RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ÉCOLE NATIONALE POLYTECHNIQUE

Département D'Électronique

End-of-study project dissertation

for obtaining the State Engineer's degree in Electronics

BMI Estimation From Face Images

BENREKIA Yaseen & DJEMMAH Imad Eddine

Presented and defended publicly on (07/07/2025)

Composition of the jury:

President:	Pr. Hicham BOUSBIA-SALAH
Examiner:	Mr. Mohamed Ousaid TAGHI
Supervisors:	Pr. Sid Ahmed BERRANI
	Dr. Nesrine BOUADJENEK

ENP 2025

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ÉCOLE NATIONALE POLYTECHNIQUE

Département D'Électronique

End-of-study project dissertation

for obtaining the State Engineer's degree in Electronics

BMI Estimation From Face Images

BENRESKIA Yaseen & DJEMMAH Imad Eddine

Presented and defended publicly on (07/07/2025)

Composition of the jury:

President:	Pr. Hicham BOUSBIA-SALAH
Examiner:	Mr. Mohamed Ousaid TAGHI
Supervisors:	Pr. Sid Ahmed BERRANI
	Dr. Nesrine BOUADJENEK

ENP 2025

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ÉCOLE NATIONALE POLYTECHNIQUE

Département d'Électronique

Mémoire de fin d'études

pour l'obtention du diplôme d'Ingénieur d'État en Électronique

Estimation de l'IMC à partir d'images de visage

BENREKIA Yaseen & DJEMMAH Imad Eddine

Présenté et soutenu publiquement le (07/07/2025)

Composition du jury :

Président :	Pr. Hicham BOUSBIA-SALAH
Examinateur :	Mr. Mohamed Ousaid TAGHI
Encadreurs :	Pr. Sid Ahmed BERRANI
	Dr. Nesrine BOUADJENEK

ENP 2025

الملخص:

يقدم هذا المشروع دراسة شاملة حول تقدير مؤشر كتلة الجسم (BMI) اعتمادًا على صور الوجه، مع التركيز بشكل خاص على تحسين الأداء التنبؤي. من خلال الاستفادة من أحدث تقنيات التعلم العميق، وتقنيات المعالجة المتقدمة للصور، وطرق استخراج الخصائص القوية، يحقق النهج المقترح تحسينات ملحوظة في دقة التقدير. وتُظهر النتائج التجريبية فعالية النموذج المحسن على مجموعات بيانات مرجعية، مما يبرز إمكانيات تحليل الوجه كأداة عملية للتقييم الآلي لمؤشر كتلة الجسم في مختلف تطبيقات الرعاية الصحية والرفاهية.

الكلمات المفتاحية:

مؤشر كتلة الجسم – الوجه – استخراج الخصائص – معالجة الصور – الانحدار – الصحة – التقدير التلقائي

Résumé

Ce projet présente une étude approfondie sur l'estimation de l'IMC à partir d'images de visage, en mettant l'accent sur l'amélioration des performances prédictives. En s'appuyant sur des architectures d'apprentissage profond de pointe, des techniques avancées de prétraitement d'images et des méthodes robustes d'extraction de caractéristiques, l'approche proposée améliore significativement la précision de l'estimation. Les résultats expérimentaux démontrent l'efficacité du modèle optimisé sur des jeux de données de référence, soulignant le potentiel de l'analyse faciale comme outil pratique pour l'évaluation automatisée de l'IMC dans diverses applications de santé et de bien-être.

Mots clés : IMC – Visage – Deep Learning – Apprentissage profond – Traitement d'images – Régression – Santé – Estimation automatique.

Abstract

This project presents a comprehensive study on BMI estimation from face images, with a particular emphasis on improving predictive performance. By leveraging state-of-the-art deep learning architectures, advanced image preprocessing techniques, and robust feature extraction methods, the proposed approach achieves significant improvements in estimation accuracy. Experimental results demonstrate the effectiveness of the optimized model on benchmark datasets, highlighting the potential of facial analysis as a practical tool for automated BMI assessment in various healthcare and wellness applications.

Keywords: BMI – Face – Feature Extraction – Image Pre-rocessing – Regression – Health – Automatic Estimation.

Acknowledgments 1 We would first like to thank our supervisors Mr. Sid-Ahmed Berrani, Mme. Nesrine Bouadjenek who have helped us along the way. We sincerely appreciate their presence, their useful tips, unwavering support and insightful ideas, We could not have asked for more from a mentor. We would also like to thank the jury members for taking their time and accepting to evaluate my long lasting work: Mr. Mohamed Ousaid TAGHI, and Mr.Hicham BOUSBIA-SALAH, our professors at the Ecole Nationale Polytechnique. We would like to extend my gratitude for every professor who helped us during my five years at the school, what we learned from them is invaluable. We would most certainly like to thank our supervisors and the head of the industrial engineering department for providing us with the necessary equipment so that we could carry this project properly. We would also give a very big to my family my Mom , My dad my sisters to tofik and rabah also . We would also give a very big thank also to all the saudi team that I worked with them and I hev learned from them also as: Bassim , Ahmed , Fahd ... and also Moammer . We would also like to thank all my colleagues at the school, what we shared together can never be replaced. To my friends, the Lycée : Ahmed Takieddine , Ameur "hadarate" , Mehdi , Fateh . "get it from the rebound", Mohamed "Browni", Moh "Englizi", Ousamma "9bayli" , and Naoufel "nutella", we have shared some great memories together, and I am glad I found people with whom I can be myself fully, , without forgetting Abdou and Ghani. To my prepa friends: Abdou lachlah , nasro , zaki , separated by continents, united by our love to each-other and to l'annexe, and our hate on prepa. To ELN, from my promotion "9rib ingénieurs", to Walid "la starte" and Lounes, we have shared the best and worst moments of our lives so far together, from happy memories to trying to calm me, I could not have asked for more. To my fellow IEEE committee members and IEEE crew, we might have had strokes trying to push forward, but we made the most out of it as we created some unforgettable moments along the way. To the whole reason I was in this school and the reason I chose ELN in the first place, my math teacher Kherraz and pr.Berroudji, you have set me into this path and gave me confidence during my choices , thank also to my friend TOUIL Mohamed Reda Finally, we would like to thank everyone that has helped our during my academic curriculum, even with a simple word.

Imad.

Acknowledgments 2 I dedicate this work to my family, who supported me throughout this journey. Their love, encouragement, sacrifices, and unconditional support have been the driving forces behind my success. I want to express my gratitude to all my friends and everyone who helped me, whether through small gestures or words that made me happy. I am grateful to every person who offered me experiences or life lessons, helping me grow as an individual. Thank you all, and big thanks also to my friend TOUIL Mohamed Reda.

Yaseen.

Thanks *We would first like to thank our supervisors Mr. Sid-Ahmed Berrani, Mme. Nesrine Bouadjenek who have helped us along the way. We sincerely appreciate their presence, their useful tips, unwavering support and insightful ideas, We could not have asked for more from a mentor. We would also like to thank the jury members for taking their time and accepting to evaluate my long lasting work: Mr. Mohamed Ousaid TAGHI, and Mr. Hicham BOUSBIA-SALAH, our professors at the Ecole Nationale Polytechnique. We would like to extend my gratitude for every professor who helped us during my five years at the school, what we learned from them is invaluable. We would most certainly like to thank our supervisors and the head of the industrial engineering department for providing us with the necessary equipment so that we could carry this project properly. We would also like to thank all my colleagues at the school, what we shared together can never be replaced. Finally, we would like to thank everyone that has helped our during my academic curriculum, even with a simple word."*

Imad and Yaseen.

Contents

List of Acronyms

List of Figures

List of Tables

General Introduction 17

1 BMI, its applications and its estimation from face images 19

1.1 What is BMI 19

1.1.1 BMI formula 19

1.1.2 BMI categories 21

1.2 BMI for biometrics 22

1.2.1 Types of biometrics 22

1.2.1.1 Hard biometrics 22

1.2.1.2 Soft biometrics 23

1.3 BMI in medical context 24

1.3.1 Importance of BMI estimation and tracking 24

1.3.2 Alarming global statistics 25

1.3.3 Challenges in traditional BMI measurement 27

1.4 BMI and face images 28

1.4.1 Facial features and BMI correlation 28

1.4.2 General principle of existing solutions 29

1.4.3 Potential applications 29

1.4.4 General approach outline . . . 30
1.4.5 Advantages of face-based BMI estimation . . . 31
1.5 Conclusion . . . 31

2 Deep learning approaches for BMI estimation from face images 33

2.1 Introduction . . . 33
2.2 1st approach by: Dhanamjayulu et al. (2021) . . . 34
2.2.1 Used dataset . . . 34
2.2.2 Data preprocessing . . . 34
2.2.2.1 Joint face detection and alignment using multitask cascaded convolutional networks MTCNN: . . . 35
2.2.3 Features extraction . . . 40
2.2.4 Regression model . . . 41
2.2.5 Performance of the method . . . 41
2.3 2nd Approach by: Yousaf et al. (2021) . . . 42
2.3.1 Used datasets . . . 42
2.3.2 Data preprocessing . . . 42
2.3.3 Features extraction . . . 42
2.3.4 Regression model . . . 42
2.3.5 Performance of the method . . . 43
2.4 3rd Approach by: Sidhpura et al. (2022) . . . 44
2.4.1 Used datasets . . . 44
2.4.2 Data preprocessing . . . 44
2.4.3 Features extraction . . . 44
2.4.4 Regression model . . . 45
2.4.5 Performance of the method . . . 46
2.5 Conclusion . . . 47

3 Datasets for BMI estimation from face images 48

3.1 Introduction . . . 48

3.2 Selected datasets 49
3.2.1 VisualBMI dataset 49
3.2.2 Arrest Records dataset 52
3.2.3 Illinois DOC labeled faces dataset 54
3.3 Conclusion 56

4 Our approaches for BMI estimation from face images 58

4.1 Introduction 58
4.2 Our methodology 59
4.2.1 Model selection: 59
4.2.2 Dataset usage methodology: 60
4.3 Experiment 01 : ViT-H/14 + Visual BMI 60
4.3.1 Model description 60
4.3.2 Data preprocessing 61
4.3.3 Model architecture 62
4.3.4 Training and optimization 73
4.3.5 Data augmentation 73
4.3.6 Experimental results, interpretations 76
4.4 Experiment 02 : VGG16, ResNet50 + Arrest Records 78
4.4.1 Model description 78
4.4.2 Data preprocessing 78
4.4.3 Model architecture 79
4.4.4 Training and optimization 90
4.4.5 Experimental results, interpretations 91
4.5 Experiment 03 : Inception-V3 + Illinois DOC Faces 93
4.5.1 Model description 93
4.5.2 Data preprocessing 94
4.5.3 Face detection and alignment method selection 94
4.5.4 Model architecture 96

4.5.5 Data augmentation option 101

4.5.6 Gender-based model 102

4.5.7 Training and optimization for (Gender-mixed model / Gender-based model) . 102

4.5.8 Experimental results, interpretations 103

4.6 Conclusion 106

Conclusion **108**

Bibliography **110**

List of Acronyms

- **ACD** : Anterior corneal curvature
- **AUC** : Area Under the Curve
- **BMI** : Body Mass Index
- **CNN** : Convolutional Neural Network
- **CPU** : Central Processing Unit
- **CV** : Computer Vision
- **CVD** : Cardiovascular disease
- **FC** : Fully Connected (Layer)
- **GeLU** : Gaussian Error Linear Unit
- **GPU** : Graphics Processing Unit
- **HOG** : Histogram of Oriented Gradients
- **IOP** : Intraocular pressure
- **MAE** : Mean Absolute Error
- **ML** : Machine Learning
- **MLP** : Multi-layer perceptron
- **MSE** : Mean Squared Error
- **MTCNN** : Multi-task Cascaded Convolutional Neural Network
- **NCD** : Non-Communicable Diseases
- **NLP** : Natural Language Processing
- **PReLU** : Parametric Rectified Linear Unit
- **ReLU** : Rectified Linear Unit
- **RMSE** : Root Mean Squared Error
- **SOTA** : State of the art

- **SVM** : Support Vector Machine

- **TPU** : Tensor Processing Unit

- **ViT** : Vision Transformer

List of Figures

1.1 Some facial images display varying BMI values and corresponding categories. As the BMI increases, noticeable increases in facial adiposity become evident [1]. 21

1.2 Types Of biometric information [5]. 24

1.3 Proportion of deaths from leading NCDs attributable to high BMI [8]. 25

1.4 Projected numbers of adults and children with high Body Mass Index (BMI) [8]. . . . 26

1.5 06 images taken from VisualBMI dataset with their BMI. 28

1.6 General road-map [6]. 30

2.1 Pipeline of the 03 stages [15]. 36

2.2 Architectures of P-Net, R-Net, and O-Net [15]. 38

2.3 ReLU vs. PReLU. For PReLU, the coefficient of the negative part is **not constant and is adaptively learned** [16] [15]. 38

2.4 Transfer learning idea [17]. 40

2.5 Regression model architecture [6]. 45

2.6 GELU, RELU, ELU activation functions curves [23]. 46

3.1 VisualBMI dataset images. 50

3.2 BMI category statistics of VisualBMI dataset. 51

3.3 Gender distribution of VisualBMI dataset. 51

3.4 Arrest Records dataset images. 52

3.5 BMI category statistics of Arrest Records dataset. 53

3.6 Gender distribution of Arrest Records dataset. 53

3.7 Illinois DOC labeled faces dataset images. 54

3.8 BMI category statistics of Illinois DOC labeled faces dataset. 55

3.9 BMI category statistics of Illinois DOC labeled faces dataset by gender. 56
3.10 Gender distribution of Illinois DOC labeled faces dataset. 57

4.1 Model description. 61
4.2 Original images / Preprocessed images with MTCNN. 62
4.3 Vision transformer ViT architecture [29]. 63
4.4 Image devided into fixed-size patches [30]. 65
4.5 Embedding step [30]. 66
4.6 Necessity of patches order [30]. 67
4.7 Calculation of dimensions example [30]. 67
4.8 Transformer encoder block [30]. 68
4.9 Dimension of multi-head attention block [30]. 69
4.10 Dimensions for attention block [30]. 70
4.11 Complete ViT architecture [30]. 71
4.12 Regression model architecture. 72
4.13 Data augmentation images. 74
4.14 Original image , rotated image by 5 degree. 74
4.15 Original image , horizontal flipped image. 75
4.16 Orignal image, image with 0.8 brightness factor. 75
4.17 Original image , image with 96.34% contrast. 76
4.18 Model description. 78
4.19 Original images / Preprocessed images with MTCNN 79
4.20 ResNet50 architecture [31]. 80
4.21 The concept of the widely popular CNN [31]. 81
4.22 VGG-16 architecture. [32] 83
4.23 Softmax curve [33]. 84
4.24 Example [33]. 84
4.25 VGG-16 architecture map [32]. 86
4.26 Regression head. 89

4.27 Classification head. 89

4.28 Model description. 94

4.29 MAE in kg of the VIP attribute test set for various face detectors and cropping margins [2]. 96

4.30 Inception-V3 model architecture [35]. 97

4.31 Mini-network replacing the 5 × 5 convolutions [35]. 98

4.32 Mini-network replacing the 3 × 3 convolutions [35]. 99

4.33 Inception modules after the factorization of the n × n convolutions [35]. 100

4.34 Inception module that reduces the grid-size while expands the filter banks [35]. 100

4.35 Regression model architecture 101

List of Tables

1.1 Global estimate (2020) and projected number of young people (2025–2035) with overweight and obesity [8]. 26

2.1 Results on Arrest Records dataset [7]. 41

2.2 Results on VisualBMI dataset [20]. 43

2.3 Testing results on 03 datasets [6]. 46

4.1 Obtained results on **VisualBMI**: Existing work (VGGFace) vs. Our model (with different Hyper-Parameters). 76

4.2 Obtained Results on **Arrest Records**: Existing work (VGGFace) vs. Our model (with different Hyper-Parameters). 92

4.3 Obtained results with **MSE vs. MAE** loss functions. 93

4.4 Obtained results on our **03 datasets**: Female-based model (with different hyper-parameters). 103

4.5 Obtained results on our **03 datasets**: Male-based model (with different hyper-parameters).104

4.6 Obtained results on our **03 datasets**: Gender-mixed model (with different hyper-parameters). 105

General Introduction

"The face is the mirror of the mind, and eyes without speaking confess the secrets of the heart."

St. Jerome – 4th Century

Body Mass Index (BMI) is a widely accepted indicator used to classify individuals according to their body weight relative to height. It provides an effective yet simple way to assess weight-related health risks, categorizing individuals into groups such as underweight, normal weight, overweight, or obese. While traditionally BMI is calculated using manually provided weight and height, recent progress in artificial intelligence and computer vision has opened new pathways for estimating BMI directly from facial images—automating the process and increasing accessibility.

Facial analysis presents a powerful, non-invasive alternative by leveraging visual cues that correlate with body weight. Features like facial adiposity, cheek fullness, jaw structure, and geometric ratios have shown strong associations with BMI. Advances in deep learning and transfer learning, combined with the availability of large-scale annotated facial datasets, have enabled the training of robust models that predict BMI directly from images.

This approach has gained interest due to its practical benefits in areas like remote health monitoring, digital wellness platforms, and biometric analytics. In contexts where scales or direct body measurements are unavailable—or even socially intrusive—face-based estimation provides a valuable alternative. However, this method brings challenges including demographic bias, variability in lighting and pose, and generalization across diverse populations, all of which must be addressed to ensure reliability and fairness.

In this project, we present a comprehensive study of BMI estimation from facial images using deep learning. We compare several state-of-the-art architectures such as Vision Transformers (ViT), VGG16, ResNet50, and Inception-V3, across multiple public datasets (VisualBMI, Arrest Records, and Illinois DOC Faces). We also investigate how data preprocessing, augmentation, and loss function choices influence performance, and explore gender-specific models to improve predictive accuracy.

The main contributions of this work include:

- A review and comparison of leading facial BMI estimation techniques based on different deep learning models and pipelines.
- Implementation of customized preprocessing and augmentation strategies, including face alignment using MTCNN.

- Empirical evaluation of models across gender-specific and mixed-gender datasets, with performance measured using MAE, RMSE, and AUC.

- A robust training strategy integrating multiple datasets to improve generalization to unseen subjects and real-world conditions.

This thesis is organized into four chapters structured as follows:
Chapter one introduces the concept of BMI, its health implications, and the motivation for estimating it from face images.
Chapter two provides a detailed review of existing deep learning approaches, including face detection and feature extraction techniques used in BMI estimation.
Chapter three presents the datasets used in this study and discusses their composition, characteristics, and relevance.
Chapter four outlines our proposed methodology, experiments, and results, including model architectures, training strategies, and performance evaluations.

Through this work, we aim to contribute to the development of accessible, non-invasive health monitoring tools using computer vision and deep learning, while highlighting the importance of fairness, transparency, and robustness in AI-driven biomedical applications.

Finally, this manuscript concludes with a general conclusion that summarizes the key findings and insights of this work. It also outlines future research directions and perspectives that could further enhance the performance, generalizability, and applicability of BMI estimation systems. Potential improvements include exploring multimodal inputs, addressing demographic biases more effectively, and extending the system to real-time or mobile health applications.

Chapter 1

BMI, its applications and its estimation from face images

1.1 What is BMI

Body Mass Index (BMI) is a simple yet essential metric used to assess whether an individual has a healthy weight relative to their height. Medical professionals across the world use BMI daily to classify individuals as underweight, of normal weight, overweight, or obese.

As weight-related health issues continue to rise globally, understanding BMI becomes increasingly important. Many people are unaware of their health status in terms of weight, and BMI provides an accessible method for quick and informative evaluation.

1.1.1 BMI formula

Recently, BMI has drawn significant interest in health-monitoring and weight loss applications. Body mass index (BMI) serves as a metric for assessing an individual's weight relative to their height. This measurement is derived by dividing the weight in kilograms by the square of the height in meters. In short, BMI provides a quantitative evaluation of body composition [1].

$$\mathrm{BMI} = \begin{cases} \dfrac{\text{weight (kg)}}{\text{height (m)}^2} \\ \\ \dfrac{\text{weight (lb)}_{\times 703}}{\text{height (in)}^2} \end{cases} \tag{1.1}$$

kg = **lbs** × 0.453592

m = **in** × 0.0254

1.1.2 BMI categories

BMI values are divided into 4 categories: underweight, normal, overweight, and obese [2].

- **Underweight:** BMI $<$ 18.5.
- **Normal:** 18.5 $<$ BMI $<$ 25.
- **Overweight:** 25 $<$ BMI $<$ 30.
- **Obese:** BMI $>$ 30 [2].

