for developing fair and reliable FER systems.



*Table 4 : F-1 Scores for Fear emotion*

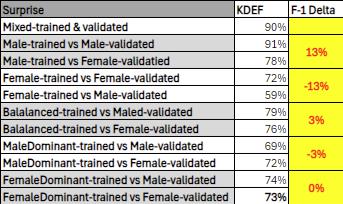


*Figure 17 Fear F-1 Scores across validation datasets*

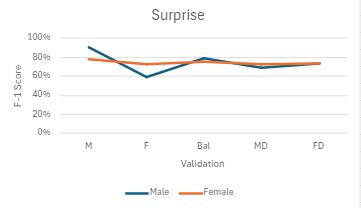
* Surprise: Table 5 provides insights into the performance of deep learning models in recognizing the emotion of surprise, focusing on F-1 score differences (deltas) across various datasets. Models trained on male data show a significant positive F-1 delta of 13% in the Male vs. Female (M vs F) scenario, indicating better performance on male expressions of surprise compared to female expressions. Conversely, the Female vs. Female (F vs F) scenario shows a negative F-1 delta of -13%, indicating better performance on female expressions with female-trained data. The Female vs. Male (F vs M) scenario records the lowest F-1 score at 59%,

48

highlighting challenges in recognizing male expressions with a female-trained model. Balanced datasets exhibit small positive F-1 deltas (3%) in the Bal vs M scenario, suggesting some reduction in bias, although disparities remain, as shown by a 76% F-1 score in Bal vs F. Male Dominant (MD) scenarios show mixed results, with a negative delta of -3% in MD vs M, indicating a slight drop in performance, while MD vs F maintains a 72% score. The Female Dominant (FD) scenario shows consistent performance with no delta (0%) and a 73% score in FD vs F and FD vs M. These findings underscore the importance of diverse datasets to achieve more equitable recognition of surprise across genders and highlight ongoing challenges in mitigating gender-specific biases in FER systems.



*Table 5: F-1 Scores for Surprise emotion*



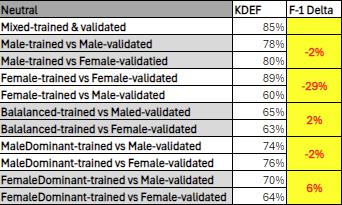
*Figure 18 Surprise F-1 Scores across validation datasets*

49

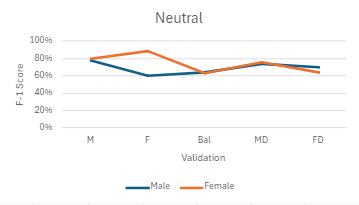
* Neutral: Table 6 provides insights into the performance of deep learning models in recognizing the neutral emotion, focusing on F-1 score differences (deltas) across various datasets. In the Male vs. Female (M vs F) scenario, a small negative F-1 delta of -2% is observed, suggesting slightly lower performance on female data when using male-trained models. Conversely, the Female vs. Female (F vs F) scenario shows a significant negative F-1 delta of -29%, indicating a substantial drop in performance on male expressions when using female-only data. The Female vs. Male (F vs M) scenario records a low F-1 score of 60%, highlighting challenges in recognizing male expressions with a female-trained model. Balanced datasets show minor positive F-1 deltas (2%) in the Bal vs M scenario, indicating some improvement in reducing bias, whereas the Bal vs F scenario remains lower at 63%. Male Dominant (MD) scenarios present consistent performance, with a -2% delta in MD vs M and a 76% score in MD vs F. The Female Dominant (FD) scenario shows a 6% positive delta in FD vs M, suggesting improved performance, while FD vs F maintains a 64% score. These insights emphasize the importance of diverse datasets to achieve fairer recognition of neutral expressions across genders and

50

highlight ongoing challenges in mitigating gender-specific biases within FER systems.



*Table 6: F-1 Socres for Neutral Emotion*

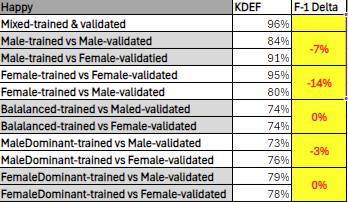


*Figure 19 Neutral F-1 Scores across validation datasets*

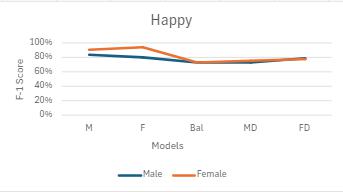
* Happy: Table 7 provides insights into the performance of deep learning models in recognizing the emotion of happiness, focusing on F-1 score differences (deltas) across various datasets. In the Male vs. Female (M vs F) scenario, a negative F-1 delta of -7% is observed, indicating reduced performance on male expressions when using male-trained models. The Female vs. Female (F vs F) scenario shows a significant negative F-1 delta of -14%, highlighting a notable drop in performance on male expressions. Conversely, the Female vs. Male (F vs M) scenario records a