BMI values correspond to specific areas, providing a broader understanding of a person's weight status. Visual models of facial images corresponding to different BMI categories provide a visual representation of different weight statuses in the population studied [1], and the Figure 1.1 shown some facial images display varying BMI values and corresponding categories :

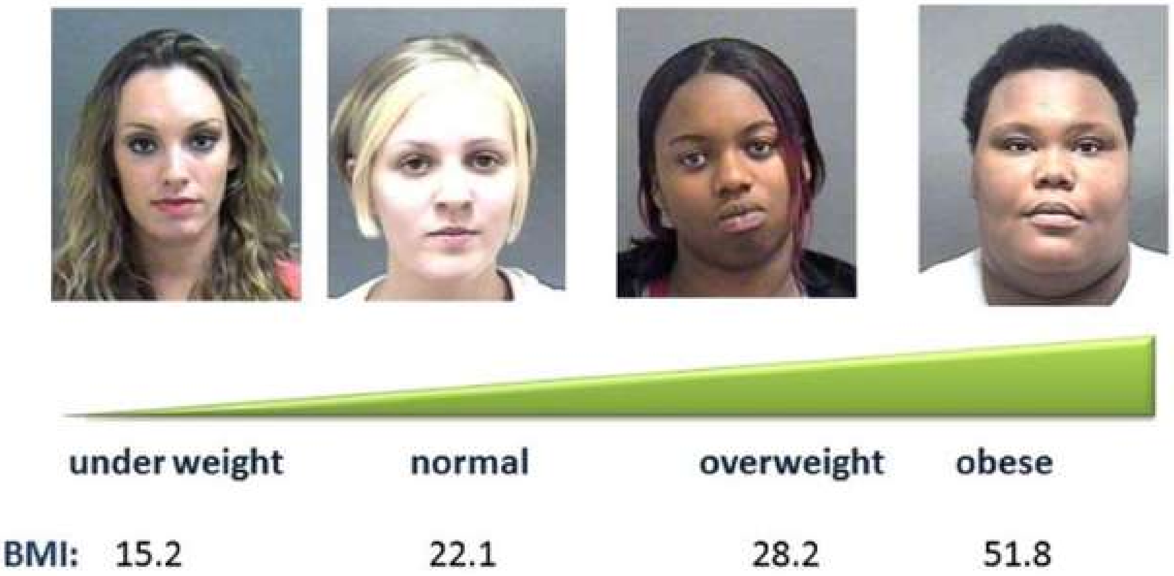


Figure 1.1: Some facial images display varying BMI values and corresponding categories. As the BMI increases, noticeable increases in facial adiposity become evident [1].

1.2 BMI for biometrics

Biometrics refers to the measurement and statistical analysis of people's physical and behavioral characteristics. It is commonly used for identification, authentication, and access control. The basic principle is that individuals can be recognized based on their intrinsic biological and behavioral traits. The term "biometrics" originates from the Greek words *bios* (life) and *metron* (measure) [3].

BMI is part of a broader field called biometrics, which involves measuring various physical and behavioral characteristics of the human body. Biometrics is generally divided into two categories: soft biometrics, which include variables such as age, gender, height, and weight—factors that can change over time; and hard biometrics, such as fingerprints, facial structure, and iris patterns, which remain largely constant.

1.2.1 Types of biometrics

Biometric traits are typically classified into two main categories: **hard biometrics** and **soft biometrics** [4].

1.2.1.1 Hard biometrics

Hard biometrics are unique, permanent, and highly discriminative traits that can reliably and unambiguously identify an individual. These traits remain stable over time and are widely used in security and authentication systems. Hard biometrics include:

- **Physical characteristics**: face geometry, fingerprint, iris, retina, hand geometry.
- **Biological characteristics**: DNA, blood type, and other physiological traits.
- **Behavioral characteristics**: signature dynamics, voice patterns, gait, and keystroke dynamics (when they are consistent and unique to individuals).

1.2.1.2 Soft biometrics

Soft biometrics are traits that are not sufficient for unique identification on their own but can provide supportive or descriptive information to improve recognition accuracy when combined with hard biometrics. They are often variable over time and include attributes such as:

- **Descriptive features**: gender, age, ethnicity, height, weight, body mass index (BMI).
- **Contextual or semantic attributes**: clothing style, accessories, tattoos, and facial hair.

Soft biometrics are particularly useful in scenarios such as surveillance, where full biometric data might not be available, and in health-related applications, such as the prediction of BMI or age from images.

The distinction between soft and hard biometrics is important, especially in the design of secure and reliable authentication systems. While hard biometrics offer high accuracy, soft biometrics can provide useful supplemental information, particularly in surveillance scenarios.

Note: In this work, we focus specifically on soft biometric estimation—particularly BMI—from facial images, which does not aim to uniquely identify individuals but rather to infer informative health-related characteristics.

The following Figure 1.2 shows the types of biometric information with some examples:

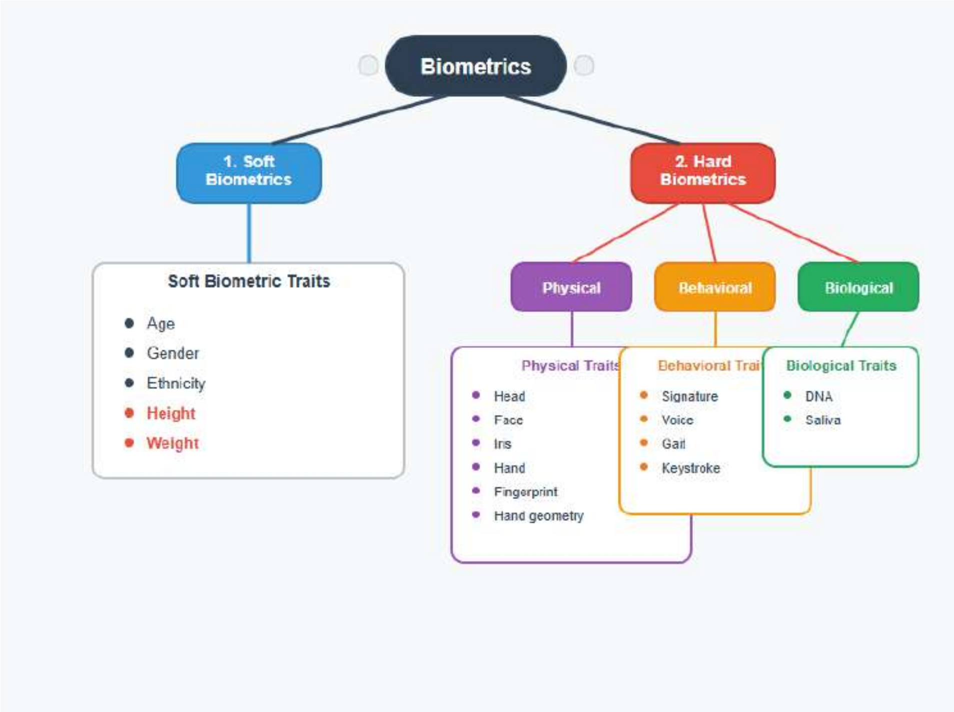


Figure 1.2: Types Of biometric information [5].

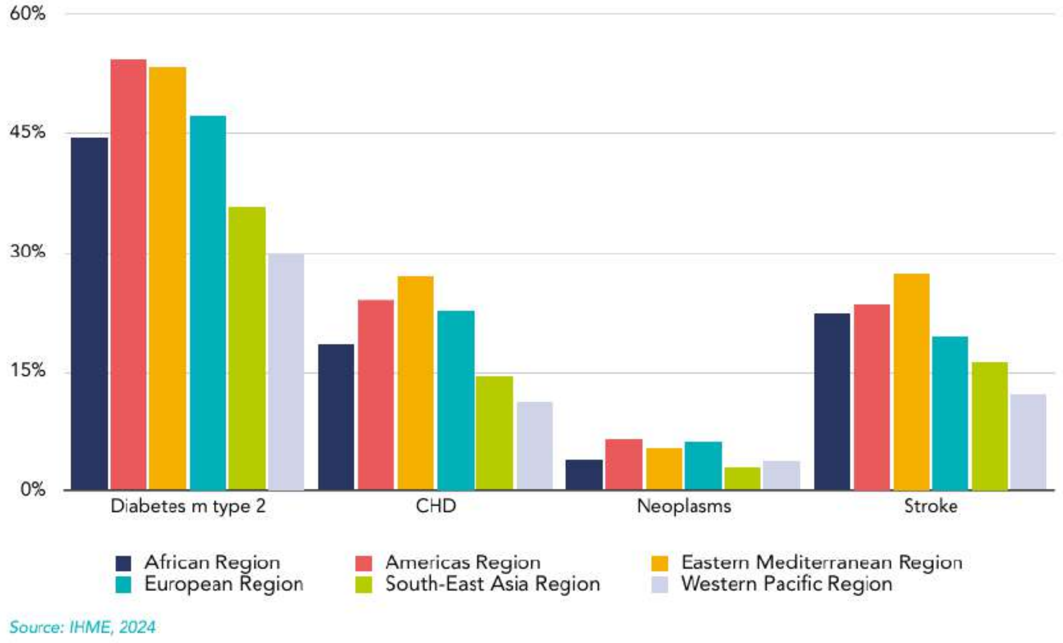
1.3 BMI in medical context

1.3.1 Importance of BMI estimation and tracking

BMI serves as a crucial indicator in various medical and health-related contexts:

- Numerous aspects like physical health, mental health, and popularity have been linked to weight and BMI [6].
- High BMI increases the risk of chronic and cardiovascular diseases, cancers like colon cancer, breast cancer, and thyroid cancer for all genders [7].
- Low BMI can indicate inadequacies or malnutrition, so BMI can help a person to keep a track record of their health [6].
- BMI is considered a risk factor for myocardial infarction in the field of medicine. In patients, it is often known as a risk factor for dysfunctional angina. The risk of type 2 diabetes, neoplasms, stroke and cardiovascular disease (CVD) can be stratified using BMI [7].

The following Figure 1.3 represent the Proportion of deaths from leading "Non-Communicable Diseases" attributable to high BMI, showing and ensuring more about the importance of tracking BMI:



Source: IHME, 2024

Figure 1.3: Proportion of deaths from leading NCDs attributable to high BMI [8].

1.3.2 Alarming global statistics

In the current scenario, health is one of the most neglected factors. Technology which has more benefits also has some drawbacks. It has made humans lazy and thus reduced their physical activity leading to a sedentary lifestyle and a rise in BMI which adversely affects their health.

- Almost 26% of adults in the United Kingdom are living with obesity [9].
- Excess weight is responsible for approximately 500,000 deaths per year in the United States [9].
- By the year 2030, an estimated 20% of the global population would be obese [9].

  This Figure 1.4 shows Algeria's projected obesity crisis from 2020-2035 for children and adults. This trend highlights the urgent need for accessible BMI monitoring tools like our face-based estimation system.

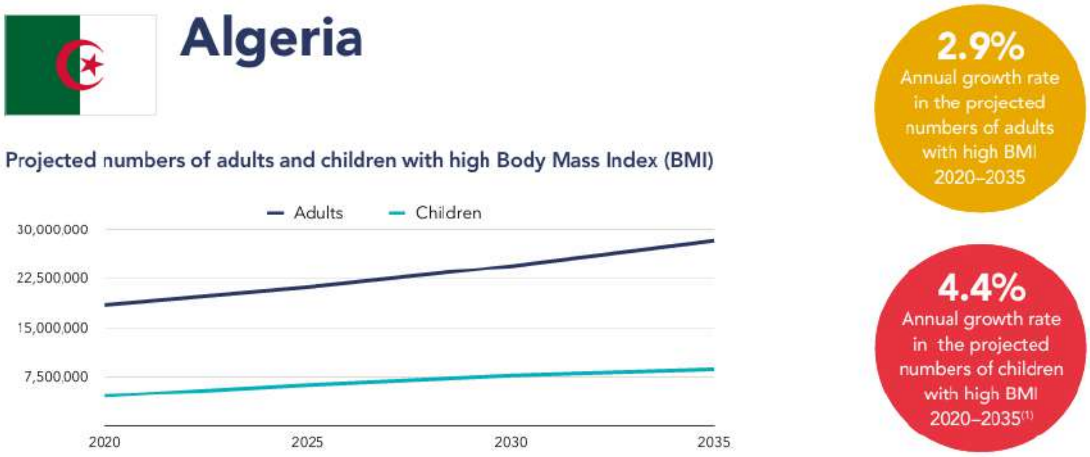


Figure 1.4: Projected numbers of adults and children with high Body Mass Index (BMI) [8].

- Trends related to obesity are alarming because obesity affects 93.3 million adults in the United States alone [10].

- Obesity is one of the biggest drivers of preventable chronic diseases and healthcare costs in the United States. Severe obesity costs the United States approximately 69 billion overall, with almost 8 billion a year being paid for via state Medicaid programs [11].

- According to research published by the "GBD 2017 Obesity Collaborators", more than 4 million people die each year from complications related to excess weight. Paradoxically, the prevalence of obesity has increased in recent decades, affecting both children and adults worldwide [1].

- The following Figure 1.5 represents Global estimate (2020) and projected number of young people (2025-2035) with overweight (BMI >1sd – 2sd)* and obesity (BMI >2sd)*

Table 1.1: Global estimate (2020) and projected number of young people (2025–2035) with overweight and obesity [8].

Year	Children with overweight	Chilfren with obesity
2020	260m	175m
2025	310m	240m
2030	350m	310m
2035	390m	380m

- The burden of obesity is not limited to developed countries; low-income and developing countries are also struggling with rising rates. In fact, the overall infection rate in these areas is a staggering 30% higher than the rate in developed countries [1].

1.3.3 Challenges in traditional BMI measurement

The main challenges in BMI estimation can be summarized into 4 key points:

0 People often fail to keep track of their weight and BMI over time. This happens because some people find it too much effort to measure themselves regularly, while others simply don't own or can't easily use basic measurement tools like scales and rulers [6].

0 Traditional methods for measuring body weight and size require people to be physically present during the assessment process, which can make many individuals feel uncomfortable or embarrassed about their bodies [12].

0 The BMI calculation relies on a specific statistical formula that requires accurate and detailed measurements of both height and weight. Without precise data, the BMI score may not reflect the person's true health status [13].

0 The whole procedure of taking precise body measurements is very time-consuming and often requires several attempts to get reliable results, which makes it inconvenient for both patients and healthcare providers [13].

Conventional measurement techniques require the cooperation of the subject to be measured, which might not be possible during medical emergencies where rapid investigation is needed, road accidents or due to different patient disabilities.

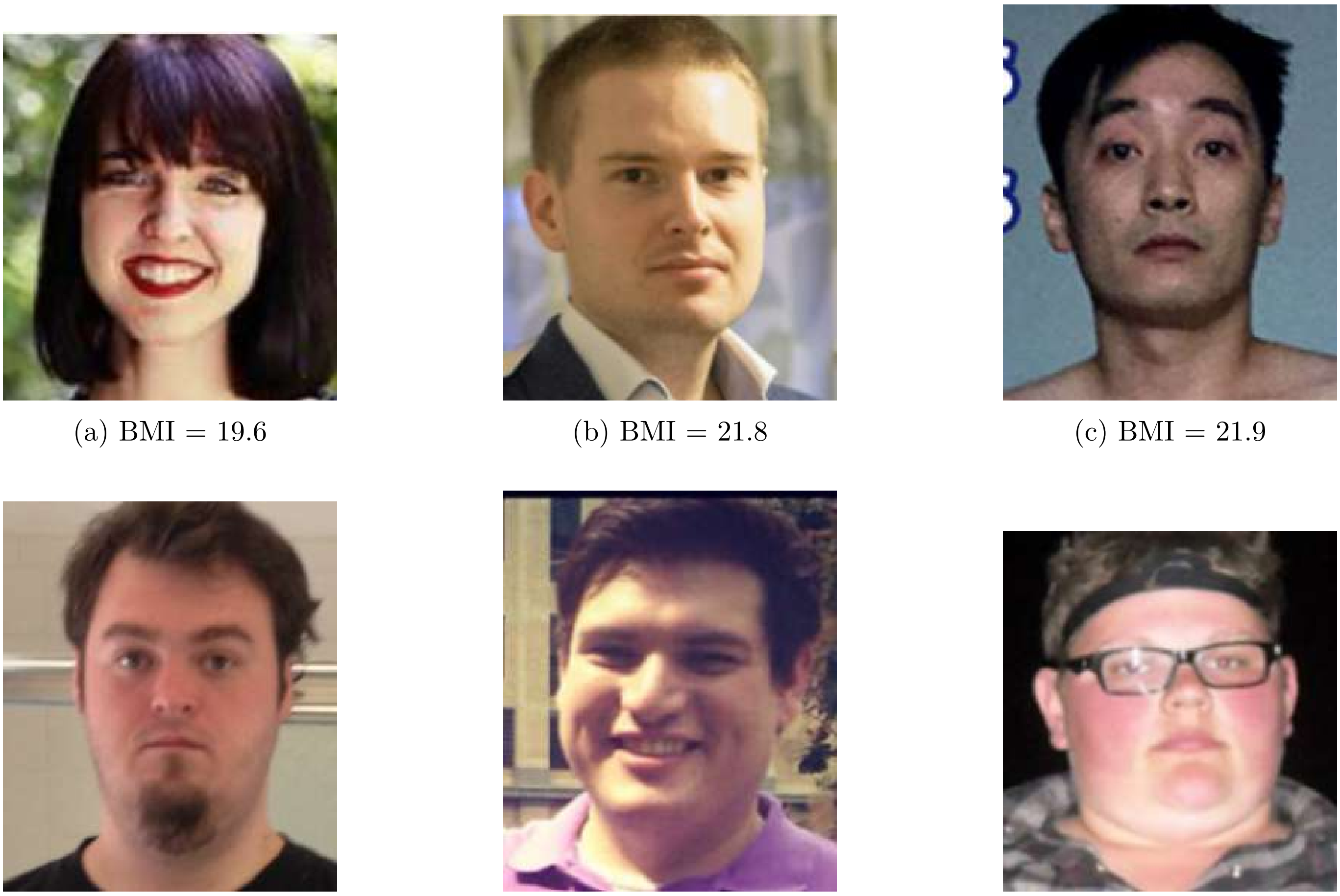
1.4 BMI and face images

1.4.1 Facial features and BMI correlation

Human faces carry a significant amount of information about a person. Recent studies specially in psychology and human emotions have shown a strong correlation between the human face and the BMI of the person. It's visually clear that for example people with skinny faces have chances of less BMI and vice versa.

Generally, obese people tend to have the middle and lower part of the face wider, and to go more specifically: BMI and weight are strongly related with the structure of eye or it uses features anterior corneal curvature (ACD) and intraocular pressure (IOP), neck circumference, and physical measurements of the face, including ratios like: width to height, perimeter to area, and cheek to jaw width [6] [13] [1].

- This is can be seen from this following Figure 1.5 :



(a) BMI = 19.6

(b) BMI = 21.8

(c) BMI = 21.9

(d) BMI = 39.6

(e) BMI = 40.4

(f) BMI = 45.6

Figure 1.5: 06 images taken from VisualBMI dataset with their BMI.

In the modern era, advances in technology have transformed lifestyles but have also led to increased sedentarism. Prolonged screen time, reduced physical activity, and unhealthy dietary habits contribute significantly to the growing prevalence of weight-related health problems worldwide.

1.4.2 General principle of existing solutions

Because of some academic researches claim that machine and human readable facial pictures can be used to determine BMI, the existing solution to facilitate and automate and simplify this process of estimating the BMI or the weight of a person can be as follows: **Predict the BMI by giving only the Face Image as input**, by using Machine Learning or Deep Learning approaches where models can extract meaningful features from Face Images, and estimate the BMI [6] [14].

1.4.3 Potential applications

- This system has the potential to revolutionize health monitoring and management. For health insurance companies, it could serve as a tool to maintain and update the health records of their customers efficiently. By automating BMI and weight estimation through facial images [6].
- Additionally, governments could leverage this system to track and analyze the health metrics of specific regions, design targeted policies to address issues like obesity, malnutrition in specific regions [6].
- Some social media analysis purposes like: Images from platforms like Instagram can be used to study BMI trends demographically or geographically.
- Many use cases necessitate the estimation of body weight without the physical measurement or presence of a person directly. For example, in health analysis to check the weight through mobile devices for a quick estimation, in forensics to gain additional identification features, in airports to estimate weight to aid dynamic baggage allowance, for physicians working remotely for rural patients [12].

1.4.4 General approach outline

This section is to summarize the whole approach in some main steps to compute the BMI from an input face image, as the Figure 1.6 shows :

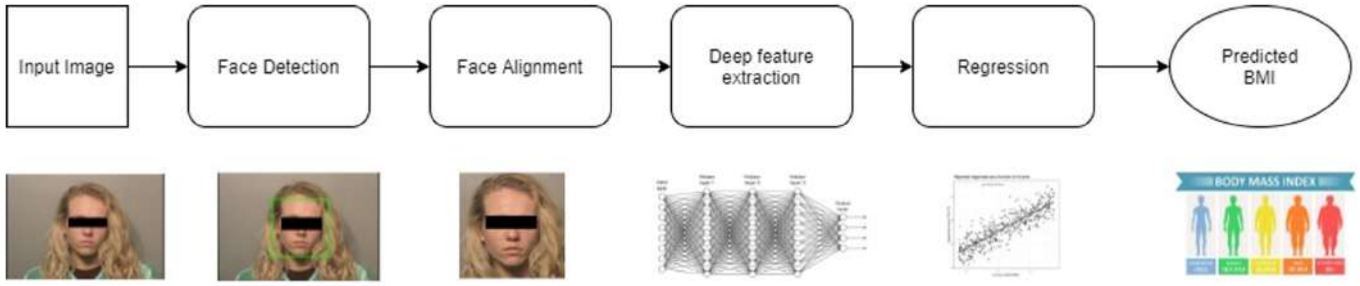


Figure 1.6: General road-map [6].

1. **Input image:**
   - A facial image of a person is provided as input.
2. **Face detection:**
   - A face detection algorithm (HOG + SVM, Hierarchical Hetero-PSO-Adaboost-SVM Model, MTCNN, Viola-Jones algorithm) is used to locate the face within the image.
   - Then blurring the background of all images while focusing only on the face.
3. **Face alignment:**
   - The detected face is aligned vertically to ensure consistency across images, making sure that landmarks like: eyes, nose, and mouth are properly oriented, or consistently positioned (all the faces must be in the same orientation direction -Upright-).
4. **Deep feature extraction:**
   - A deep learning or a machine learning model (Pretrained Models, Custom CNN end-to-end) is used to extract features from the aligned faces.
   - These features encode facial structure, fat distribution, and other visual cues related to BMI.
5. **Regression model:**

- A regression algorithm (Support Vector regression, Deep Neural Network Model) is used to extrapolate or maps the previous extracted deep features to BMI values.

6. **Predicted BMI:**

    - The model outputs the predicted BMI value.

1.4.5 Advantages of face-based BMI estimation

This new approach solves many problems with traditional BMI measurement:

- People don't need scales or measuring tools
- They don't have to feel embarrassed about measuring their body
- The process is quick and can work even when someone cannot cooperate, like in medical emergencies
- It provides a non-invasive method for health monitoring
- It can be implemented in various settings without specialized equipment

1.5 Conclusion

BMI is a very important tool for checking and understanding our health. The numbers and facts we have seen in this chapter are truly alarming and should make us all reflect seriously on the growing weight-related health issues facing the world today.

The statistics clearly show that millions of people globally are overweight or obese, and this number continues to rise each year. Worryingly, this is no longer just a problem for wealthy countries; low-income nations are also experiencing increased rates of obesity, even among children.

Being overweight or obese is associated with numerous serious health conditions, including heart disease, diabetes, and several types of cancer such as breast and colon cancer. Additionally, high BMI can negatively affect mental health and self-esteem. Conversely, being underweight can also be dangerous, often indicating malnutrition or other underlying health problems.

The advantage of BMI lies in its simplicity—it requires only basic information like height and weight, and can be calculated easily. This empowers individuals to take control of their health through better lifestyle choices such as balanced diets and regular physical activity.

This study shows that it is possible to estimate BMI from face images using modern technology. The method works by taking a photo of someone's face, finding the face in the picture, and then using computer programs to analyze facial features. These programs can learn the connection between how a face looks and what the person's BMI might be.

The technology has many useful applications in healthcare, insurance, government health monitoring, and remote patient care. However, this method is not perfect yet and gives an estimate, not an exact measurement. More research is needed to make it more accurate and reliable.

Despite these limitations, face-to-BMI technology represents an important step forward in making health monitoring easier and more accessible for everyone. Maintaining a healthy BMI contributes to both physical and mental well-being. Promoting awareness among individuals, families, schools, and governments is essential to tackling this growing public health challenge.

In the next chapter, we will explore the technical foundations of our approach, including the challenges involved, the facial features linked to BMI, and the detailed methodology we propose to perform this estimation.

Chapter 2

Deep learning approaches for BMI estimation from face images

2.1 Introduction

In recent years, there has been a growing interest in estimating health-related attributes such as Body Mass Index (BMI) directly from facial images using deep learning techniques. This interest is driven by the increasing availability of visual data, advances in computer vision, and the potential applications in non-invasive health monitoring, biometric systems, and personalized healthcare services. Traditional methods of BMI estimation require manual input of weight and height, which may not always be accessible or reliable. In contrast, facial analysis offers a passive and automated approach that can be integrated into modern digital platforms with minimal user effort.

Several research studies have explored various methodologies to predict BMI from facial features, employing different datasets, preprocessing strategies, feature extraction techniques, and machine learning models. These approaches range from basic regression frameworks to more complex architectures involving multitask learning, facial region segmentation, and fine-tuned deep neural networks.

This chapter presents a comprehensive review of three prominent approaches in the literature: the works of Dhanamjayulu et al. (2021), Yousaf et al. (2021), and Sidhpura et al. (2022). Each of these studies brings forward unique contributions and methodologies in tackling the challenge of BMI estimation from facial images. Through a comparative analysis of their datasets, preprocessing pipelines, model architectures, and results, this chapter aims to highlight current trends, strengths, and limitations in the state of the art, providing valuable insights for future research directions.

The chapter is organized around the detailed presentation of each approach, followed by a comparative discussion of their outcomes and key takeaways.

2.2 1st approach by: Dhanamjayulu et al. (2021)

2.2.1 Used dataset

The dataset used in this study is a small Arrest Records dataset consisting of 1544 person images scraped from the web. Each record includes metadata such as:

- Age
- Height (in feet and inches)
- Weight (in pounds)
- Gender
- Race

Each image is named according to a unique ID and is linked to a metadata row. The dataset includes a variety of races and is dominated by male samples (approximately 80%) [7].

2.2.2 Data preprocessing

- Converting height to meters and weight to kilograms.
- Calculating BMI: BMI $= \frac{\text{weight (kg)}}{\text{height (m)}^2}$.
- Encoding gender numerically.
- Splitting the dataset into training (1227 images) and testing (317 images).
- Using the **MTCNN face detection model**, to detect and cropp only the region of interest (face region), from the whole image [7].

2.2.2.1 Joint face detection and alignment using multitask cascaded convolutional networks MTCNN:

- In 2016 a powerful framework called MTCNN have introduced, It consists of 03 stages, each with its own CNN: **P-Net, R-Net, and O-Net**. The key innovation is using multitask learning to simultaneously: [15]

- **Detect faces** (classification).
- **Refine bounding boxes** (regression).
- **Locate facial landmarks** (alignment).

- The following Figure 2.1 shows the Pipeline of the cascaded framework that includes 03 stage multitask deep convolutional networks. First, candidate windows are produced through a fast **P-Net**. After that, we refine these candidates in the next stage through a **R-Net**. In the third stage, the **O-Net** produces final bounding box and facial landmarks position : [15]

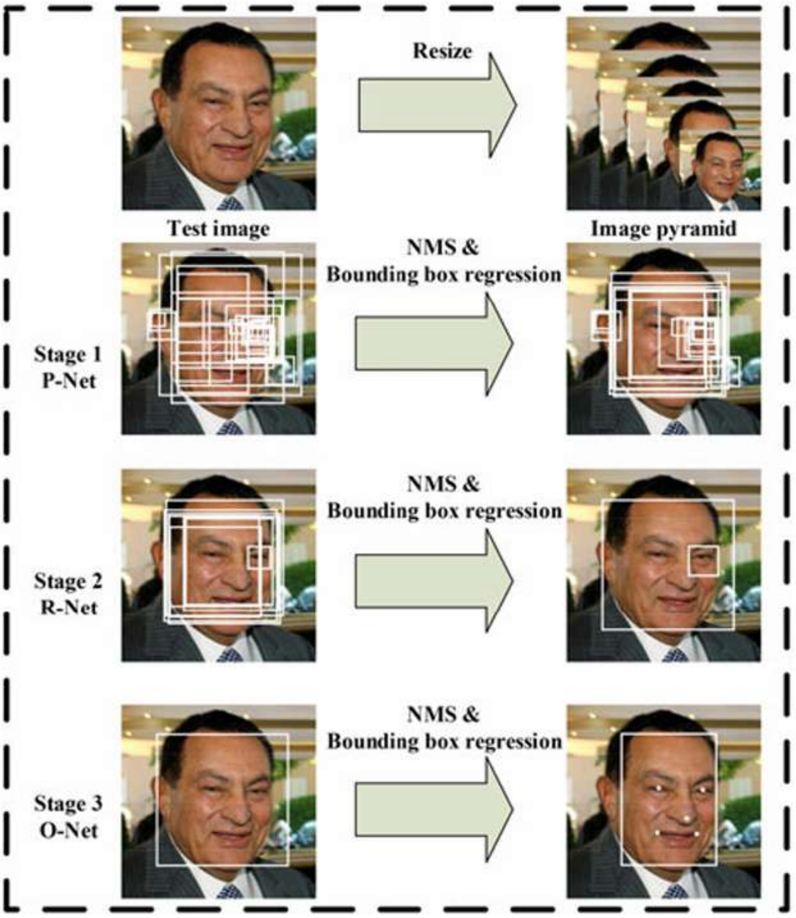


Figure 2.1: Pipeline of the 03 stages [15].

Stage 1: P-Net (Proposal network) [15]

- Input image is resized to multiple scales to detect faces of varying sizes, called: **Image pyramid**.

- Fully convolutional network processes each scale, to obtain the candidate facial windows and their **bounding box regression vectors**. Then candidates are **calibrated** based on the estimated bounding box regression vectors. After that, we employ **non maximum suppression (NMS)** to merge highly overlapped candidates.

- This 1st stage is for doing Fast screening **eliminates obvious non-face regions**.

Stage 2: R-Net (Refine network) [15]

- All candidates are fed to another CNN, called refine network (R-Net), which further rejects a large number of false candidates, that mean **rejecting more false positives** (Actual: negative / Predicted: positive).

- Finally, its performs calibration with bounding box regression (After the bounding box regression outputs the offsets, **calibration means**: applying those offsets to adjust the original boxes), and conducts NMS (NMS removes **duplicate or overlapping detections** for the same object, so it's applied to reduce redundancy, If multiple bounding boxes are detecting the same object (a face), we want to: Keep only the best box (highest confidence score), Remove redundant ones that overlap too much based on a specific threshold), This stage allow better descriminations.

Stage 3: O-Net (Output network) [15]

- This 3rd stage is similar to the second stage, but in this stage we aim to identify face regions with more supervision (Final decision making). In particular, the network will output 05 facial landmarks' positions (including: **left eye**, **right eye**, **nose**, **left mouth corner**, and **right mouth corner**).

- This is because that the 3rd stage Performs the most precise face bounding box regression and landmark localization, du to the **Most complex network** (Deepest network with most parameters).

CNN architectures of MTCNN stages

The CNN architectures of MTCNN stages is shown in the following Figure 2.2, and it will be explained later:

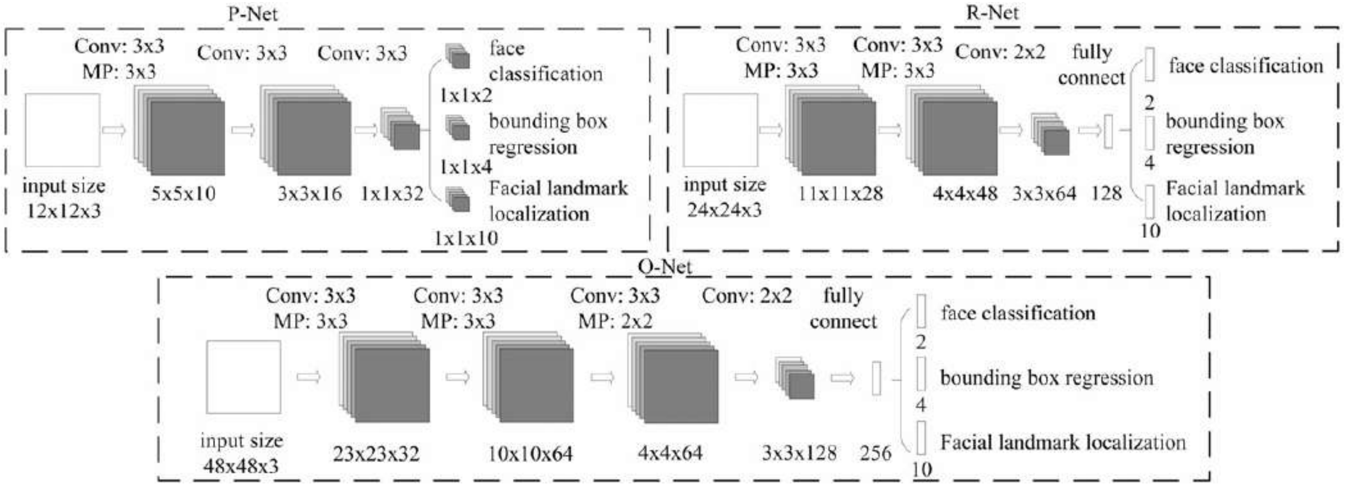


Figure 2.2: Architectures of P-Net, R-Net, and O-Net [15].

- **PReLU** as a non-linearity activation function is applied after the convolution and fully connection layers (except output layers), They wanted to avoid dead neurons to improve convergence. The difference between this activation function "PReLU" and "ReLU" is shown here in this Figure 2.3 :

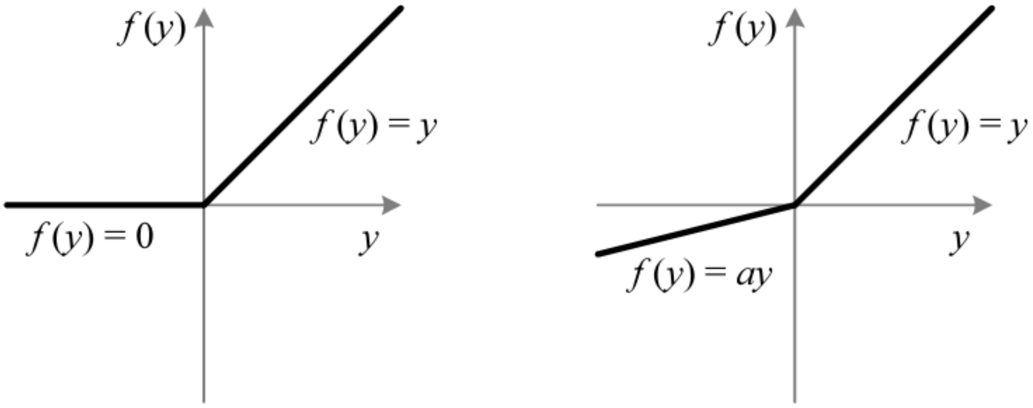


Figure 2.3: ReLU vs. PReLU. For PReLU, the coefficient of the negative part is **not constant and is adaptively learned** [16] [15].

- Each network performs three tasks: [15]

- **Face classification** (is it a face?).
- **Bounding box regression** (adjust face box location).
- **Facial landmark localization** (predict 5 keypoints).

1/- Face classification [15]

The learning objective is formulated as a two-class classification problem. For each sample x_i, we use the **cross-entropy loss** as

$$L_i^{\text{det}} = -(y_i^{\text{det}} \log(p_i) + (1 - y_i^{\text{det}})(1 - \log(p_i))) \tag{2.1}$$

where p_i is the probability produced by the network that indicates sample x_i being a face. The notation $y_i^{\text{det}} \in \{0, 1\}$ denotes the ground-truth label.

2/- Bounding box regression [15]

For each candidate window, we predict the offset between it and the nearest ground truth (i.e., the bounding boxes' left, top, height, and width). The learning objective is formulated as a regression problem, and we employ the Euclidean loss for each sample x_i

$$L_i^{\text{box}} = ||\hat{y}_i^{\text{box}} - y_i^{\text{box}}||_2^2 \tag{2.2}$$

where $\hat{y}_i^{\text{box}}$ is the regression target obtained from the network and y_i^{box} is the ground-truth coordinate. There are four coordinates, including: (left, top, height and width), and thus $y_i^{\text{box}} \in \mathbb{R}^4$.

- **Left position (x):** The x-coordinate of the left edge of the bounding box.
- **Top position (y):** The y-coordinate of the top edge of the bounding box.
- **Height:** How tall the bounding box is (vertical dimension).
- **Width:** How wide the bounding box is (horizontal dimension).

3/- Facial landmark localization [15]: Similar to bounding box regression task, facial landmark detection is formulated as a regression problem and we minimize the Euclidean loss as

$$L_i^{\text{landmark}} = ||\hat{y}_i^{\text{landmark}} - y_i^{\text{landmark}}||_2^2 \tag{2.3}$$

where $\hat{y}_i^{\text{landmark}}$ is the facial landmark's coordinates obtained from the network and y_i^{landmark} is the ground-truth coordinate for the ith sample. There are five facial landmarks, including: (left eye, right eye, nose, left mouth corner, and right mouth corner), and thus $y_i^{\text{landmark}} \in \mathbb{R}^{10}$.

2.2.3 Features extraction

Deep features were extracted using ResNet-50 and VGG16 pre-trained models (will be explained later) via transfer learning from facial images [7].

Transfer learning using pre-trained models: Transfer learning is an approach to machine learning where a model trained on one task is used as the starting point for a model on a new task. This is done by transferring the knowledge that the first model has learned about the features of the data to the second model [17], as it shown in the Figure 4.9.

BMI calculation from facial images is rather complicated so learning all required features from relatively small datasets would be infeasible. Many tasks in computer vision have used transfer learning to boost performance and reduce the training time. Hence a state of the art pre-trained models such as: (**ResNet50, VGG16 ..ect**) are used.

In feature extraction, the pre-trained model is used to extract features from the data. These features are then used to train a new model on the target task. This is a good approach in the case of a limited data for the target task.

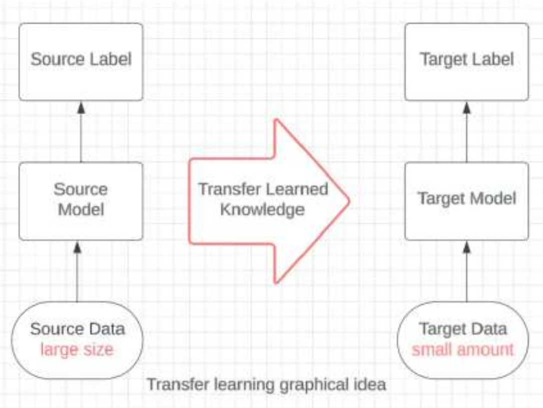


Figure 2.4: Transfer learning idea [17].

2.2.4 Regression model

A multi-task learning framework was used for simultaneous prediction of:

- **BMI** (regression).
- **Age** (regression).
- **Gender** (classification).

Each task was handled by dedicated fully connected layers [7].

2.2.5 Performance of the method

To evaluate the performance of their BMI estimation model from facial images, they adopted the widely-used regression metric MAE [18] and AUC area under the curve [19] as classification metric. These evaluation metrics help quantify the accuracy and reliability of the predicted values compared to the ground-truth values.

The following Table 2.1 shows the obtained results :

Table 2.1: Results on Arrest Records dataset [7].

Models \ Results	**BMI (MAE)**	**Age (MAE)**	**Gender (AUC)**
ResNet50	5.02	7.16	0.998
VGG16	6.13	9.57	0.99

Interpretation

- ResNet50 model clearly outperform the VGG16 Model.
- Low BMI regression preformances, due to the insufficient size of training dataset (very small = 1227).
- Dataset imbalance and range limitations may affect generalization.

2.3 2nd Approach by: Yousaf et al. (2021)

2.3.1 Used datasets

The study evaluated the method across three public datasets:

- **VisualBMI:** 4206 user-shared transformation face images.
- **VIP-Attribute:** 1026 celebrity face images (balanced gender).
- **Bollywood Dataset:** 237 celebrity images with multiple samples per identity [20].

2.3.2 Data preprocessing

- The first 3368 images are used for training and the rest of the images are used for testing.
- Face detection using MTCNN.
- Semantic segmentation using a modified BiSeNet trained on CelebAMask-HQ to extract precise facial regions (eyes, nose, lips, hair, etc).
- Mask generation and interpolation for region-based processing [20].

2.3.3 Features extraction

- Deep features were extracted using VGGFace and FaceNet pre-trained models backbone.
- Final convolution layer features were pooled region-wise using masks.
- Region-Aware Global Average Pooling (Reg-GAP) was performed over segmented areas to enhance local feature representation [20].

2.3.4 Regression model

- Fully connected network with layers (512 → 256 → 1), using ReLU activation and dropout (0.4).

- Final output layer for BMI regression with linear activation.
- Optimization using MSE loss and Adam optimizer [20].

2.3.5 Performance of the method

To evaluate the performance of their BMI estimation model from facial images, they adopted the widely-used regression evaluation metrics MAE [18] and RMSE [21].

The following Table 2.2 shows the obtained results :

Table 2.2: Results on VisualBMI dataset [20].

Dataset	**Model**	**MAE**	**RMSE**
VisualBMI	VGGFace (Reg-GAP)	4.99	6.94
VisualBMI	FaceNet (Reg-GAP)	5.03	6.92

Interpretation

- The Reg-GAP method significantly improved BMI prediction across all datasets.
- Region-based pooling captures local features (eyes, nose, mouth) which are more correlated with BMI.
- Demonstrated superior performance over GAP and classical regression baselines.
- Strong gender classification potential (as visualized via t-SNE plots), though male BMI prediction lagged due to dataset bias [20].

2.4 3rd Approach by: Sidhpura et al. (2022)

2.4.1 Used datasets

- **Illinois DOC Faces:** 56,200 male and 3,649 female images with BMI calculated from height/weight metadata [6].
- **Arrest Records:** 1544 images. BMI mean = 26.41, BMI Standard deviation = 5.2 [6].
- **VIP-Attribute:** 1026 celebrity images (513 male, 513 female) [6] [22].

2.4.2 Data preprocessing

- Cleaning malformed or missing entries.
- Converting height to meters and weight to kilograms.
- Calculating BMI: $\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2}$.
- Face alignment using **DLIB 68 landmark detector**.
- Dataset split for training (Illinois DOC) and evaluation (Arrest Records, VIP-Attribute) [6].