51

lower F-1 score of 80%, indicating slight difficulty in recognizing male expressions with a female-trained model. Balanced datasets exhibit no F-1 delta (0%) in Bal vs M and Bal vs F scenarios, suggesting consistent performance across genders. Male Dominant (MD) scenarios show a slight negative delta of -3% in MD vs M, while MD vs F maintains a 76% score. The Female Dominant (FD) scenarios display zero deltas in both FD vs M with a 79% score in FD vs F, indicating stable performance. These insights underscore the effectiveness of diverse datasets in achieving equitable recognition of happiness across genders, while also highlighting persistent challenges in addressing gender-specific biases in FER systems.



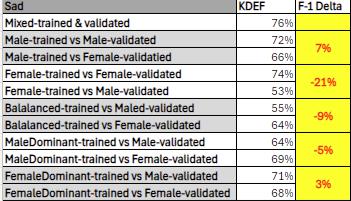
*Table 7 : F-1 Scores for Happy emotion*



*Figure 20 Happy F-1 Scores across validation datasets*

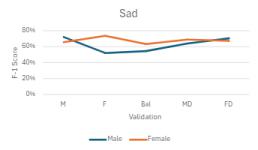
52

* Sad: Table 8 provides insights into the performance of deep learning models in recognizing the emotion of sadness, focusing on F-1 score differences (deltas) across various datasets. The Male vs. Female (M vs F) scenario shows a positive F-1 delta of 7%, indicating better recognition of male expressions of sadness with male-trained models. In contrast, the Female vs. Female (F vs F) scenario exhibits a significant negative F-1 delta of -21%, highlighting a substantial drop in performance on male expressions. The Female vs. Male (F vs M) scenario records a low F-1 score of 53%, illustrating challenges in recognizing male expressions with a female-trained model. Balanced datasets show a negative F-1 delta of -9% in the Bal vs M scenario, indicating reduced effectiveness, while Bal vs F achieves a 64% score. Male Dominant (MD) scenarios present a negative delta of -5% in MD vs M, with MD vs F maintaining a 69% score. The Female Dominant (FD) scenario shows a positive delta of 3% in FD vs M, indicating improved performance, with FD vs F achieving a 68% score. These insights underscore the importance of diverse datasets for better recognition of sadness across genders and highlight persistent challenges in mitigating gender-specific biases in FER systems.



*Table 8 : F-1 Score for Sad Emotion*

53

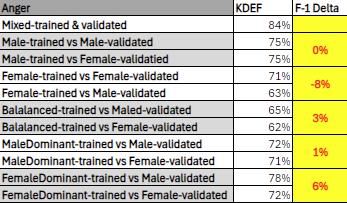


*Figure 21 Sad F-1 Scores across validation datasets*

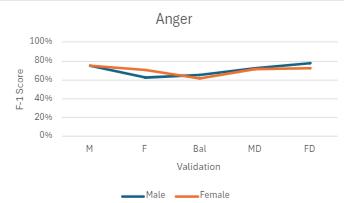
* Anger: Table 9 provides insights into the performance of deep learning models in recognizing the emotion of anger, focusing on F-1 score differences (deltas) across various datasets. The mixed dataset achieves the highest F-1 score at 84%, indicating effective performance when models are trained on both genders. In the Male vs. Female (M vs F) scenario, there is no F-1 delta, suggesting balanced performance between genders. However, the Female vs. Female (F vs F) scenario shows a negative F-1 delta of -8%, indicating decreased performance on male expressions. The Female vs. Male (F vs M) scenario records a lower F-1 score of 63%, highlighting challenges in recognizing male expressions with a female-trained model. Balanced datasets present a positive F-1 delta of 3%, suggesting decreased bias, while Bal vs F scores 62%. Male Dominant (MD) scenarios show a slight positive delta of 1% in MD vs M, with MD vs F maintaining a 71% score. The Female Dominant (FD) scenario displays a positive delta of 6% in FD vs M, indicating better performance, and a 72% score in FD vs F. These insights underscore the importance of mixed datasets for achieving equitable recognition of

54

anger across genders and highlight ongoing challenges in addressing gender-specific biases in FER systems.