2.4.3 Features extraction

- Transfer learning using pre-trained models:
  - Inception-v3.
  - VGG-Face.
  - VGG-19.
  - Xception [6].

2.4.4 Regression model

The same fully connected layers is used at the end of all pre-trained models. The added layers at the end are shown in the Figure 2.5 :

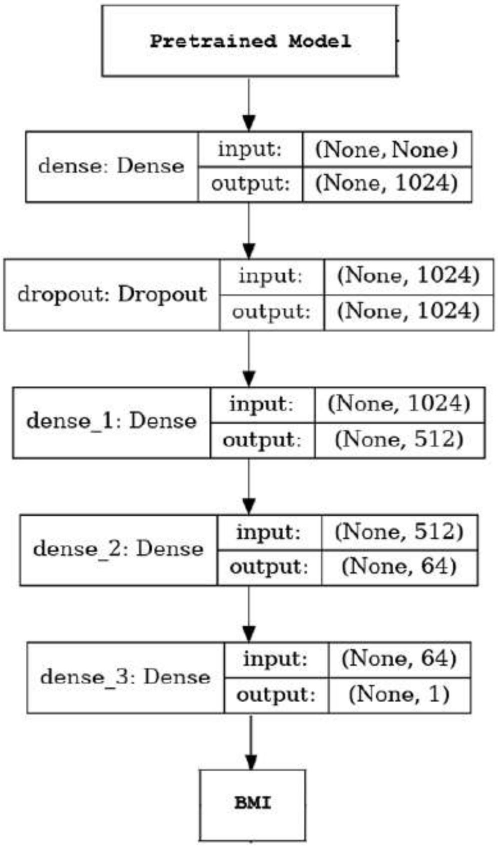


Figure 2.5: Regression model architecture [6].

To prevent overfitting, we added one dropout layer with a dropout of 50% into the model architecture. They also used Gaussian Error Linear Unit (Gelu) as an activation function, it combines the properties of the RELU activation function, Dropout, and Zoneout. Due to this, it tends to generalize better when there is more noise in the data so they used it in our models [6]. The folowwing Figure 2.6 represent the GELU, RELU, ELU activation functions curves.

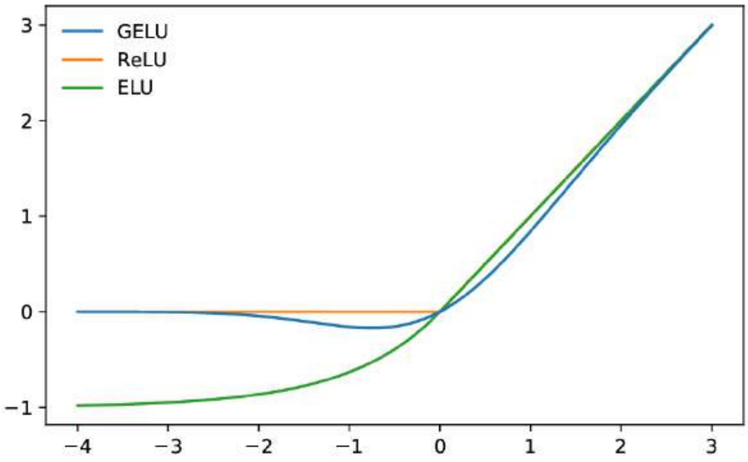


Figure 2.6: GELU, RELU, ELU activation functions curves [23].

As a comparatively larger dataset is found for their study they also fine-tuned the models. In deep convolutional networks, layers near the input learn basic features such as edges and corners. As we move towards the output, the layers generally learn advanced features from the images used for training it. They used a higher learning rate for the new fully connected layers and a much lower learning rate for some of the final layers of the pre-trained model so that it extracts more features from the images. Due to these reasons, Adam optimizers is used (with decreasing learning rates as we move to deeper layers of the model) [6].

They fixed a batch size of 128. To prevent overfitting, Early Stopping Callback is used to stop training if validation loss does not improve for 5 consecutive epochs [6].

2.4.5 Performance of the method

The following Table 2.3 represent the best 02 results on each datasets :

Table 2.3: Testing results on 03 datasets [6].

Dataset	**Model**	**MAE Overall**	**MAE Male**	**MAE Female**
Illinois DOC	Xception	2.82	2.79	3.54
Illinois DOC	Inception-v3	2.86	2.83	3.59
VIP Attribute	Inception-v3	3.10	3.04	3.17
VIP Attribute	VGG-19	3.20	3.33	3.07
Arrest Records	VGG-19	3.79	3.75	3.99
Arrest Records	VGG-Face	3.73	3.35	5.09

Interpretation:

- On the **VIP Attribute Dataset**: The Inception-v3 model gave the best scores on all 03 cases, with very small differences between male and female, indicating the high quality of the extracted features.
- On the **Arrest Records Dataset**: For all models, MAE-Female > MAE-Male. The difference is due to gender imbalance in the training dataset.
- Inception-v3 performed best overall, followed closely by VGG-19.
- Inception-v3 outperform all other pretrained-models, across testing on the 03 datasets.

2.5 Conclusion

All three approaches demonstrate the potential of using facial features for estimating BMI. The first uses basic regression with multitask learning, the second adds region-aware pooling for performance improvement, and the third leverages large-scale transfer learning with multiple models. Each contributes uniquely to BMI prediction from facial images and suggests promising directions for health assessment tools.

In the next chapter, we present our own contribution to this field, including the models we developed, the experiments conducted, and the results obtained on different datasets.

Chapter 3

Datasets for BMI estimation from face images

3.1 Introduction

The accuracy and generalizability of any machine learning or deep learning model depend heavily on the quality and diversity of the datasets used for training and evaluation. In the context of BMI estimation from facial images, having access to representative and well-annotated datasets is essential for building robust and reliable models.

Collecting such datasets, however, presents various challenges. These include ensuring demographic diversity, maintaining consistent image quality, and obtaining accurate labels such as height, weight, and gender. Furthermore, ethical considerations like data privacy and consent must be taken into account, especially when dealing with facial images.

Fortunately, several publicly available datasets have emerged in recent years, enabling researchers to experiment with and benchmark their models in this domain. Each dataset comes with its own structure, characteristics, and limitations, which must be thoroughly understood to choose the appropriate dataset for the task.

In this chapter, we present the different datasets used in our study, provide detailed descriptions of their sources, structures, and statistics, and explain how we preprocessed them to suit our BMI estimation task.

3.2 Selected datasets

We were able to get and download 03 publicly available and very famous datasets which are:

- **Illinois DOC labeled faces Dataset** [24] [6].
- **VisualBMI dataset** [25] [2] [20].
- **Arrest Records Dataset** [26] [7].

3.2.1 VisualBMI dataset

- These images are collected from **Reddit posts** that link to the imgur.com service, with examples of the underlying Reddit posts available in the 'progresspics' subreddit [27].

- The VisualBMI dataset comprises a total of 16,483 images containing pairs of 'before' and 'after' images, annotated with gender, height, and previous and current body weights. all image URLs are manually processed and the faces are cropped, retaining only images with two faces since we required both previous and current body weight information. After this manual cleaning process, a 2,103 pairs of faces obtained with corresponding gender, height, and previous and current body weights.

- Here are 08 images token from the VisualBMI dataset in this Figure 3.1:

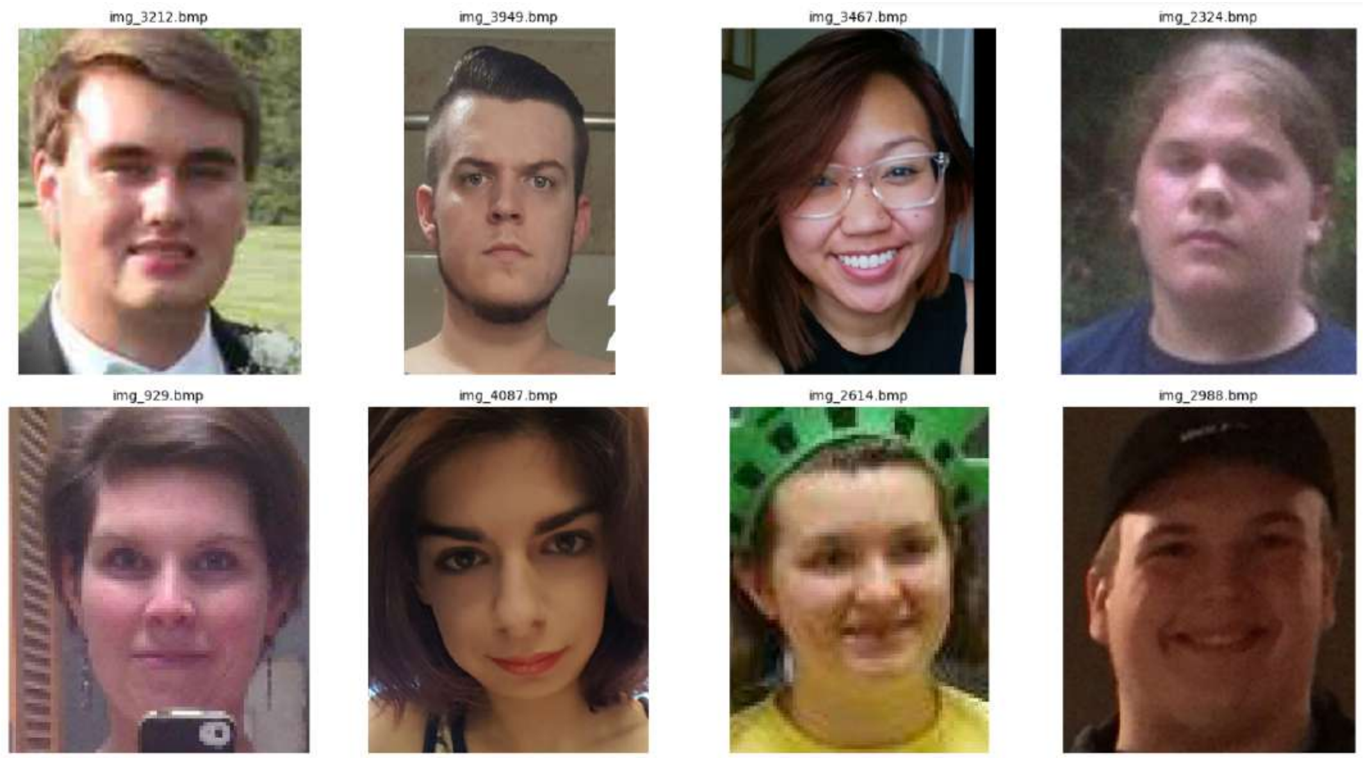


Figure 3.1: VisualBMI dataset images.

- This process yielded a total of 4,206 faces with corresponding gender and BMI information. The distribution of BMI categories in our dataset is as the Figure 3.2 shows:

  - Underweight (BMI $<$ 18.5): 7 faces
  - Normal weight (18.5 $\leq$ BMI $<$ 25): 680 faces.
  - Overweight (25 $\leq$ BMI $<$ 30): 1,151 faces.
  - Moderately obese (30 $\leq$ BMI $<$ 35): 941 faces.
  - Severely obese (35 $\leq$ BMI $<$ 40): 681 faces.
  - Very severely obese (BMI $\geq$ 40): 746 faces.

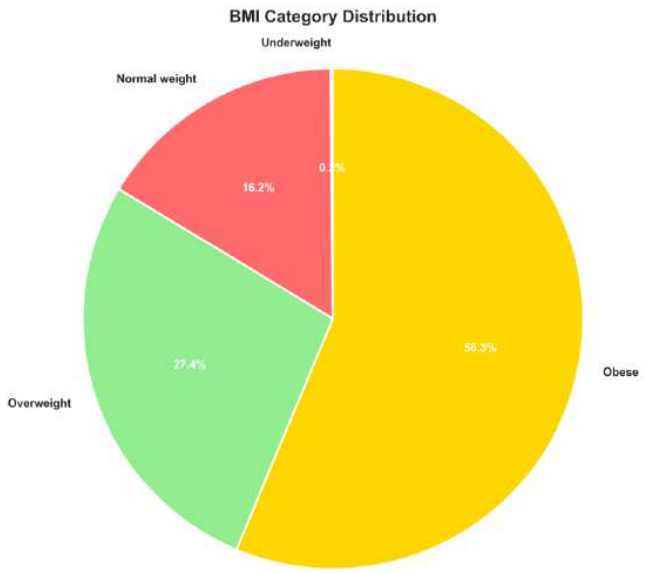


Figure 3.2: BMI category statistics of VisualBMI dataset.

- The dataset contains **2,438 male** and **1,768 female** subjects as the Figure 3.3 of gender distribution shows:

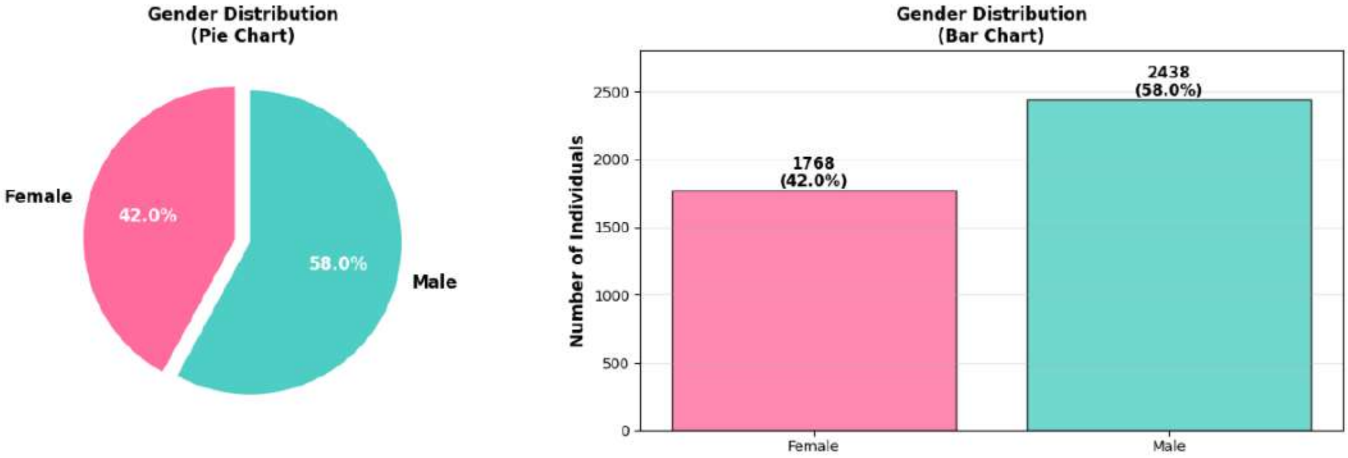


Figure 3.3: Gender distribution of VisualBMI dataset.

3.2.2 Arrest Records dataset

- This small dataset consists of 1544 person images was scraped from the web. Each containing demographic information including: name, location, age, height, weight, race, sex, eye color, and hair color, along with corresponding facial photographs.

- We performed extensive data pre-processing to ensure data quality. Height measurements were converted from imperial format (feet and inches) to metric units, and weight measurements were converted from pounds to kilograms.

- Here are 08 images token from the Arrest Records dataset in this Figure 3.4:

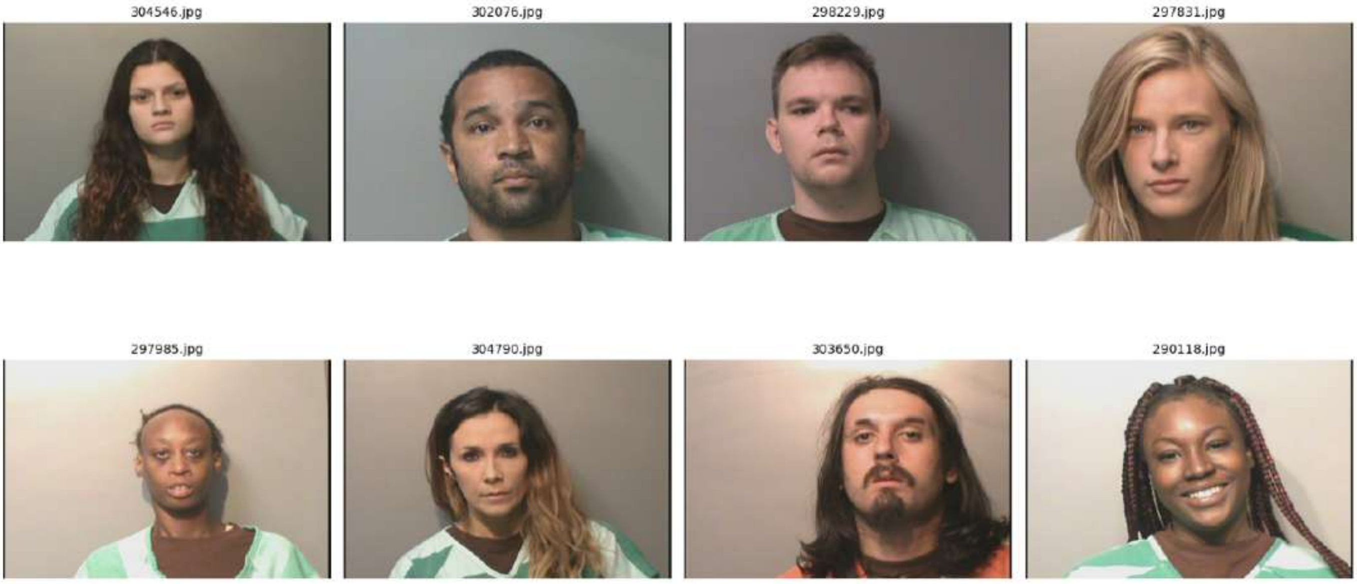


Figure 3.4: Arrest Records dataset images.

- The BMI values in our dataset range from 16.95 to 68.35, with a **mean BMI equal to 26.32** and a **standard deviation of 5.35**. The distribution of BMI categories in our dataset is as follows and as the Figure 3.5 shows:

  - Underweight ($BMI < 18.5$): 60 individuals.

  - Normal weight ($18.5 \leq BMI < 25$): 505 individuals.

  - Overweight ($25 \leq BMI < 30$): 600 individuals.

  - Very severely obese ($BMI \geq 30$): 379 individuals.

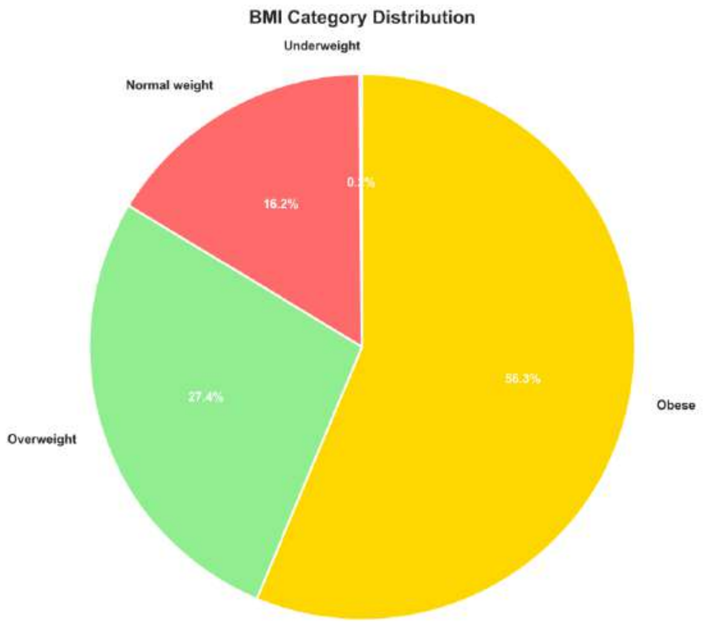


Figure 3.5: BMI category statistics of Arrest Records dataset.

- The dataset contains **1,244 male** and **300 female** subjects as it is represented by the Figure 3.6, with ages ranging from 18 to 73 years (average age: 34.7 years).

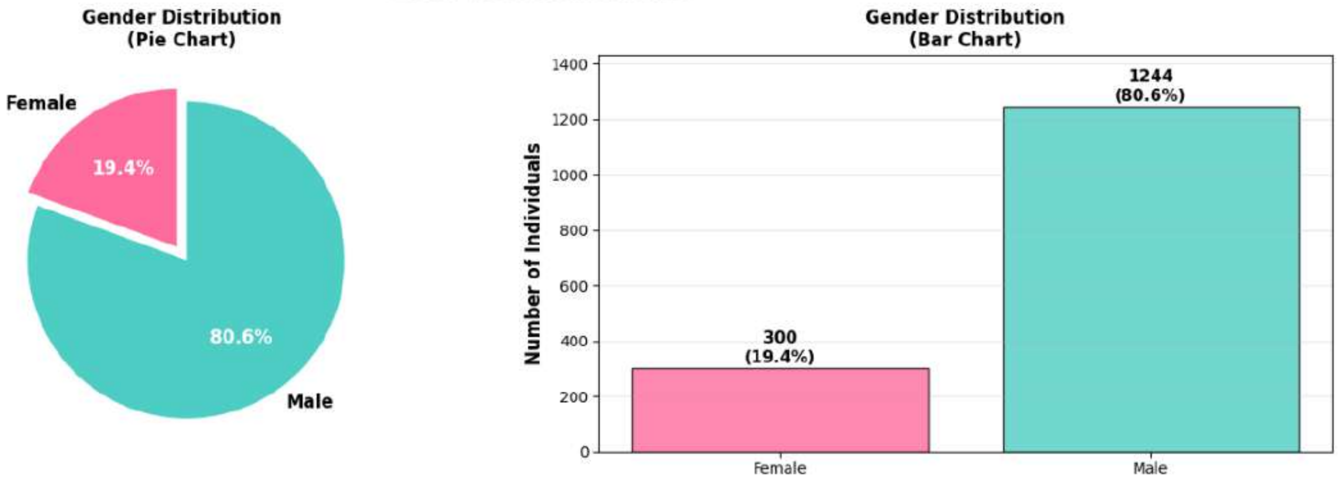


Figure 3.6: Gender distribution of Arrest Records dataset.

3.2.3 Illinois DOC labeled faces dataset

- The source of this dataset is the Illinois Dept. of Corrections. This dataset comprises frontal and side views of **68149 prisoners** and has additional information such as gender, height (in inches), weight (in lbs), and date of birth.

- Here are 08 images token from the Illinois DOC labeled faces dataset in this Figure 3.7:

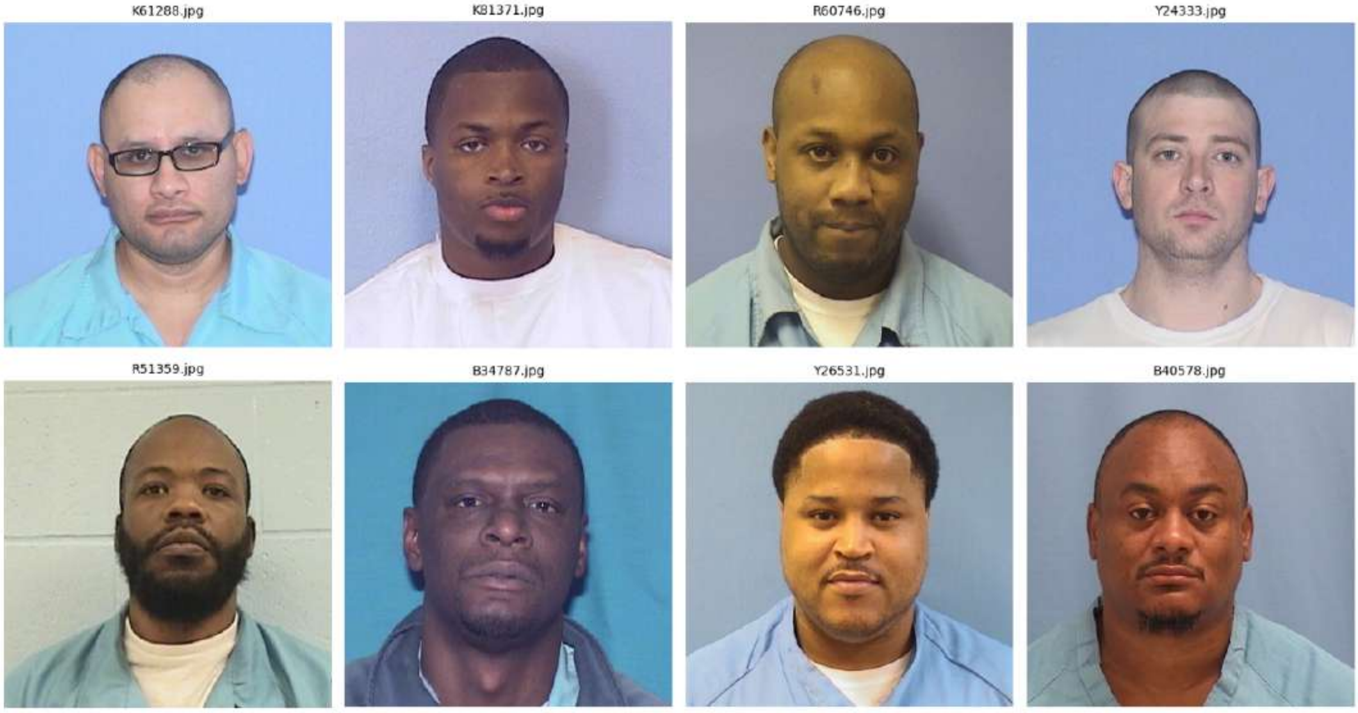


Figure 3.7: Illinois DOC labeled faces dataset images.

- There were **1365 corrupt images** and **7309 images** that did not have the height and weight information. They were not included in our research study. The dataset in the end we used had 56200 males and 3649 females, with a **mean BMI equal to 27.88** and a **standard deviation of 5.20**.

- The statistics of BMI categories in our entire dataset is represented by the Figure 3.8 as follows:

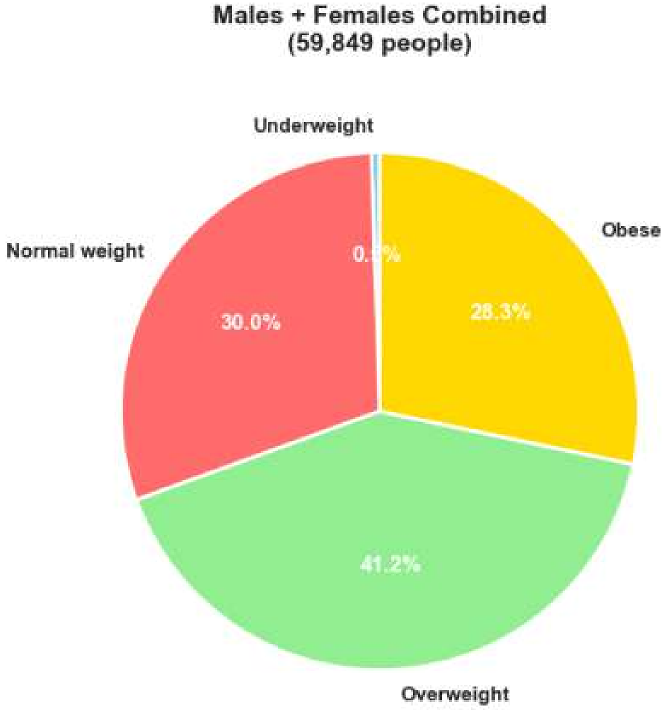


Figure 3.8: BMI category statistics of Illinois DOC labeled faces dataset.

As shown, the combined dataset (males and females) includes:

- 41.2% Overweight,
- 30% Normal weight,
- 28.3% Obese,
- 0.5% Underweight.

- The statistics of BMI categories in our dataset (devided by 02 : Male / Female) is represented by the Figure 3.9 as follows:

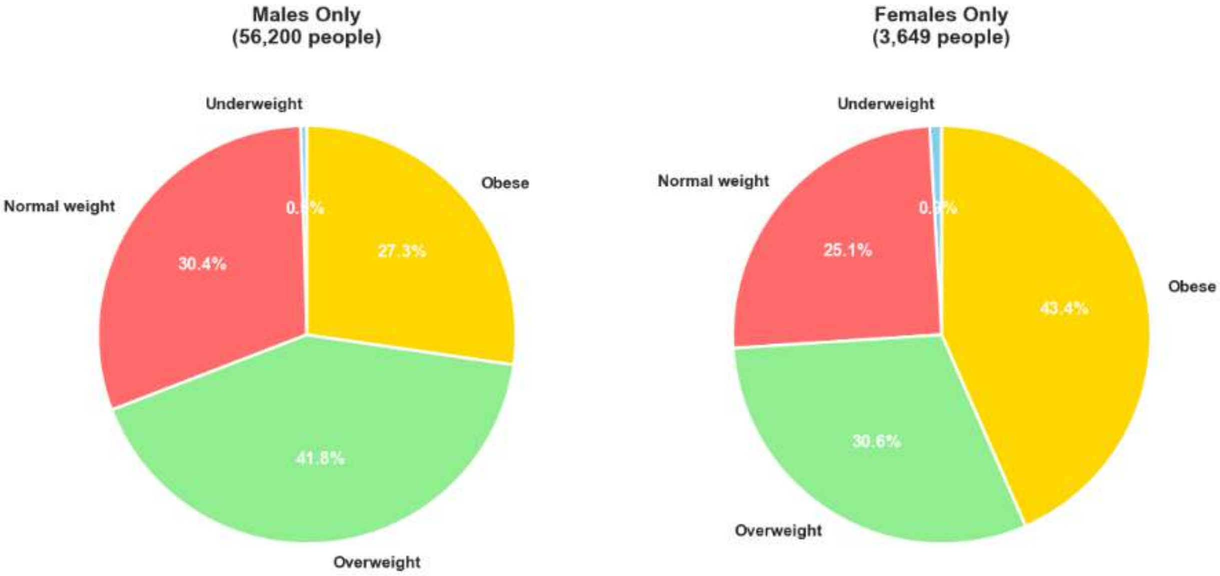


Figure 3.9: BMI category statistics of Illinois DOC labeled faces dataset by gender.

- **Males**: 41.8% Overweight, 30.4% Normal weight, 27.3% Obese, 0.5% Underweight.
- **Females**: 43.4% Obese, 30.6% Overweight, 25.1% Normal weight, 0.9% Underweight.

We observe that obesity is significantly higher among females, while males have a higher proportion of overweight individuals.

- This Figure 3.10 represent the Gender distribution in the entire dataset:

3.3 Conclusion

In this chapter, we introduced the three main datasets used in our study: VisualBMI, Arrest Records, and Illinois DOC labeled faces. Each dataset brings its own characteristics, such as diversity in age, gender, and BMI distribution, which are essential for training a robust and generalizable model.

We explored the sources of the data, the types of annotations available, and the

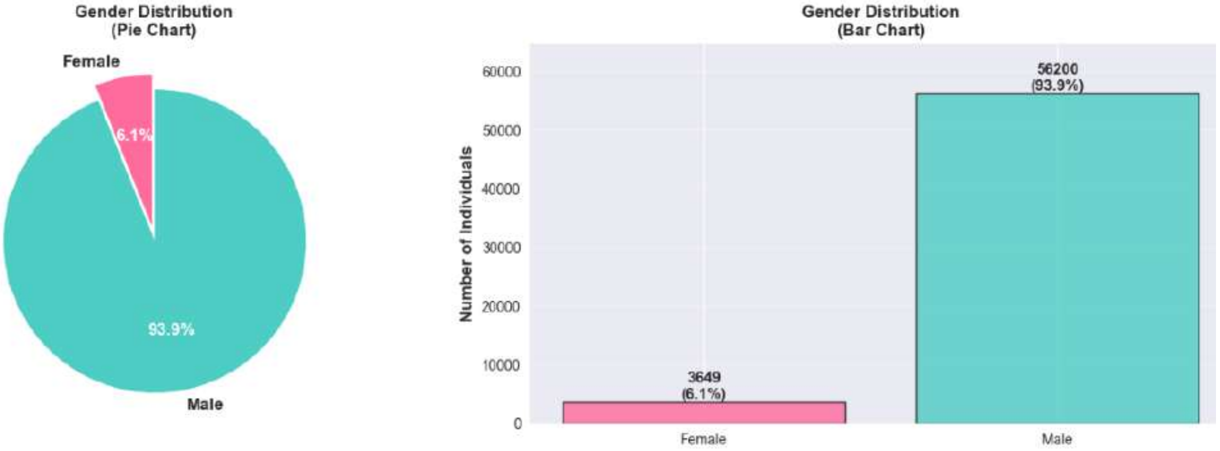


Figure 3.10: Gender distribution of Illinois DOC labeled faces dataset.

BMI categories represented. These datasets vary in size, structure, and data quality, but together they provide a solid foundation for building and testing our BMI estimation models.

We also highlighted the importance of preprocessing and cleaning steps to ensure data consistency and reduce noise, as well as the significance of gender and category balance when developing health-related machine learning applications.

The richness and variability of the datasets play a crucial role in the model's ability to learn meaningful patterns and make accurate predictions.

In the next chapter, we will delve into the technical approach used in our work, including preprocessing methods, model selection, training procedures, and evaluation metrics for estimating BMI from facial images.

Chapter 4

Our approaches for BMI estimation from face images

4.1 Introduction

Estimating Body Mass Index (BMI) from facial images offers a non-invasive alternative to traditional methods, with potential applications in telemedicine and digital health. Recent advances in deep learning, including CNNs and Vision Transformers, have enabled progress in facial analysis, yet BMI prediction remains challenging due to limited datasets, subtle feature correlations, and demographic variability.

In this chapter, we present our contributions based on three diverse datasets (VisualBMI, Arrest Records, and Illinois DOC), combining advanced preprocessing, novel model architectures, and data augmentation techniques. We also explore gender-based modeling and multi-task learning to improve accuracy and generalizability.

The chapter details our experimental setup, model design, training strategies, and a comparative evaluation of the results across all datasets.

4.2 Our methodology

- We selected 03 of the most important and recent approaches from the state of the art . We were able to obtain and download 03 publicly available and widely used datasets in this field:

- **Illinois DOC labeled faces dataset** [24] [6].
- **VisualBMI dataset** [25] [20].
- **Arrest Records dataset** [26] [7].

- Our objective in this work is to analyze results across these 03 datasets by studying several criteria, including: model generalization capacities, gender separation impact, pre-trained model impact, impact of internal model parameters (such as data augmentation and model fine-tuning), impact of MTCNN usage for face detection and alignment from the whole image, and impact of the nature of Dataset images (like: Resolution variety, face poses ..ext).

4.2.1 Model selection:

The model have been selected based on the following main considerations: model complexity, dataset size, and available computing device (GPU, CPU).

1/- For the **VisualBMI dataset**, we chose to work with the large ViT-H-14 vision transformer model (with 633M parameters, indicating high complexity), which operates differently from CNN models. We wanted to test its performance on this task, particularly after our research on this transformer model revealed that Vision Transformers have recently emerged as competitive alternatives to Convolutional Neural Networks (CNNs), outperforming current state-of-the-art CNNs by almost 4x in terms of computational efficiency and accuracy in some computer vision tasks [28]. This model is applied to the VisualBMI dataset, which is relatively smaller. We could not apply it to a large dataset like the Illinois DOC Faces dataset because this process would require high computational cost (dramatically increased training time and very powerful hardware such as GPUs/TPUs), and our GPU was fully utilized most of the time.

2/- For the **Arrest Records dataset**, we chose VGG16 (138M parameters) and ResNet50 (25M parameters) pre-trained models, which are less complex models, particularly ResNet50. We worked on this dataset using the CPU device because the GPU was frequently occupied by ourselves or other users. This made the usage of these models very compatible with this small-sized dataset. We could have used the Inception-v3 model again, but we chose not to in order to test the performance of as many different models as possible. Due to these computational constraints, we were also unable to increase the size of our datasets via data augmentation techniques, which would have required high computational cost.

3/- For the **Illinois DOC labeled faces dataset**, we have chosen the Inception-v3 as a pretrained model for the feature extraction step, which is a relatively lightweight model (27 million parameters — that means it's not a particularly complex model.), which makes it compatible with this large-sized dataset, also due to its proven performances in recent research papers.

4.2.2 Dataset usage methodology:

- Due to the comparatively large size of the Illinois DOC Faces Dataset, we used it as a complete dataset initially, and then split it by gender (male and female) to verify or deduce certain properties.
- For the other two datasets, we could not apply the same approach due to their smaller sizes.

4.3 Experiment 01 : ViT-H/14 + Visual BMI

4.3.1 Model description

The system first processes input images from the VisualBMI dataset through MTCNN (Multi-task Convolutional Neural Network) for face detection and alignment. The detected and aligned faces are then fed into a ViT-H/14 (Vision Transformer) model for deep feature extraction, capturing high-level facial characteristics. Finally, a regression layer maps these extracted features to BMI predictions, enabling the estimation of body mass index directly from facial appearance, This is shown in the

Following Figure 4.1:

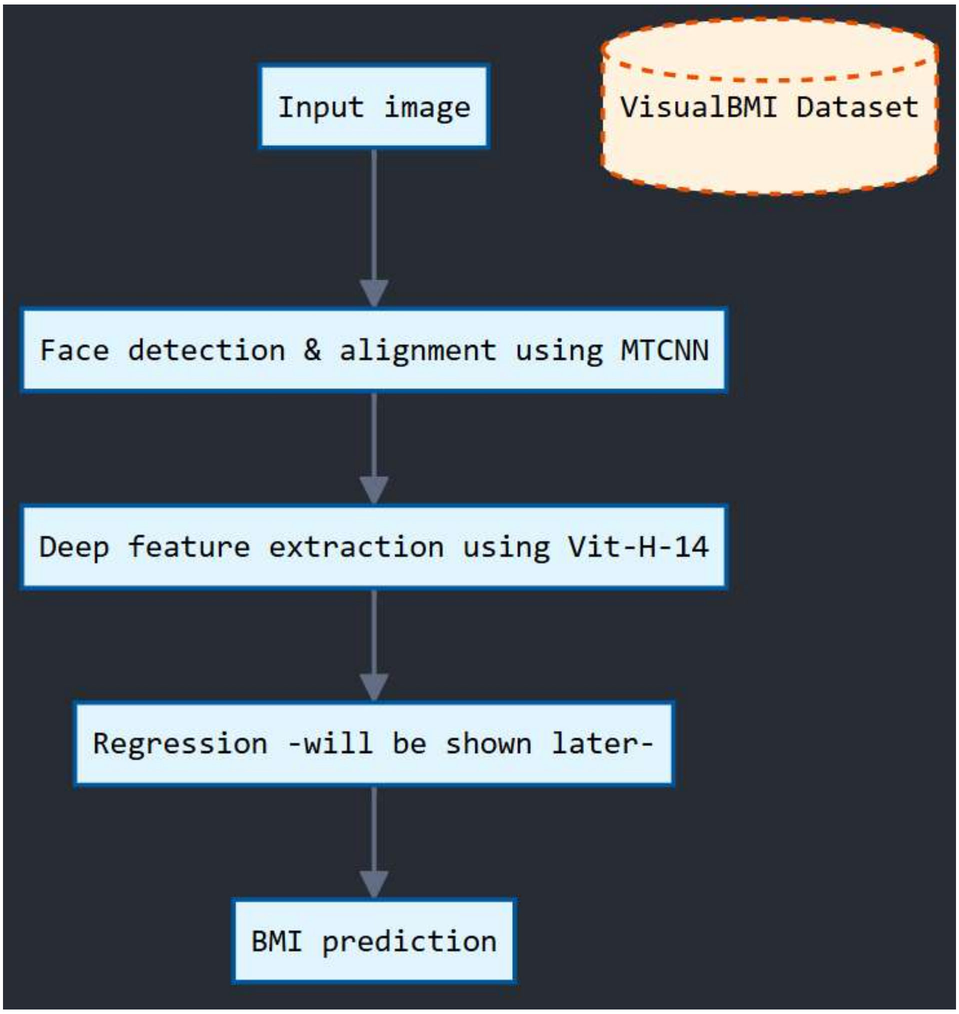
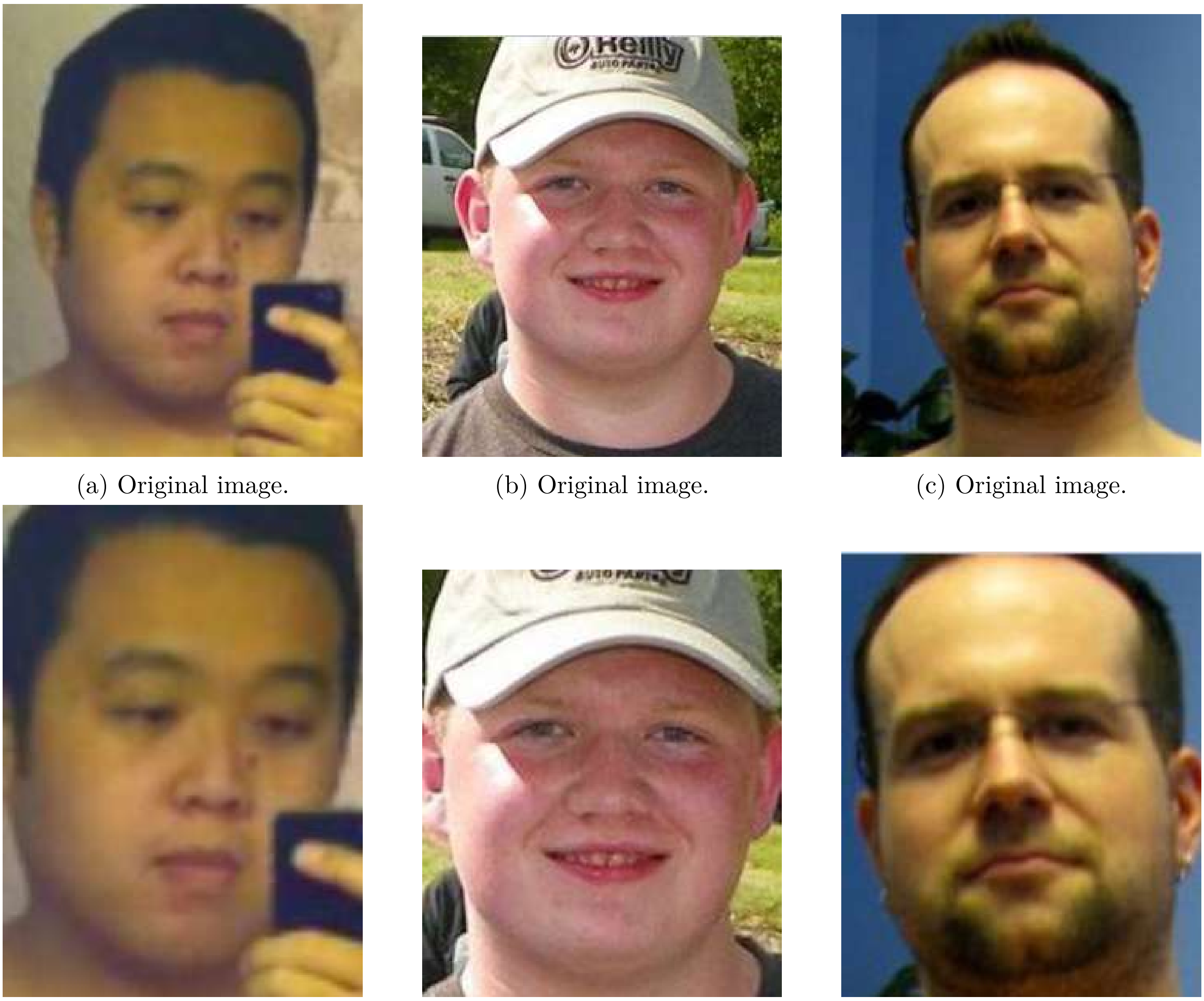

Input image
VisualBMI Dataset
Face detection & alignment using MTCNN
Deep feature extraction using Vit-H-14
Regression -will be shown later-
BMI prediction

Figure 4.1: Model description.