*Table 9: F-1 Score for Anger emotion*

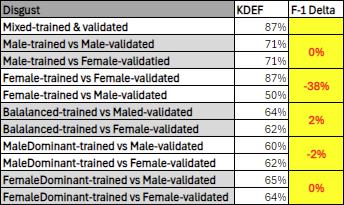


*Figure 22 Anger F-1 Scores across validation datasets*

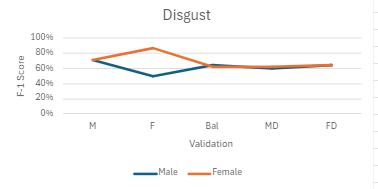
* Disgust: Table 10 provides insights into the performance of deep learning models in recognizing the emotion of disgust, focusing on F-1 score differences (deltas) across various datasets. In the Male vs. Female (M vs F) scenario, there is no F-1 delta, indicating balanced performance between genders. However, the Female vs. Female (F vs F) scenario displays a significant negative F-1 delta of -38%, highlighting a substantial drop in performance on male expressions. The Female vs. Male (F vs M) scenario records a low F-1 score of 50%, illustrating difficulties

55

in recognizing male expressions with a female-trained model. Balanced datasets show a positive F-1 delta of 2% in the Bal vs M scenario, indicating slight improvement, while Bal vs F scores 62%. Male Dominant (MD) scenarios exhibit a negative delta of -2% in MD vs M, with MD vs F maintaining a 62% score. The Female Dominant (FD) scenario shows no delta in FD vs M and a 64% score in FD vs F, indicating consistent performance. These insights emphasize the importance of diverse datasets for better recognition of disgust across genders and highlight persistent challenges in mitigating gender-specific biases in FER systems.



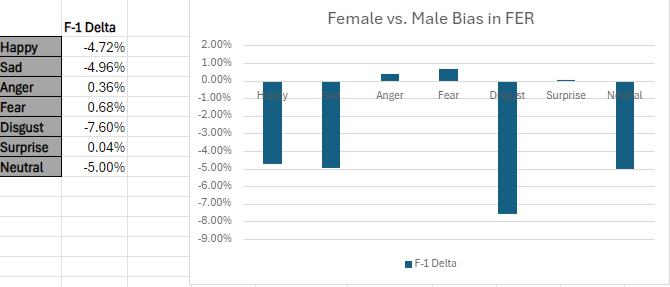
*Table 10: F-1 Score for Disgust emotion*



*Figure 23 Disgust F-1 Scores across validation datasets*

56

**4.3 Discussion**



*Figure 24 Gender bias on Emotion across DL models*

The analysis of gender bias in facial expression recognition (FER) systems highlights that the emotion "Disgust" is particularly vulnerable to gender bias, with an average F-1 score delta of -7.6% across various datasets and deep learning models. This indicates a significant discrepancy in the model's ability to recognize disgust in male versus female expressions. The detailed examination of the data reveals that in scenarios where models are trained on female data, there is a substantial drop in performance when recognizing male expressions of disgust, as evidenced by a negative F-1 delta of -38% in the Female vs. Female (F vs F) scenario. This suggests that models trained predominantly on female expressions struggle to generalize to male expressions, highlighting a critical area for improvement in FER systems.

In addition to "Disgust," other emotions such as "Happy," "Sad," and "Neutral" also exhibit notable gender biases, with average F-1 score deltas nearing -5%. These biases consistently favor female facial image data, suggesting that the models are more adept at recognizing these emotions in females than in males. For instance, the emotion

57

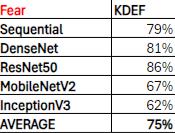
"Happy" shows a negative F-1 delta of -7% in the Male vs. Female (M vs F) scenario, indicating reduced performance on male expressions when using male-trained models. Similarly, "Sad" and "Neutral" emotions display significant negative deltas in scenarios where models are trained on female data, further underscoring the challenges in accurately predicting these emotions in male subjects.

These findings emphasize the critical need for improved model training and dataset balancing to address gender disparities in emotion recognition systems. The use of balanced datasets has been shown to mitigate gender bias, as they provide more equitable performance across genders. By ensuring that training datasets are representative of both male and female expressions, FER systems can achieve more consistent and fair outcomes. This approach not only enhances the accuracy of emotion recognition but also contributes to the development of fairer and more reliable FER technologies, which is essential for applications in sectors such as healthcare, security, and media. Addressing these biases is crucial for ensuring that FER systems which do not perpetuate existing societal inequalities and can be used responsibly in decision-making processes. (Fig. 24)

Several key initial insights were observed when conducting our experimentation, including the lowest and highest performing emotions for the Balanced, Female, and

58

Male KDEF subsets. Beginning with the experimentation conducted on mixed datasets, we observed that fear was the emotion resulting in the lowest F-1 score overall.