4.3.2 Data preprocessing

- Splitting the dataset: The first 3368 images are used for training (80%) and the rest of the images are used for testing (like is done on the state of the art previous work).
- Using the **MTCNN face detection model**, to detect and cropp only the region of interest (face region), from the whole image.

- The following Figure 4.2 showing the MTCNN usage on Visual BMI dataset :



(a) Original image.

(b) Original image.

(c) Original image.

(d) Preprocessed with MTCNN.

(e) Preprocessed with MTCNN.

(f) Preprocessed with MTCNN.

Figure 4.2: Original images / Preprocessed images with MTCNN.

4.3.3 Model architecture

- Our method for predicting Body Mass Index (BMI) from Visual BMI dataset images utilizes a Vision Transformer (ViT) as a feature extractor, specifically the **vit-h-14** model from the torchvision library, extracting a feature vector of **size = 1280**. This backbone is chosen for its powerful capability in modeling global dependencies in image data.

Vit-H-14 model architecture:

Vision Transformers (ViT) shown in the Figure 4.3 have recently emerged as a competitive alternative to Convolutional Neural Networks (CNNs) that are currently state-of-the-art (SOTA) in different image recognition and computer vision tasks. ViT models outperform the current SOTA CNNs by almost **x4 in terms of computational efficiency and accuracy** [28].

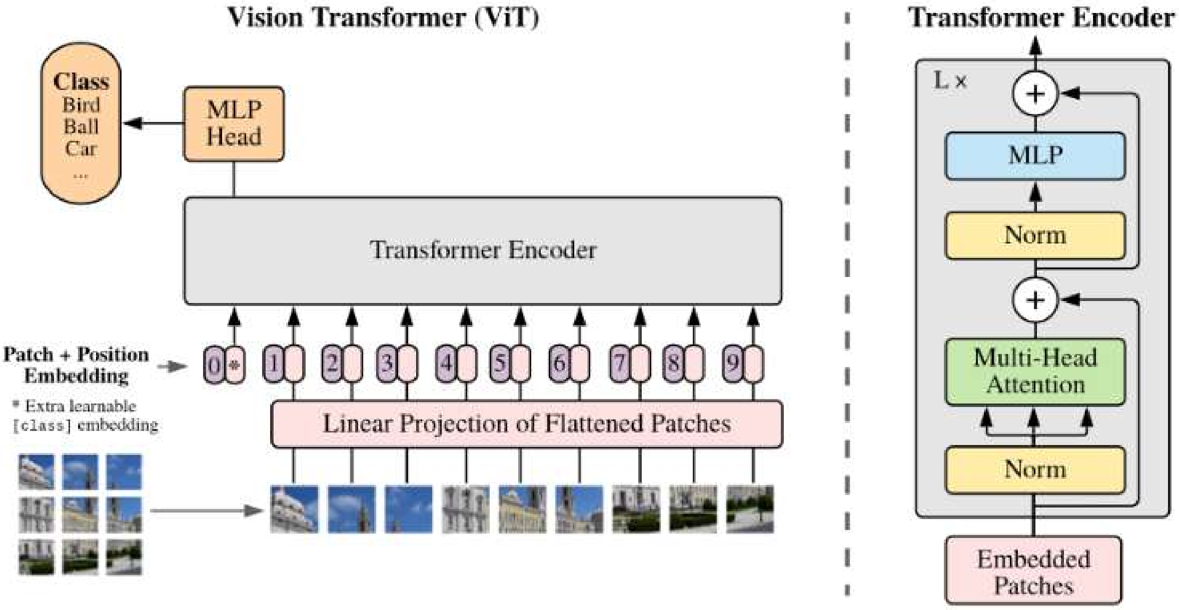


Figure 4.3: Vision transformer ViT architecture [29].

Transformer models have become the de facto status quo in Natural Language Processing (NLP). For example, the popular ChatGPT AI chatbot is a transformer-based language model. Specifically, it is based on the GPT (Generative Pre-trained Transformer) architecture. This uses **self-attention mechanisms** to model the dependencies between words in a text [28].

A transformer in machine learning is a deep learning model that uses the mechanisms of attention, differentially **weighing the significance of each part of the input sequence of data**. Transformers in machine learning are composed of multiple self-attention layers. They are primarily used in the AI subfields of natural language processing (NLP) and computer vision (CV) [28].

1/- Difference between CNN and ViT (ViT vs. CNN):

The ViT is a visual model based on the architecture of a transformer originally designed for text-based tasks. The ViT model **represents an input image as a series of image patches**, like the series of word embeddings used when using transformers to text, and directly predicts class labels for the image. ViT exhibits an extraordinary performance when trained on enough data, breaking the performance of a similar SOTA CNN with 4x fewer computational resources [28].

These transformers have high success rates when it comes to NLP models and are now also applied to images for image recognition tasks. **CNN uses pixel arrays**, whereas ViT splits the input images into visual tokens. The visual transformer divides an image into fixed-size patches, correctly embeds each of them, and includes positional embedding as an input to the transformer encoder [28].

The self-attention layer in ViT makes it possible to embed information globally across the overall image. The model also learns from training data to encode the relative location of the image patches to reconstruct the structure of the image [28].

- The Vision Transformer (ViT) model architecture was introduced in a research paper published as a conference paper at ICLR 2021 titled “An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale”. It was developed and published by Neil Houlsby, Alexey Dosovitskiy, and 10 more authors of the Google Research Brain Team [28].

The fine-tuning code and pre-trained ViT models are available on the GitHub of the Google Research team. The ViT models were pre-trained on the **ImageNet** and **ImageNet-21k** datasets [28].

2/- How does ViT-H/14 work ?

Vit-H-14 meaning:
ViT = Vision Transformer , **H** = Huge (Model Size) , **14** = Patch Size.

- Embedding:

In this step, the input image is devided into fixed-size patches of [P, P] dimension like it represented in the Figure 4.4, and linearly flatten them out, by concatenating the channels.

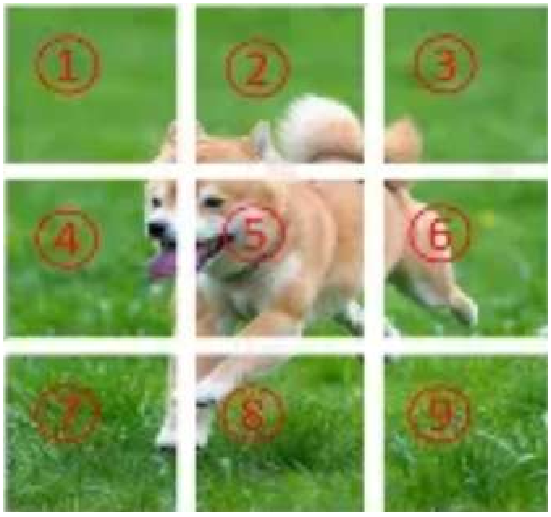


Figure 4.4: Image devided into fixed-size patches [30].

For example, a patch of size [P, P, C]=[14*14*3] is converted to **[P*P*C, 1]**. This linearly flattened patch is further passed through a Feed-Forward layer with a linear activation function to get a linear patch projection of the dimension **[D, 1]**. D is the hyperparameter called as embedding dimension used throughout the transformer [30], as its shown in the Figure 4.5.

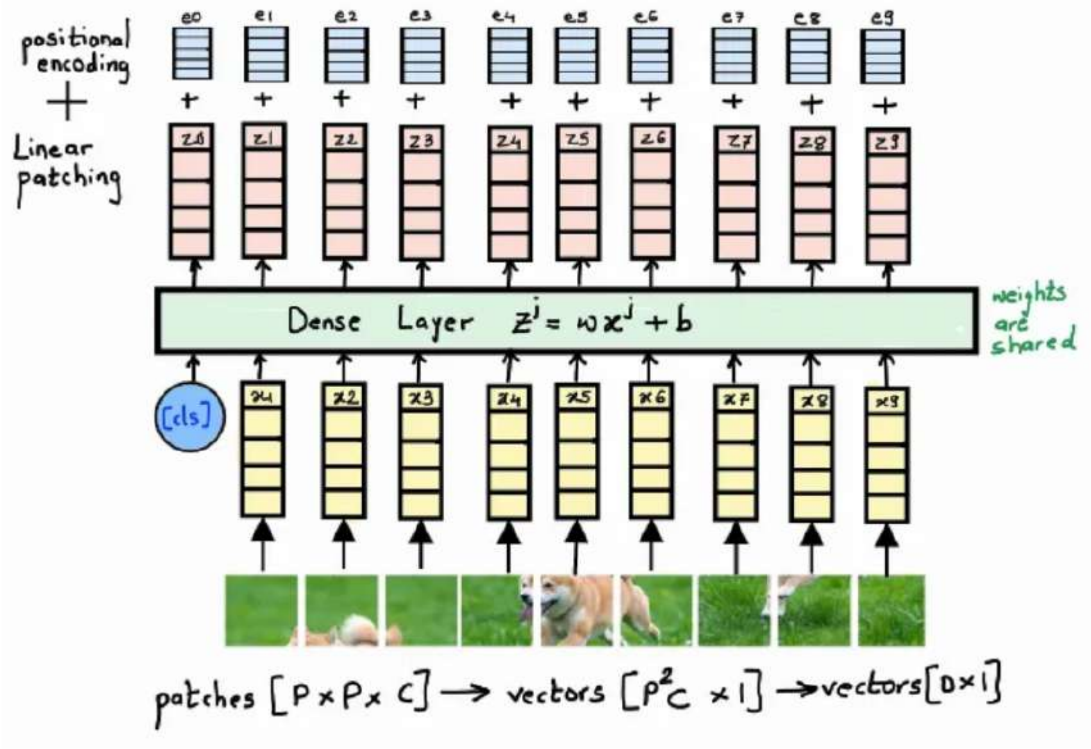


Figure 4.5: Embedding step [30].

For classification purposes, taking inspiration from the original BERT paper, we concatenate a **learnable class embedding** with the other patch projections, whose state at the output serves as class information. This extra **class token** is added to the set of image tokens which is responsible for **aggregating global image information and final classification**. It is able to learn this global aggregation while it passes and learns through the **attention layers** (The class token learns how to selectively gather and combine information from patches. For example, it might prioritize certain patches (like a dog's head) over others (like the background) [30].

1D positional embedding is added also to the linear patches, to establish a certain order in the input patches [30].

- Why is positional encoding necessary?

Transformers are not capable of remembering the order or sequence of the inputs. If the image patches are re-ordered the meaning of the original image is lost, as it is shown in the Figure 4.6. Hence, we add a positional embedding to our linearly embedded image patches to keep track of the sequence [30].

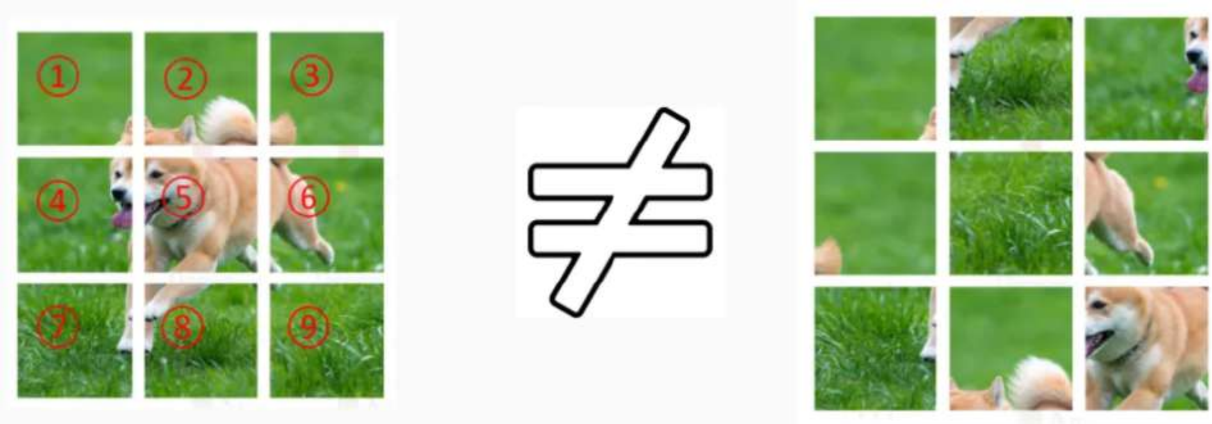


Figure 4.6: Necessity of patches order [30].

To understand the embedding step a bit better let us see the dimensions:

Suppose, we have an input image of size 224x224x1, we divide it into fixed-size patches of size 16x16. Let us denote the patch size as P and the image channels as C. The total number of patches N that we get is 196 [30].

After linearly flattening all the patches to get a vector X of dimension [N, P^2C], we pass it through a Dense Layer to convert it to a D dimensional vector called embedding E [**N, D**]. We then append a **learnable class embedding [1, D]** to convert the E vector to dimension [**N+1, D**]. The last step is adding positional encoding **to get the final vector Z**, as it's shown in he Figure 4.7. Both the **class and positional embeddings** are randomly initialized vectors, **learned during the training** of the network [30].

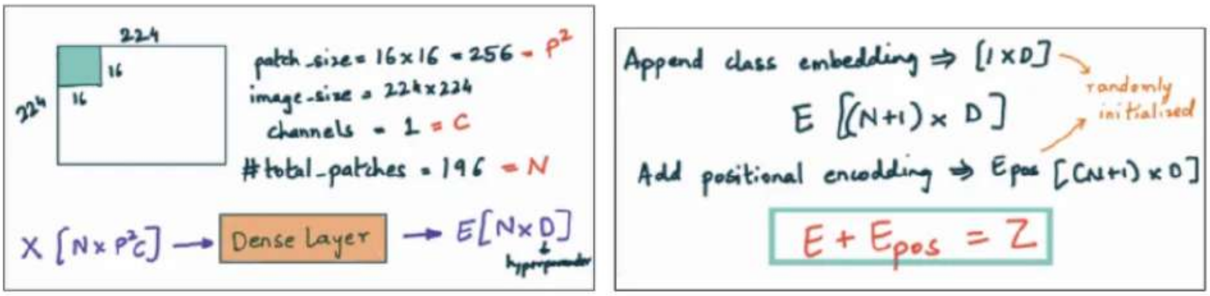

224
224
16
16
patch_size = 16x16 = 256 = P²
image-size = 224x224
channels = 1 = C
#total_patches = 196 = N
X [N x P²C] → Dense Layer → E[N x D]
hyperparameter
Append class embedding ⇒ [1 x D]
E [(N+1) x D]
randomly initialised
Add positional encodding ⇒ Epos [(N+1) x D]
E + Epos = Z

Figure 4.7: Calculation of dimensions example [30].

– Once we have our vector Z we pass it through a Transfomer encoder layer.

- Transformer encoder:

The Transformer Encoder is core component of the model, the architecture consists of multiple of **identical encoder blocks**, where each block has a **Multi-Head Attention unit** and a **Feed-Forward Network**. Each layer is also followed by a **normalization layer**, to stabilize training and improve convergence [30].

Note: The attention output is added back to the original input (Residual Connection), that's mean: **skip connections**.

The Z vector from the previous step is passed through the transformer Encoder architecture to get the context vector C [30], as it's shown in the following Figure 4.8.

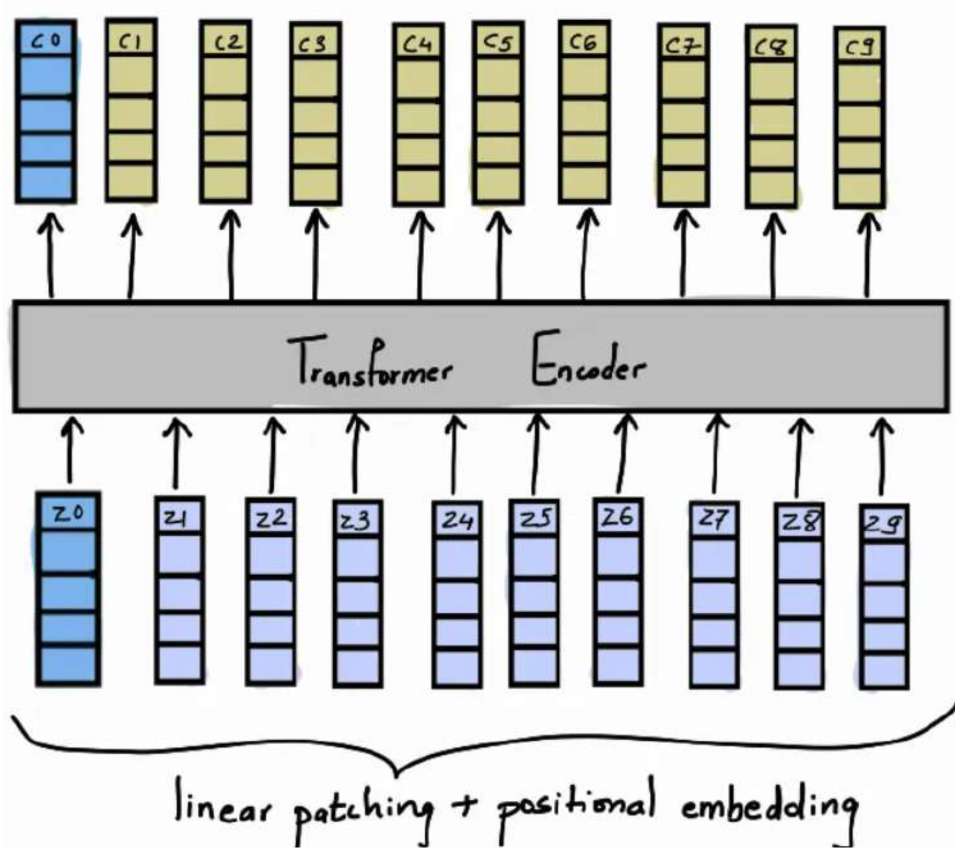


Figure 4.8: Transformer encoder block [30].

- Multi-head attention:

The main component of a Multi-Head Attention unit is the **Scaled Dot-Product Attention**. At first, the input vector Z is duplicated 3 times and multiplied by weights **Wq**, **Wk**, and **Wv**, to get the **Queries**, **Keys**, and **Values** respectively, like its shown in the Figure 4.23. The Queries are then multiplied by the Keys, and the result is divided by the square root of the dimension, to avoid the vanishing gradient problem. This matrix goes through a **Softmax layer** and gets multiplied by the Values to give us the final output called **Head H** [30], these steps are done and represented in the Figure 4.9.

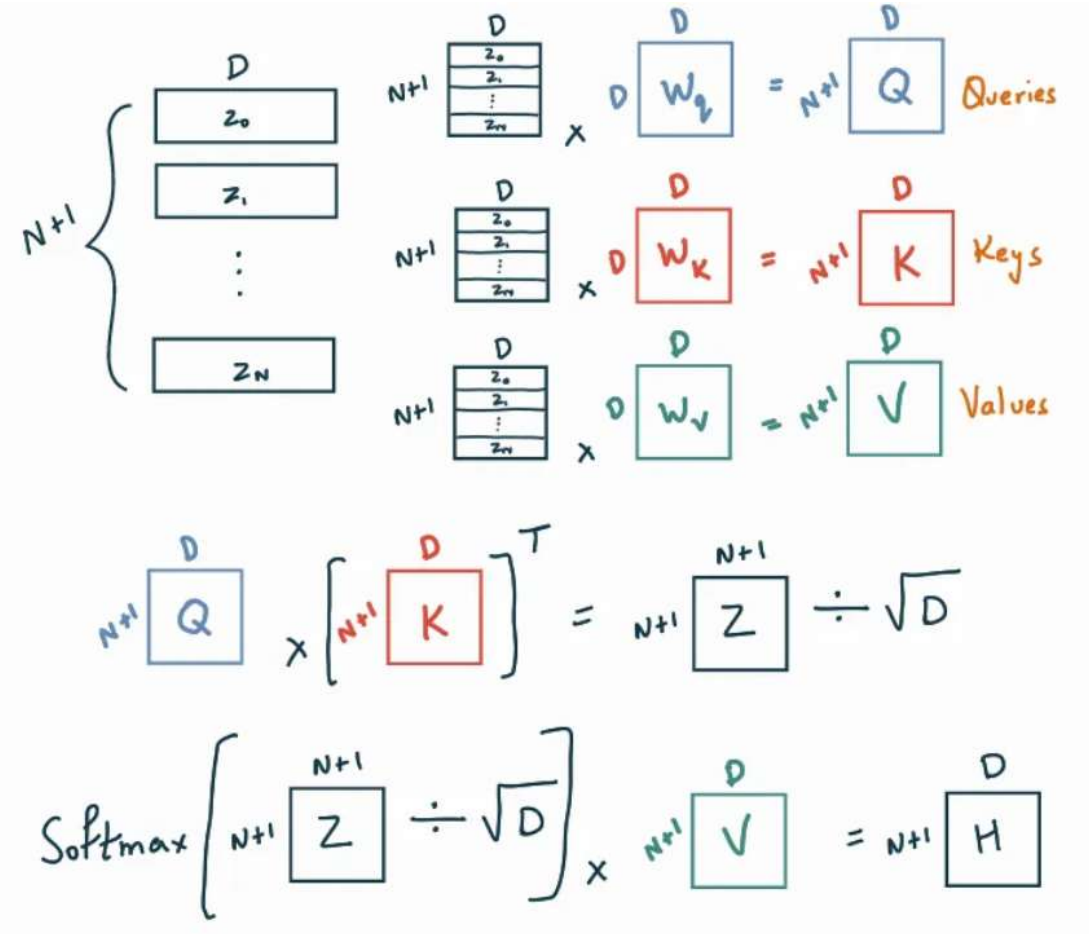


Figure 4.9: Dimension of multi-head attention block [30].