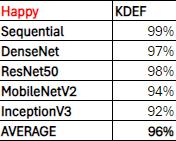


*Table 11 F-1 Score of Fear for Balanced Dataset*

It is possible fear is a much more difficult emotion to predict, particularly when using mixed or balanced datasets, due to the subtlety and complexity of its expression. Unlike more overt emotions such as happiness or anger, fear can manifest through nuanced facial cues that are less pronounced and more difficult for deep learning models to detect

1. This subtlety requires models to discern slight variations in facial features, which can be challenging, especially when the dataset lacks sufficient examples of fear expressions. Additionally, the complexity of fear expressions involves a combination of facial features that may vary significantly between individuals, adding another layer of difficulty for models to generalize effectively. This complexity is compounded by the need for models to accurately interpret these features across diverse demographic groups, as highlighted in the thesis, which emphasizes the importance of addressing gender bias and ensuring fair recognition across different populations [14]. The intricate nature of fear expressions, combined with the subtlety of its cues, underscores the necessity for well-annotated and balanced datasets that can provide the depth and variety needed for models to learn and accurately recognize this emotion.

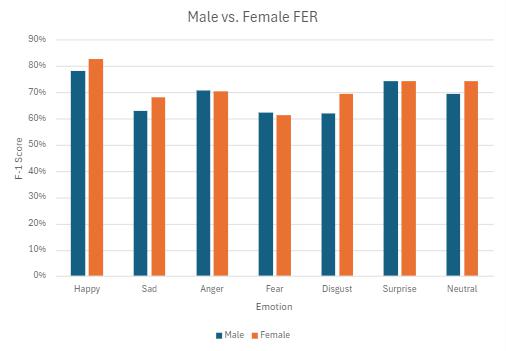
59



*Table 12 F-1 scores for Happiness using Mixed or balanced dataset*

The Highest performing emotion when training using a mixed or balanced dataset was happiness. Happiness is often the highest-performing emotion in facial expression recognition tasks due to its distinct features and consistency across individuals [1]. The expression of happiness is characterized by clear and recognizable facial features, such as a broad smile and raised cheeks, which are easily detectable by deep learning models [14]. These features are not only pronounced but also consistent across different individuals, reducing variability and aiding models in learning the associated patterns effectively. This consistency means that models can generalize well across diverse demographic groups, as the expression of happiness does not vary significantly from person to person. All models achieve high accuracy, with Sequential reaching 99%, DenseNet at 97%, and ResNet50 at 98%. These high scores reflect the models' ability to easily detect the clear and recognizable features of happiness, such as broad smiles and raised cheeks, which are consistently expressed across individuals.

60



*Figure 25 Male vs Female F-1 Score Bar Chart*

Figure 22 compares the performance of facial emotion recognition (FER) systems for male and female emotions across various categories using the F1 score as a metric. For females, the highest performing emotion is "Happy," with an average F1 score slightly above 80%, indicating that the system is highly effective at recognizing happiness in female expressions. This high score could be attributed to the universal and distinct nature of happy expressions [17], which are typically characterized by obvious features such as smiling. On the other hand, the lowest performing emotion for females is "Fear," with an average F1 score around 65%. The complexity and subtlety involved in expressing fear, which often shares facial characteristics with surprise, might contribute to this lower recognition accuracy [8].

For males, similar patterns are observed, with "Happy" also being the highest performing emotion, scoring just above 80%. This suggests that happiness is a universally

61

recognizable expression, regardless of gender, likely due to its clear and distinct facial indicators. Conversely, the lowest performing emotion for males is "Sad," with an average F1 score around 70%. The relatively lower performance in recognizing sadness could be due to the more nuanced and less pronounced facial cues associated with sadness in men, which might be less overt compared to emotions like happiness. This discrepancy highlights potential areas for improvement in FER systems, emphasizing the need for enhanced sensitivity to subtle emotional cues across genders [8].