The Scaled Dot-Product Attention as explained above is applied h times (h=8) to get h attention heads. These attention heads are concatenated and passed through a dense Layer to get the final vector of embedded dimension D [30], as the Figure 4.10 shows.

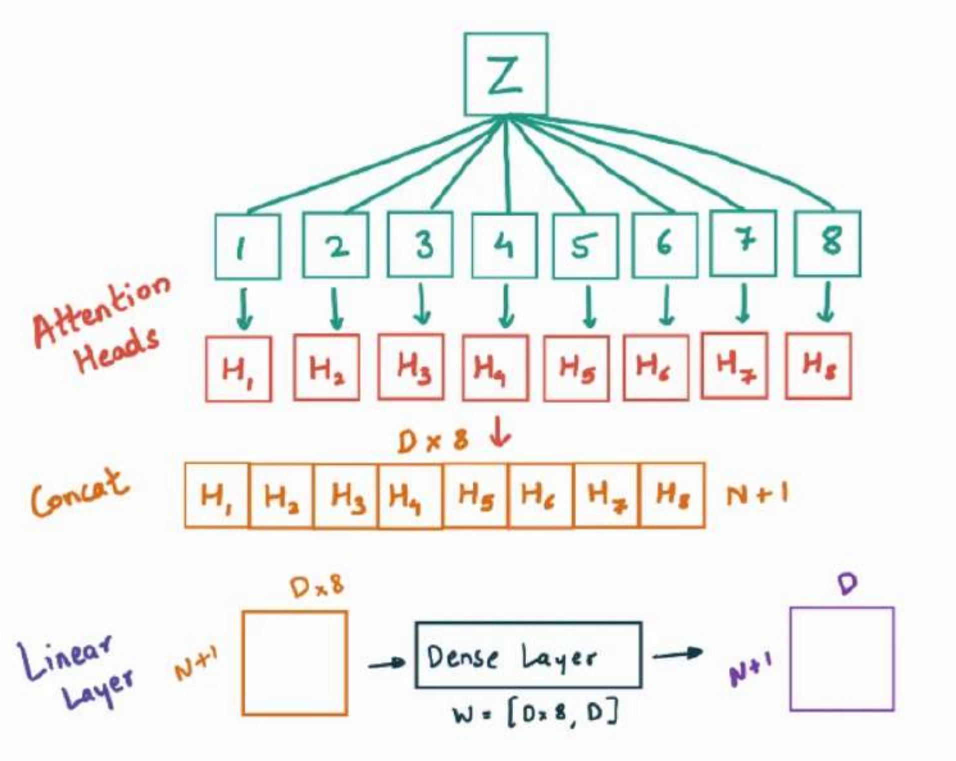


Figure 4.10: Dimensions for attention block [30].

Note: The CLS token learns to represent the entire image by attending to the different patches through the **self-attention mechanism**. At the output of the transformer layers, the CLS token is extracted and passed to a classifier for the final prediction.

- MLP head:

After the transformer encoders process the sequence of patches and the CLS token (class token c0), the output corresponding to the CLS token is used for classification [30].

The output of the CLS token is fed into an MLP, typically consisting of one or two fully connected layers. A softmax layer is applied at the end of the MLP for classification tasks, predicting the image's label [30].

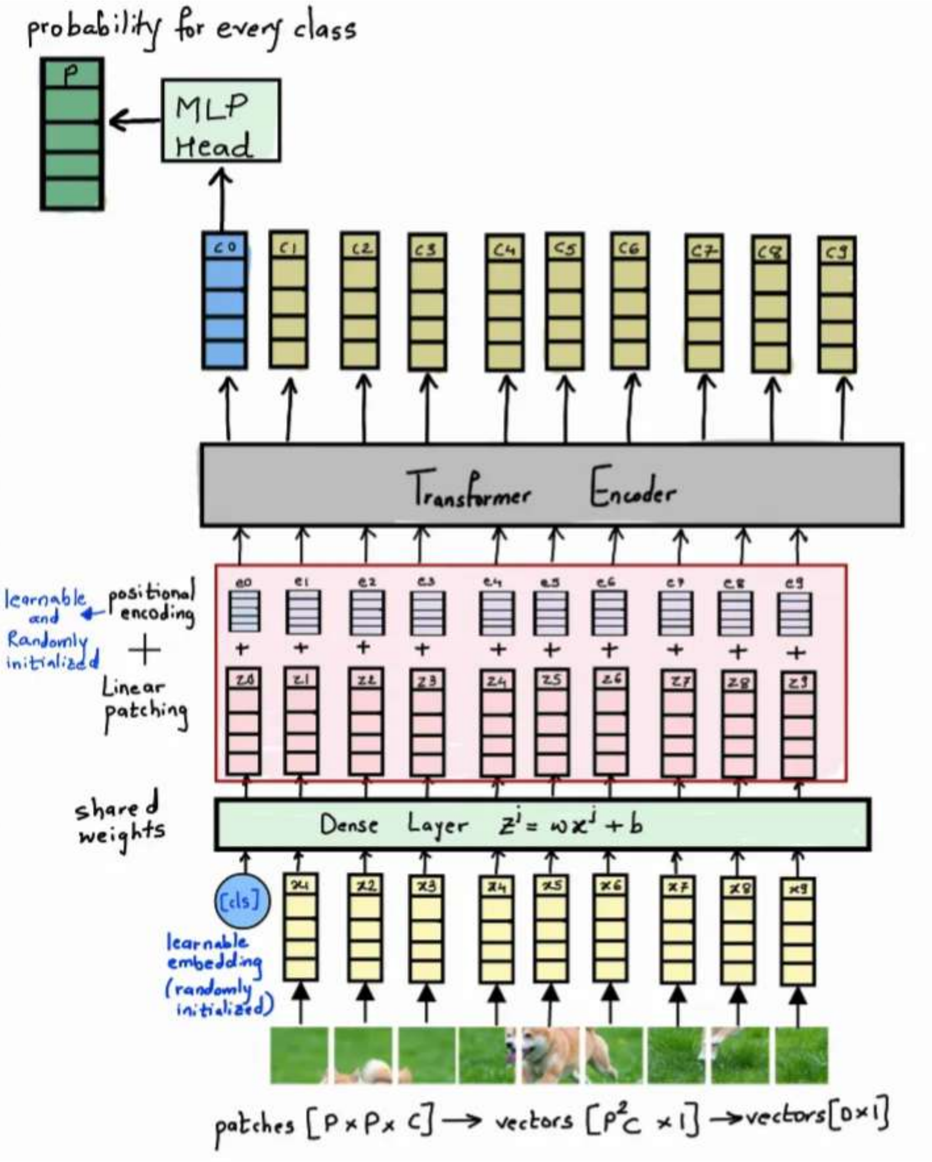

probability for every class
P
MLP Head
c0
c1
c2
c3
c4
c5
c6
c7
c8
c9
Transformer Encoder
learnable and Randomly initialized
positional encoding
+
Linear patching
e0
e1
e2
e3
e4
e5
e6
e7
e8
e9
z0
z1
z2
z3
z4
z5
z6
z7
z8
z9
shared weights
Dense Layer $z^i = wx^i + b$
[cls]
x1
x2
x3
x4
x5
x6
x7
x8
x9
learnable embedding (randomly initialized)
patches [P × P × C] → vectors [P²C × 1] → vectors [D × 1]

Figure 4.11: Complete ViT architecture [30].

- To adapt the ViT model for the regression task of BMI prediction, we replaced its classification head with a custom regression head, which is a deep fully connected layers consisting of:

05 layers with decreasing dimensionalities: 1280 → 640 → 320 → 160 → 80 → 1. This architecture is shown in this Figure 4.12 :

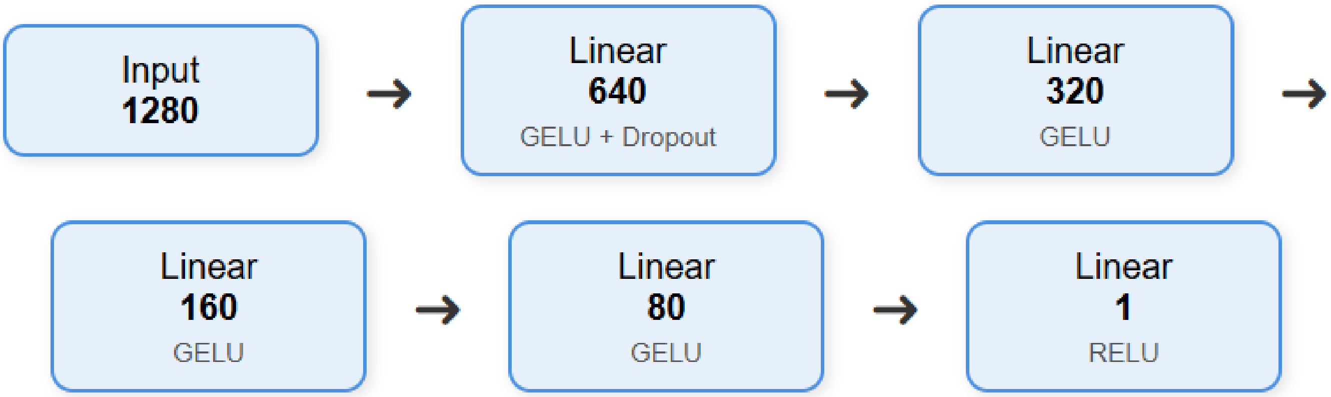


Figure 4.12: Regression model architecture.

- **GELU** activation function after each layer, except the last one **RELU**, because the BMI has a positive range values.

- To prevent overfitting, we added one dropout layer, with a dropout of 50% into our model architecture, which applied only after in first linear layer. It's usually more effective to regularize at the beginning of deep networks (the largest part of the network, where overfitting is most likely) than in later narrow layers (where underfitting is most likely).

- The final output is a single scalar value representing the predicted BMI.

- We have leveraged the pretrained knowledge, all parameters (632 M parameters) of the vit-h-14 model are one time frozen and other time are trainable during training. In all cases the parameters of the custom BMI-Head are updated.

4.3.4 Training and optimization

The models are trained using the **Mean Squared Error (MSE)** loss function, appropriate for regression tasks, for **30 epochs**.

We have fixed the **batch size = 16**.

The optimization is handled using the **Adam optimizer** with a learning rate of **1e-3** and **1e-4** in case of fine-tuning the Vit-H-14 model, and to reduce or penalize the complexity of the model weights which can reduce the overfitting, we have used a **weight decay** (Regularization Parameter L2) of λ = **1e-4**.

$$\mathcal{L}_{\text{total}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}_{\text{MSE Loss}} + \underbrace{\lambda\sum_{j}\|w_j\|_2^2}_{\text{Weight Decay (L2)}} \tag{4.1}$$

- y_i: Ground truth BMI.

- $\hat{y}_i$: Predicted BMI.

- N: Number of samples.

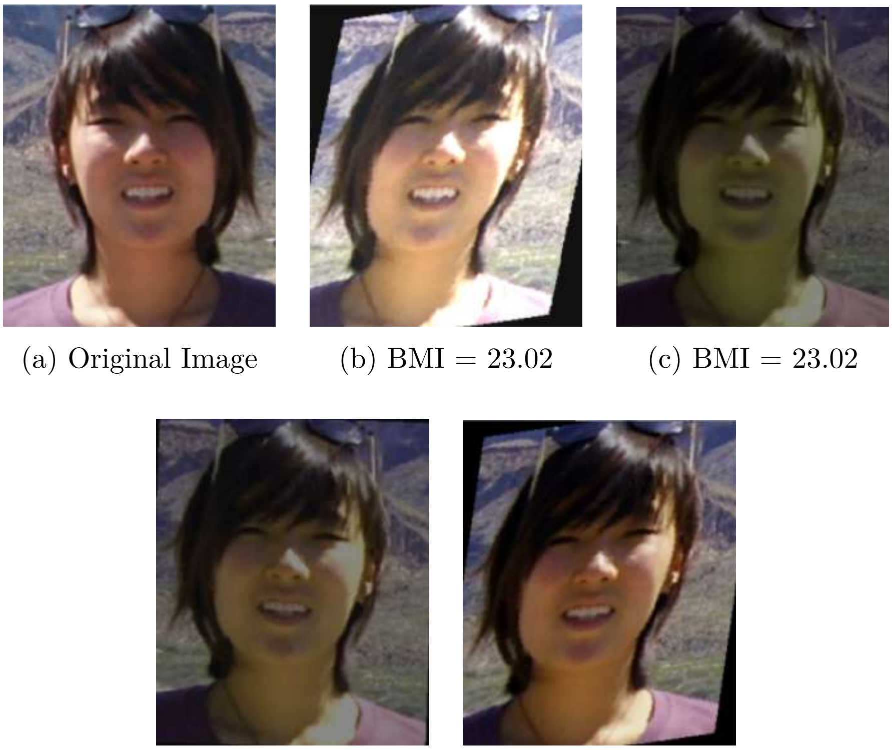
- w_j: Trainable weight parameters (e.g., in your regression head).

- $\lambda = 0.0001$: Weight decay factor.

4.3.5 Data augmentation

- The dataset is splitted on 03 different sets (70% , 10% , 20%) for training, validation, testing respectively.

- In fact, the training dataset is small (2944 images), because of that we have applied a Data augmentation technique for each training image (1x5) so we will have (2944x5=14720 training images) like it's shown here in this Figure 4.13:



(a) Original Image (b) BMI = 23.02 (c) BMI = 23.02

(d) BMI = 23.02 (e) BMI = 23.02

Figure 4.13: Data augmentation images.

- More precisely, in order to improve the robustness and generalization ability of the BMI estimation model, a variety of **data augmentation techniques** have been applied for the training phase. These transformations are designed to artificially increase the diversity of the training dataset by applying **random but controlled modifications** to the input images. We have applied :

1/- Random rotation: Rotates the image randomly within a range of ±5 degrees. It Introduces slight angular variations to simulate different head tilts and orientations. This helps the model become invariant to minor rotations, improving robustness. This is shown by the following Figure 4.14 :

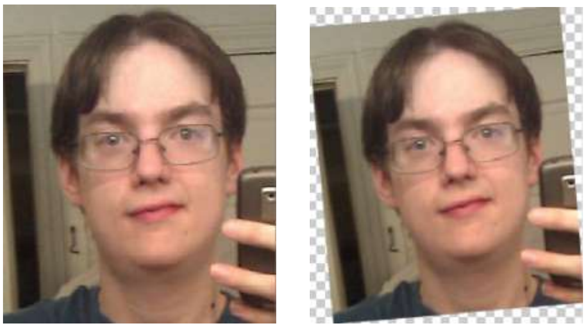


Figure 4.14: Original image , rotated image by 5 degree.

2/- Random horizontal flip: The image is flipped horizontally (mirrored). It mimics left-right symmetry, allowing the model to learn features regardless of face orientation. This can prevent the model from overfitting to a specific face orientation. This is shown by the following Figure 4.15 :

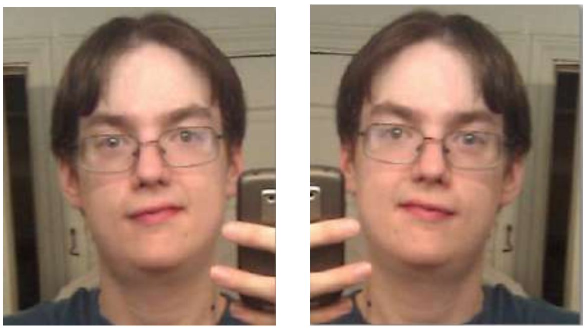


Figure 4.15: Original image , horizontal flipped image.

3/- Random brightness adjustment using Lambda: Brightness refers to the overall darkness or lightness of the image Brightness is adjusted using a random factor between **0.5 and 1.5** (50% darker to 50% brighter), to increase variation in image illumination. Preventing the model from relying too heavily on specific brightness levels. This is shown by the following Figure 4.16 :

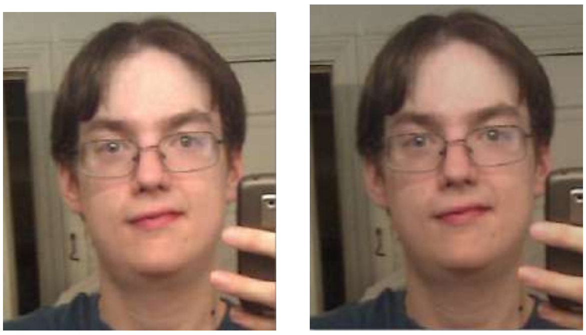


Figure 4.16: Orignal image, image with 0.8 brightness factor.

4/- Random adjust contrast (Custom class): Contrast is the difference in brightness between objects or regions (the difference between maximum and minimum pixel intensity in the image). This technique randomly changes how much light/dark separation is in an image (the image contrast is adjusted randomly between 80% and 120% of the original) and helps the model become less sensitive to contrast changes in real-world images. This is shown by the following Figure 4.17 :

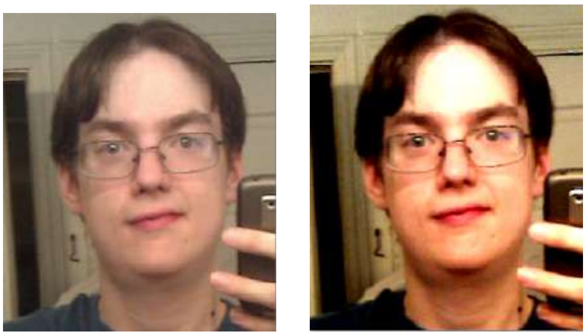


Figure 4.17: Original image , image with 96.34% contrast.

- Input images and labels are transferred to the appropriate device CUDA, to use the GPU of the PC, reducing time consuming of the training process, and increasing the memory capacity.

4.3.6 Experimental results, interpretations

After training models for 30 epochs, with the Hyper-Parameters montioned before, we get the results represented by the Table below :

Table 4.1: Obtained results on **VisualBMI**: Existing work (VGGFace) vs. Our model (with different Hyper-Parameters).

Works	**Existing Work** [20]	**Our Work**		
Models	**VGGFace**	**Vit_H_14**	**Vit_H_14**	**Vit_H_14**
MTCNN for Preprocessing	✓	✓	✓	✓
Data Augmentation	✗	✗	✓	✓
Fine-Tuning	✗	✗	✗	✓
BMI RMSE	6.94	6.82	6.27	6.14
BMI MAE	**4.99**	**4.85**	**4.36**	**4.18**

- Interpretations:

- **Baseline comparison – VGGFace vs. ViT_H_14:**
  The ViT_H_14 model outperforms the VGGFace model in both RMSE (6.82 vs. 6.94) and MAE (4.85 vs. 4.99). This demonstrates that Vision Transformers are more effective than traditional CNN-based models (like VGGFace) for BMI estimation from facial images.

- **Effect of data augmentation:**
  Applying data augmentation to ViT_H_14 improves the results significantly: RMSE drops from 6.82 to **6.27**, and MAE from 4.85 to **4.36**. This indicates that data augmentation helps the model generalize better by increasing the diversity of the training data.

- **Effect of fine-tuning:**
  Adding fine-tuning on top of data augmentation further enhances the model's performance. The best results are achieved with an RMSE of **6.14** and MAE of **4.18**. This confirms that fine-tuning the pretrained ViT_H_14 on the specific dataset leads to more accurate BMI predictions.

- **Progressive improvement through customization:**
  The results show a clear step-by-step improvement:

  $$\text{ViT\_H\_14 (Baseline)} \rightarrow + \text{ Data augmentation} \rightarrow + \text{ Fine-tuning}$$

  Corresponding MAE values: 4.85 → 4.36 → 4.18. This progression highlights the importance of tailoring the model to the task.