Female F-1 scores outperform males overall. In the emotions “Happy” and “Sad,” females often exhibit more expressive facial cues, which can enhance the performance of FER systems in recognizing happiness. Cultural and social norms frequently encourage women to display their emotions more openly, leading to more pronounced and varied expressions. These expressive cues, such as broader smiles and more active facial engagement, can be easier for recognition systems to detect and classify accurately [8]. Additionally, there may be inherent differences in how happiness is expressed through facial features between genders. Female facial expressions might align more closely with the patterns that FER systems are trained to recognize, such as the curvature of smiles and eye movements [7]. Together, these factors of expressiveness and distinct facial features may contribute to the higher performance observed in recognizing happiness in females compared to males.

62

**CHAPTER 5 CONCLUSION & FUTURE**

**WORK**

This research has made significant strides in understanding and addressing gender bias in facial expression recognition (FER) systems. By employing a range of deep learning models, including Sequential, ResNet-50, DenseNet, MobileNetV2, and InceptionV3, we have systematically evaluated the performance and fairness of these algorithms across diverse datasets. The findings underscore the existence of gender bias in FER systems, highlighting the need for more equitable machine learning models that perform consistently across different demographic groups. Through comprehensive experimentation and analysis, this study provides actionable insights for developing fairer and more accurate FER technologies. The implications of this research extend beyond the technical domain, contributing to the broader discourse on fairness and ethics in artificial intelligence.

The experimental results demonstrate that using well-balanced datasets significantly reduces gender bias in FER models. Balanced datasets, which equally represent male and female subjects, help mitigate the skewed learning that often occurs when models are trained on imbalanced data. This approach ensures more equitable performance across genders, as evidenced by minimal F-1 score differences in balanced scenarios. The study highlights the importance of dataset composition in achieving fair and accurate FER systems, emphasizing that balanced datasets are crucial for reducing bias and enhancing model performance across demographic groups.

Building on the insights gained from this study, future research should focus on several key areas to further advance the field of FER. Firstly, expanding the scope of datasets to include a wider range of demographic variables, such as age and ethnicity, could

63

provide a more holistic understanding of bias in FER systems. Additionally, exploring the integration of ensemble learning techniques may enhance model robustness and mitigate bias more effectively. Future studies should also investigate the real-world application of these models in various industries, ensuring that ethical considerations are prioritized in the deployment of FER technologies. Finally, ongoing collaboration with interdisciplinary teams, including ethicists and social scientists, will be crucial in developing FER systems that are not only technically sound but also socially responsible.

The document provides compelling evidence that balanced datasets significantly reduce bias in facial expression recognition (FER) systems.

Firstly, the study reveals that balanced datasets result in minimal F-1 score differences, indicating more equitable performance across genders. For instance, in scenarios comparing balanced datasets with female (Bal vs F) and male (Bal vs M) data, the F-1 score delta is -1% while F-1 score deltas for male-only trained model and female-only trained model are 6% and -17% respectively. This suggests that balanced datasets effectively mitigate gender bias, ensuring consistent performance across different demographic groups. Secondly, the balanced trained model shows a small negative accuracy difference of -1.2%, which points to more equitable performance across genders. Although the disparate impact is 121.6%, indicating some residual bias, the balanced dataset still achieves closer parity compared to other training models, highlighting its effectiveness in reducing bias. Lastly, in the context of emotion recognition, particularly for happiness, balanced datasets exhibit no F-1 delta (0%) in Bal vs M and Bal vs F scenarios. This consistency across emotions further supports the argument that balanced datasets contribute to fairer recognition of emotions across diverse demographic groups.

64

Collectively, these findings underscore the critical role of balanced datasets in reducing gender bias in FER systems, leading to more equitable and accurate outcomes. Other key insights were made beyond the study of gender bias, when exploring the resulting F-1 scores.

Future research in facial expression recognition (FER) systems could focus on several key areas to enhance understanding and reducing gender bias as well as other demographic biases:

* **Utilizing Additional Datasets**: Expanding research to include datasets such as CK+ and FER2013 could strengthen the findings. These datasets provide diverse demographic compositions and a range of facial expressions, crucial for evaluating model generalizability. Training and testing models on these datasets would help researchers understand the impact of dataset composition on gender bias, leading to more accurate and fair FER systems [1][16].
* **Incorporating Additional Bias Metrics**: Introducing metrics like equalized odds could offer a more comprehensive evaluation of fairness in FER systems. Equalized odds assess whether true positive and false positive rates are consistent across demographic groups, providing a nuanced view of model fairness. By using this metric, researchers can better identify and mitigate biases, ensuring equitable performance across genders [35].
* **Utilizing Ensemble Learning:** Exploring the integration of ensemble learning techniques may enhance model robustness and mitigate bias more effectively.

65

* **Investigating Other Biases:** In addition to gender bias, investigating other biases due to demographic attributes such as race and ethnicity are considered among the potential future work.
* **Mitigating Bias:** This research study can be extended to apply techniques to effectively mitigate gender bias besides extensive experimentation with various datasets and deep leaning models to investigate gender bias in facial expression recognition.

These future research directions aim to deepen the understanding of gender bias in FER systems and contribute to the development of more equitable and effective technologies.

66

**Bibliography**

[1] Kumar, Abhaya. “Facial Expression Recognition.” Medium, Medium, 16 Nov. 2019, medium.com/@abhayakumar169/facial-expression-recognition-cf6a61f8045f.

1. Prado, K. S. do. (2018, February 3). Face recognition: Understanding LBPH algorithm. Medium. <https://towardsdatascience.com/face-recognition-how-lbph-works-90ec258c3d6b>
2. Rosebrock, A. (2021, April 17). Local Binary Patterns with Python & OpenCV - PyImageSearch. PyImageSearch. <https://pyimagesearch.com/2015/12/07/local-binary-patterns-with-python-opencv/>
3. Sharma, H. (2019, September 30). Facial Expressions Recognition - Hina Sharma - Medium. Medium. <https://medium.com/@hinasharma19se/facial-expressions-recognition-b022318d842a>
4. V. Jacintha, J. Simon, S. Tamilarasu, R. Thamizhmani, K. Thanga yogesh and J.

Nagarajan, "A Review on Facial Emotion Recognition Techniques," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0517-0521, doi: 10.1109/ICCSP.2019.8698067. keywords: {Face;Emotion recognition;Feature extraction;Eyebrows;Skin;Support vector machines;Task analysis;Emotion recognition;FACS;JAFFE;Min-Max classifier;KNN;SVM},

1. Ko BC. A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors. 2018; 18(2):401. <https://doi.org/10.3390/s18020401>
2. Fan, Y., Lam, J.C.K. & Li, V.O.K. Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness. Sci Rep 11, 5214 (2021). <https://doi.org/10.1038/s41598-021-84632-9>

67

1. Manresa-Yee, C., Ramis, S., & Buades, J. M. (2023). Analysis of gender differences in facial expression recognition based on deep learning using explainable artificial intelligence. International Journal of Interactive Multimedia and Artificial Intelligence, In Press(In Press), 1. <https://doi.org/10.9781/ijimai.2023.04.003>
2. Vemou, K. (2021). Facial Emotion Recognition. EDPSTechDispatchonFacialEmotionRecognition, Issue 1, 1-5. doi:10.2804/519064
3. Manresa-Yee, C., Guarinos, S. R., & Buades Rubio, J. M. (2022). Facial Expression Recognition: Impact of Gender on Fairness and Expressions. In XXII International Conference on Human Computer Interaction (Interaccion 2022), September 07-09, 2022, Teruel, Spain. ACM, New York, NY, USA, 8

pages. <https://doi.org/10.1145/3549865.3549904>

1. Xu, T., White, J., Kalkan, S., & Gunes, H. (2022). Investigating bias and fairness in facial expression recognition. Department of Computer Science and Technology, University of Cambridge; Department of Computer Engineering, Middle East Technical University.
2. Domnich, Artem & Anbarjafari, Gholamreza. (2021). Responsible AI: Gender bias assessment in emotion recognition.
3. Chen, Y., & Joo, J. (Year). Understanding and mitigating annotation bias in facial expression recognition. University of California, Los Angeles.
4. S. Li and W. Deng, "A Deeper Look at Facial Expression Dataset Bias," in IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 881-893, 1 April-June 2022, doi: 10.1109/TAFFC.2020.2973158. keywords: {Face recognition;Databases;Adaptation models;Training;Semantics;Data models;Task analysis;Cross dataset;facial expression recognition (FER);dataset bias;domain adaption},

68

1. Raina, R., Monares, M., Xu, M., Fabi, S., Xu, X., Li, L., Sumerfield, W., Gan, J., & de Sa, V. R. (2022). Exploring biases in facial expression analysis. In NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research. Retrieved

from <https://openreview.net/forum?id=jrddoq9NDP6>

1. Shape and Texture Aware Facial Expression Recognition Using Spatial Pyramid

Zernike Moments and Law’s Textures Feature Set - Scientific Figure on ResearchGate. Available from: <https://www.researchgate.net/figure/Sample-images-a-the-emotions-of-an-individual-KDEF-database-and-b-five-different_fig2_350508917>[accessed 8 Nov 2024]

1. Nursel Yalçin , Muthana Alisawi ,(2024). Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques.

Heliyon. [https://www.sciencedirect.com/science/article/pii/S24058440241494444.](https://www.sciencedirect.com/science/article/pii/S24058440241494444)

1. Christopher, A. (2024a, August 15). Exploring the different architectures of Deep Learning. Medium. <https://medium.com/dataseries/exploring-the-different-architectures-of-deep-learning-abc5eabafb8d>
2. Convolutional Neural Networks with Transfer Learning for Pneumonia Detection - Scientific Figure on ResearchGate. Available

from: <https://www.researchgate.net/figure/Architecture-of-Inception-v3_fig3_364173580>[accessed 11 Nov 2024]

1. Efficient Approach towards Detection and Identification of Copy Move and Image Splicing Forgeries Using Mask R-CNN with MobileNet V1 - Scientific Figure on ResearchGate. Available from: <https://www.researchgate.net/figure/The-architecture-of-MobileNet-V1-30_fig5_357623745>[accessed 11 Nov 2024]

69

1. Using Deep Learning to Identify High-Risk Patients with Heart Failure with Reduced Ejection Fraction - Scientific Figure on ResearchGate. Available

from: <https://www.researchgate.net/figure/Deep-Learning-Model-Design-Sequential-Model-Architecture_fig2_353566729>[accessed 11 Nov 2024]

1. IBM developer. (n.d.). <https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/>
2. Chaure, N. (2024, May 16). Variants of resnet: A comparative analysis. Medium. <https://medium.com/@nayanchaure601/variants-of-resnet-a-comparative-analysis-63fdc1573b34>
3. Arora, A. (2020, August 2). DenseNet architecture explained with pytorch implementation from torchvision. index. <https://amaarora.github.io/posts/2020-08-02-densenets.html>
4. Ensemble-based face expression recognition approach for image sentiment analysis - Scientific Figure on ResearchGate. Available

from: <https://www.researchgate.net/figure/Block-diagram-of-the-proposed-FER-model_fig1_361008368>[accessed 13 Nov 2024]

1. Chhua, K., Went, Z., Hathalia, V., Zhu, K., & O'Brien, S. (2024). From Bias to Balance: Detecting Facial Expression Recognition Biases in Large Multimodal Foundation Models. Retrieved from [https://kvjvhub.github.io/FERRacialBias/.](https://kvjvhub.github.io/FERRacialBias/)
2. Anaconda documentation. Anaconda Documentation - Anaconda documentation. (n.d.). <https://docs.anaconda.com/#capabilities>
3. Real Time Emotion Recognition from Facial Expressions Using CNN Architecture - Scientific Figure on ResearchGate. Available

70

from: <https://www.researchgate.net/figure/Example-of-images-from-KDEF-facial-expression-dataset_fig2_336287978>[accessed 14 Nov 2024]

1. Hosseini, M. P., Lu, S., Kamaraj, K., Slowikowski, A., & Venkatesh, H. C. (2020). Deep learning architectures. Deep learning: concepts and architectures, 1-24.
2. Deep learning using rectified linear units (ReLU). (n.d.-

b). <https://arxiv.org/pdf/1803.08375>

1. Adam: A method for stochastic optimization. (n.d.-a). <https://arxiv.org/pdf/1412.6980>
2. Google. (n.d.). Model bias metrics for vertex AI | google cloud.

Google. <https://cloud.google.com/vertex-ai/docs/evaluation/model-bias-metrics>

1. F1 score in Machine Learning: Intro & Calculation. V7. (n.d.). <https://www.v7labs.com/blog/f1-score-guide>
2. *Top bias metrics and how they work*. RSS. (n.d.). https://www.monitaur.ai/blog-posts/top-bias-metrics-and-how-they-work
3. Mustafa Atay, Megh Poudyel and Saul Evora, 'Evaluation of Gender Bias in Masked Face Recognition with Deep Learning Models', 2023 IEEE Symposium Series on Computational Intelligence (IEEE SSCI), Mexico City, Mexico, December 5-8, 2023. <https://ieeexplore.ieee.org/document/10371811>
4. Mustafa Atay, Hailey Gipson, Tony Gwyn and Kaushik Roy, Evaluation of Gender Bias in Facial Recognition with Traditional Machine Learning Algorithms, IEEE Symposium Series on Computational Intelligence (IEEE

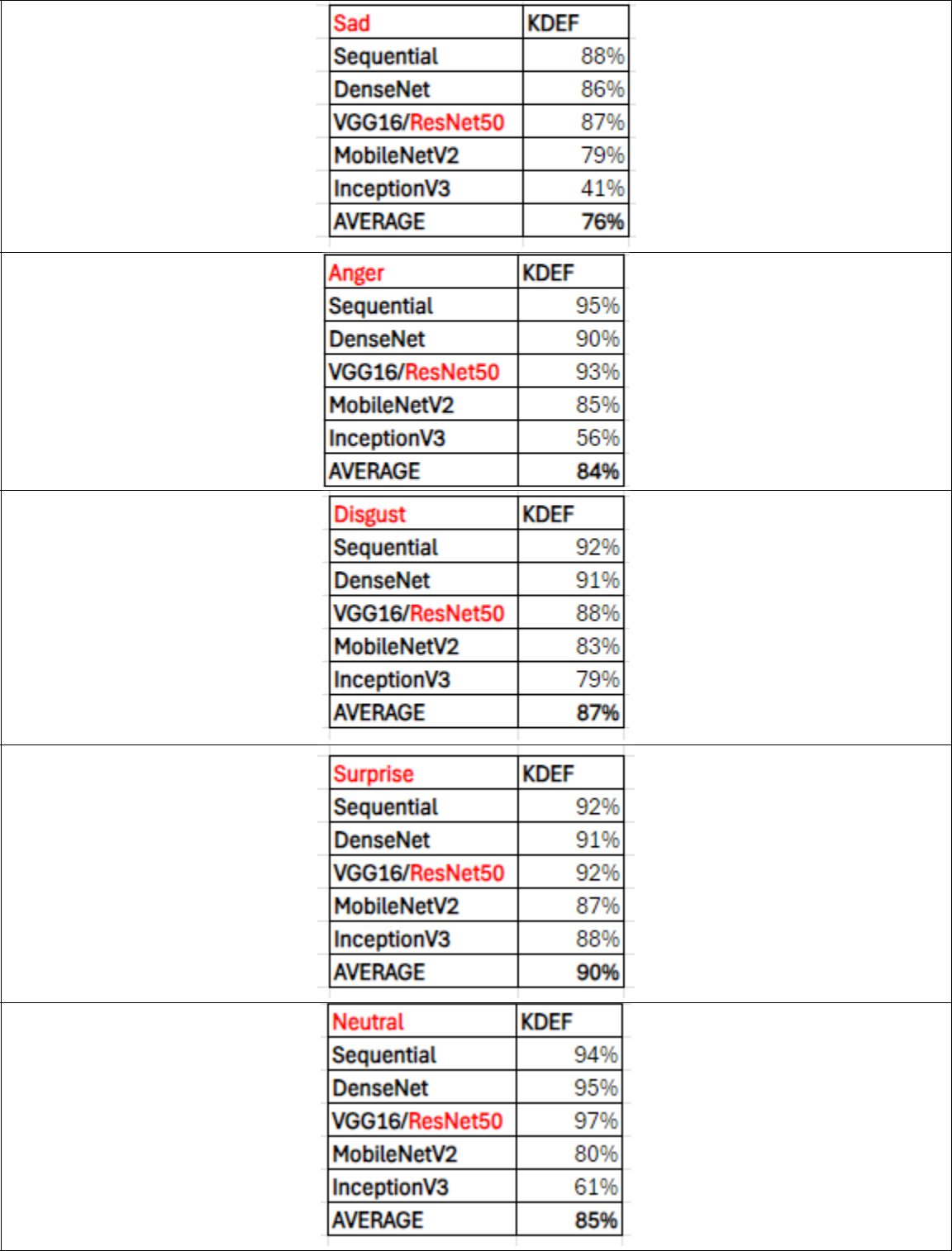
71

SSCI'21), December 4-7, 2021, Orlando,

FL. <https://doi.org/10.1109/SSCI50451.2021.9660186>

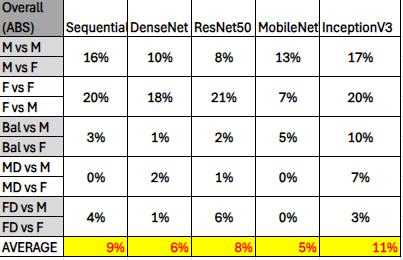
72

**APPENDIX A: MIXED DATASET TRAINED MODELS F-1 SCORES**



73

**APPENDIX B: ALL MODELS ABSOLUTE VALUE F-1 DELTA**



74

**APPENDIX C: ALL EMOTIONS ABSOLUTE VALUE F-1 DELTAS**