- **Conclusion:**
  Our proposed method (highlighted in red) shows that customized training strategies (such as augmentation and fine-tuning) yield better performance than simply switching to a more advanced model. Model adaptation to the dataset proves crucial for achieving optimal BMI estimation accuracy.

4.4 Experiment 02 : VGG16, ResNet50 + Arrest Records

4.4.1 Model description

The system first processes input images from the Arrest Records dataset through MTCNN for face detection and alignment. The detected and aligned faces are then fed into a VGG16, ResNet50 models for deep feature extraction, capturing high-level facial characteristics. Finally, a regression layers maps these extracted features to BMI predictions (Mutli-Task Learning will be explained later), enabling the estimation of body mass index directly from facial appearance, This is shown in the Following Figure 4.18:

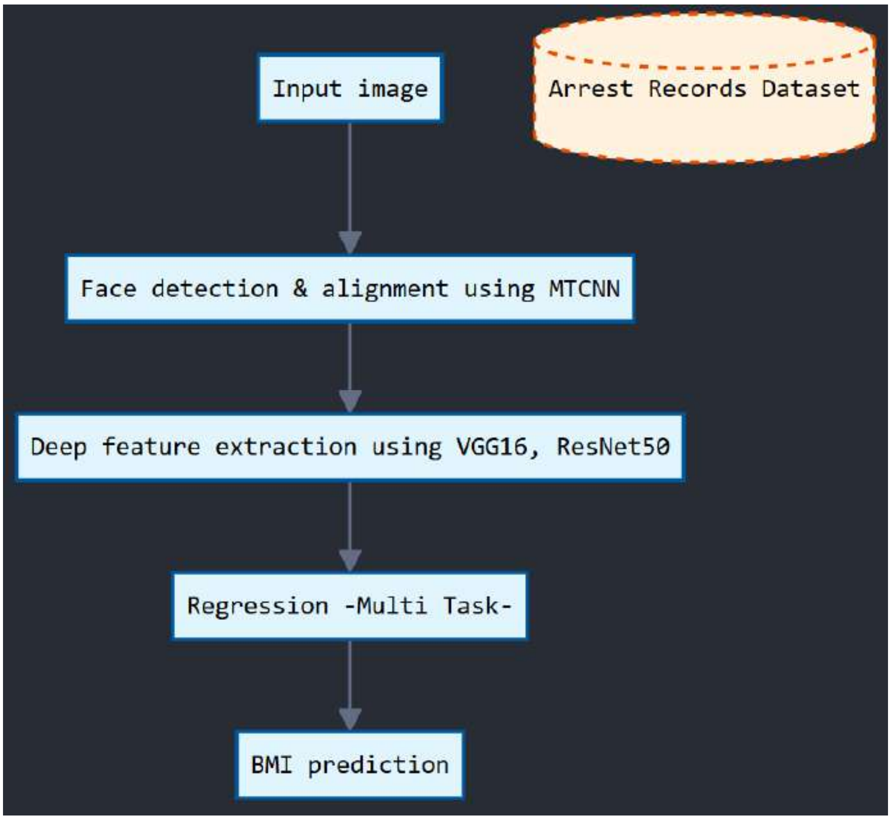
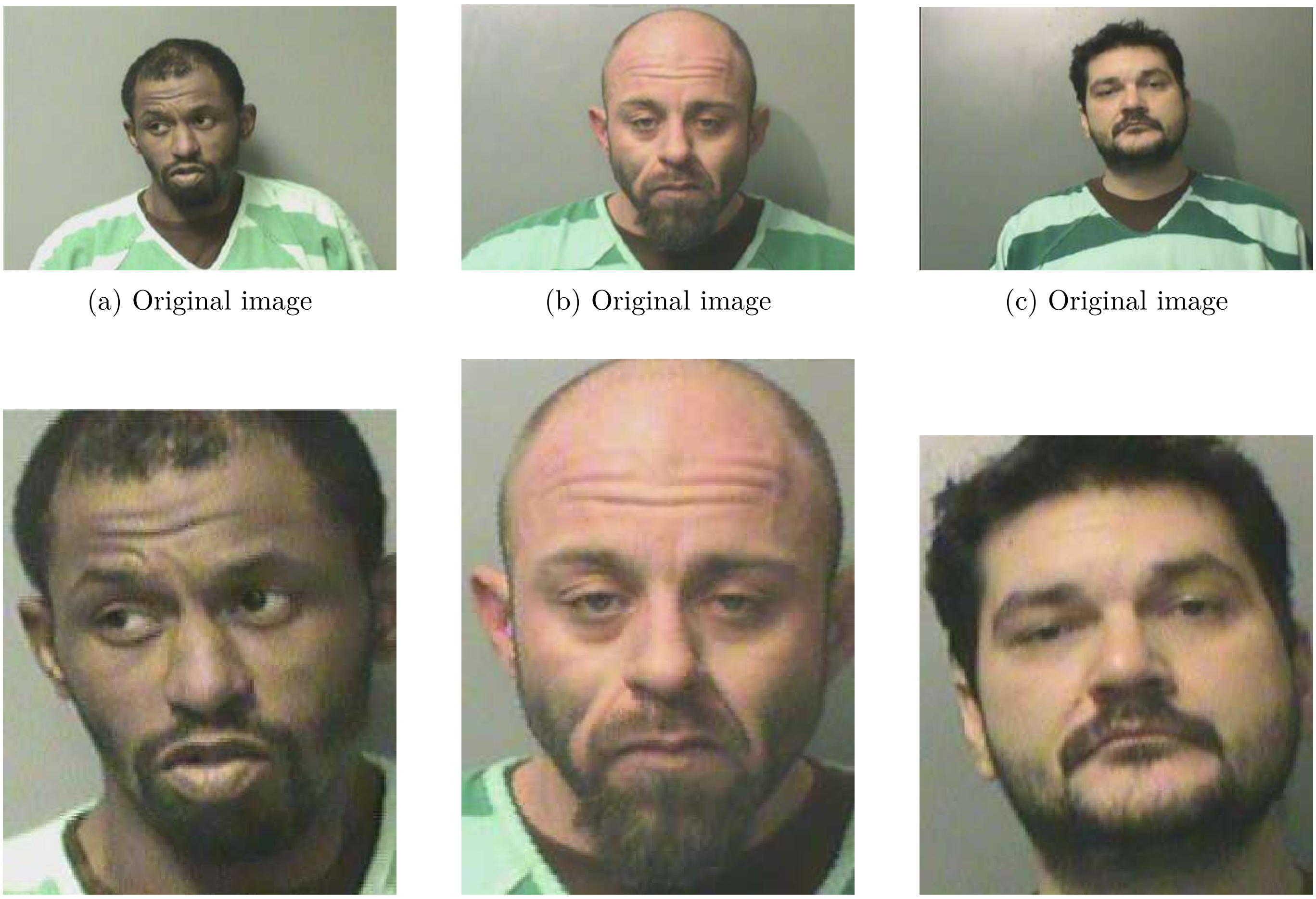

Input image
Arrest Records Dataset
Face detection & alignment using MTCNN
Deep feature extraction using VGG16, ResNet50
Regression -Multi Task-
BMI prediction

Figure 4.18: Model description.

4.4.2 Data preprocessing

- Converting height to meters and weight to kilograms.
- Calculating BMI: BMI $= \frac{\text{weight (kg)}}{\text{height (m)}^2}$.
- Encoding gender numerically.

- Splitting the dataset into training (1227 images) and testing (317 images).
- Using the **MTCNN face detection model**, to detect and cropp only the region of interest (face region), from the whole image.



(a) Original image

(b) Original image

(c) Original image

(d) Preprocessed with MTCNN

(e) Preprocessed with MTCNN

(f) Preprocessed with MTCNN

Figure 4.19: Original images / Preprocessed images with MTCNN

4.4.3 Model architecture

- Our method for predicting Body Mass Index (BMI) from Visual BMI dataset images utilizes VGG16 and ResNet50 pre-trained models as a feature extractor models via Transfer Learning, imported from Keras python library, extracting a feature vector of **size = 25088**, and a feature vector of **size = 100352**.

1/- ResNet50 model architecture:

- The Figure 4.20 represent a general or a basic overview on the ResNet50 pre-trained model architecture :

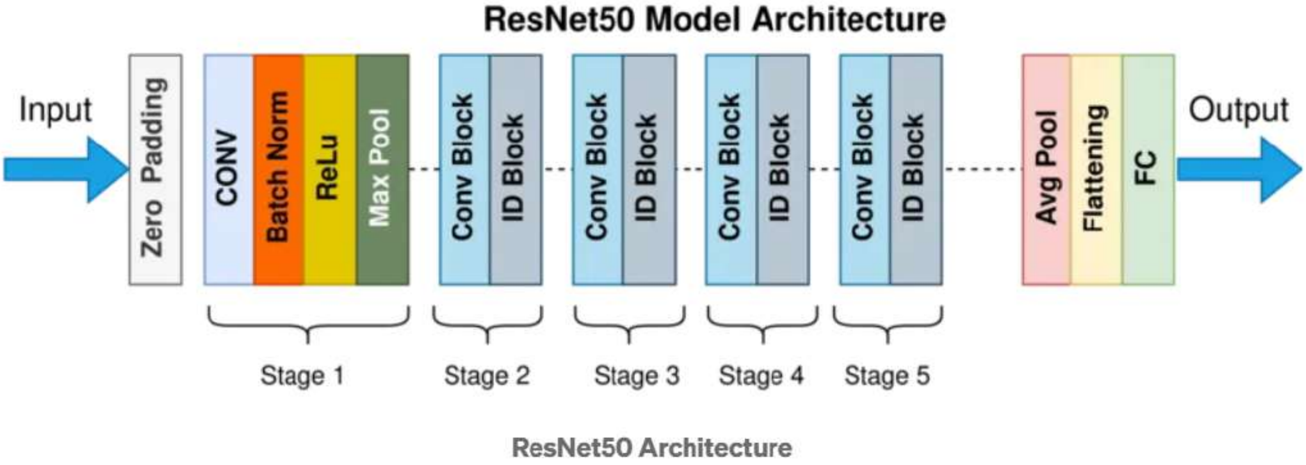


Figure 4.20: ResNet50 architecture [31].

Resnet50 is a deep convolutional neural network (CNN) architecture that was developed by Microsoft Research in 2015. It is a variant of the popular ResNet architecture, which stands for "Residual Network." The "50" in the name refers to the number of layers in the network, which is 50 layers deep [31].

Resnet50 is a powerful image classification model that can be trained on large datasets and achieve state-of-the-art results. One of its key innovations is the use of residual connections, which allow the network to learn a set of residual functions that map the input to the desired output. These residual connections enable the network to learn much deeper architectures than was previously possible, without suffering from the problem of vanishing gradients [31].

The architecture of ResNet50 is divided into four main parts: the convolutional layers, the identity block, the convolutional block, and the fully connected layers. The convolutional layers are responsible for extracting features from the input image, while the identity block and convolutional block are responsible for processing and transforming these features. Finally, the fully connected layers are used to make the final classification [31].

Resnet50 begins with convolutional layers, each followed by batch normalization and ReLU activation to extract features like edges and textures. Max pooling layers then reduce spatial dimensions while preserving important information. The identity block applies convolutions and adds the input back to the output to learn residual mappings. The convolutional block includes a 1×1 convolution to adjust filter dimensions before applying 3×3 convolutions. These residual connections help with deeper training and prevent vanishing gradients. Finally, fully connected layers make predictions, ending in a softmax layer for class probabilities [31].

- How it solved the problem of vanishing gradients

Skip connections, or residual connections as is represented in the Figure 4.21, are essential in **Resnet50** for enabling deep network training without vanishing gradients. Vanishing gradients occur when deeper layers receive very small updates, hindering learning. Skip connections help by allowing data to bypass certain layers, preserving gradient flow.

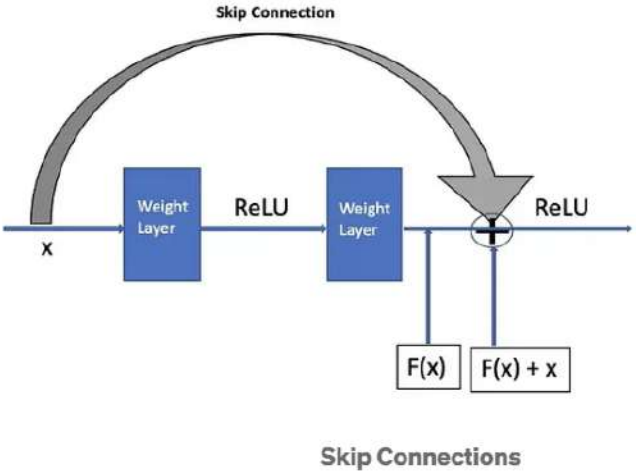


Figure 4.21: The concept of the widely popular CNN [31].

They let the network learn residual functions—differences between input and output—instead of full mappings. In **Resnet50**, identity blocks pass input unchanged alongside convolutions and then add it back to the output. Convolutional blocks use a 1×1 convolution to align dimensions before performing this addition. This design allows for effective training of deeper architectures by stabilizing learning [31].

- Summary

In summary, ResNet50 is a cutting-edge deep convolutional neural network architecture that was developed by Microsoft Research in 2015. It is a variant of the popular ResNet architecture and comprises of 50 layers that enable it to learn much deeper architectures than previously possible without encountering the problem of vanishing gradients. The architecture of ResNet50 is divided into four main parts: the convolutional layers, the identity block, the convolutional block, and the fully connected layers. The convolutional layers are responsible for extracting features from the input image, the identity block and convolutional block process and transform these features, and the fully connected layers make the final classification. ResNet50 has been trained on the large ImageNet dataset, achieving an error rate on par with human performance, making it a powerful model for various image classification tasks such as object detection, facial recognition and medical image analysis. Additionally, it has also been used as a feature extractor for other tasks, such as object detection and semantic segmentation.

2/- VGG16 model architecture:

- The VGG-16 model shown in the Figure 4.22 is a convolutional neural network (CNN) architecture that was proposed by the Visual Geometry Group (VGG) at the University of Oxford. It is characterized by its depth, consisting of 16 layers, including 13 convolutional layers and 3 fully connected layers. VGG-16 is renowned for its simplicity and effectiveness, as well as its ability to achieve strong performance on various computer vision tasks, including image classification and object recognition. The model's architecture features a stack of convolutional layers followed by max-pooling layers, with progressively increasing depth. This design enables the model to learn intricate hierarchical representations of visual features, leading to robust and accurate predictions. Despite its simplicity compared to more recent architectures, VGG-16 remains a popular choice for many deep learning applications due to its versatility and excellent performance [32].

- The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an annual competition in computer vision where teams tackle tasks including object localization and image classification. VGG16, proposed by Karen Simonyan and Andrew Zisserman in 2014, achieved top ranks in both tasks, detecting objects from 200 classes and classifying images into 1000 categories [32].

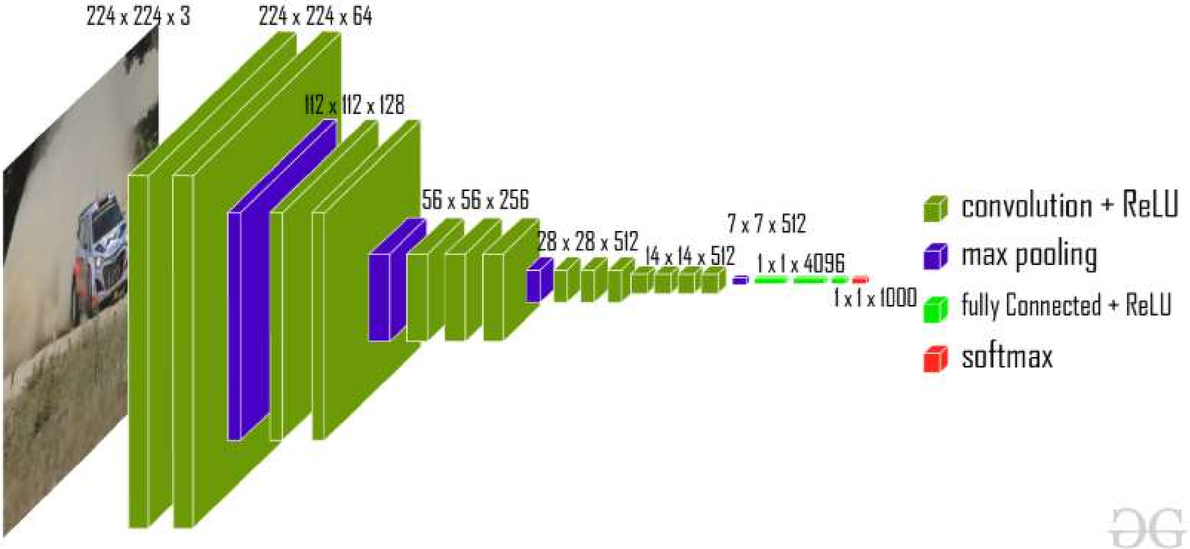


Figure 4.22: VGG-16 architecture. [32]

- This model achieves 92.7% top-5 test accuracy on the ImageNet dataset which contains 14 million images belonging to 1000 classes [32].

- The ImageNet dataset contains images of fixed size of 224 × 224 and have RGB channels. So, we have a tensor of $(224, 224, 3)$ as our input. This model process the input image and outputs the a vector of 1000 values: [32]

$$\hat{y} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \vdots \\ \vdots \\ \hat{y}_{999} \end{bmatrix}$$

- This vector represents the **classification probability** for the corresponding class, and to make sure these probabilities add to 1, a softmax function is used [32].

- Softmax activation function:

The Figure 4.23 represents the Softmax activation function curve.

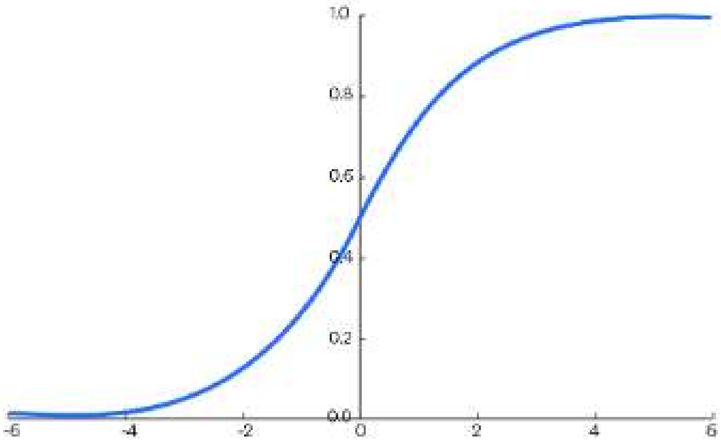


Figure 4.23: Softmax curve [33].

- It gives the probability distribution of multiple classes, making the decision-making process straightforward and effective [33].

- By converting raw scores to probabilities, it not only provides a value to be worked with but also brings clarity to interpreting results [33].

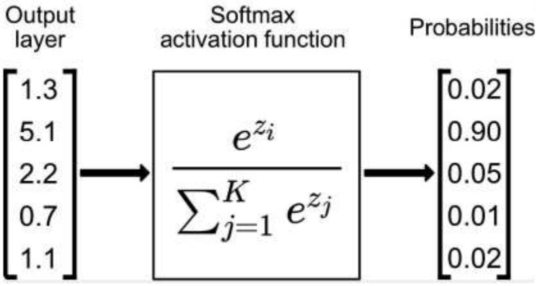

Output layer
Softmax activation function
Probabilities
1.3
5.1
2.2
0.7
1.1
$\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$
0.02
0.90
0.05
0.01
0.02

Figure 4.24: Example [33].

The softmax function converts a vector of real numbers into a probability distribution:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4.2}$$

where z_i is the i-th element of the input vector and K is the number of classes.

Example: Given input vector $\mathbf{z} = [1.3, 5.1, 2.2, 0.7, 1.1]$:

$$\text{softmax}(z_1) = \frac{e^{1.3}}{e^{1.3} + e^{5.1} + e^{2.2} + e^{0.7} + e^{1.1}} = \frac{3.67}{185.39} = 0.02 \tag{4.3}$$

$$\text{softmax}(z_2) = \frac{e^{5.1}}{185.39} = \frac{164.02}{185.39} = 0.90 \tag{4.4}$$

$$\text{softmax}(z_3) = \frac{e^{2.2}}{185.39} = \frac{9.03}{185.39} = 0.05 \tag{4.5}$$

$$\text{softmax}(z_4) = \frac{e^{0.7}}{185.39} = \frac{2.01}{185.39} = 0.01 \tag{4.6}$$

$$\text{softmax}(z_5) = \frac{e^{1.1}}{185.39} = \frac{3.00}{185.39} = 0.02 \tag{4.7}$$

The output probabilities sum to 1: $0.02 + 0.90 + 0.05 + 0.01 + 0.02 = 1.00$

- The Softmax Function comes into its own when dealing with multi-class classification tasks in machine learning. In these scenarios, we need our model to predict one out of several possible outcomes [33].

- VGG16 architecture:

- The VGG-16 architecture is a deep convolutional neural network (CNN) designed for image classification tasks. It was introduced by the Visual Geometry Group at the University of Oxford. VGG-16 is characterized by its simplicity and uniform architecture, making it easy to understand and implement [32].

- The VGG-16 configuration typically consists of 16 layers as its shown in the Figure 4.25, including 13 convolutional layers and 3 fully connected layers. These layers are organized into blocks, with each block containing multiple convolutional layers followed by a max-pooling layer for downsampling [32].

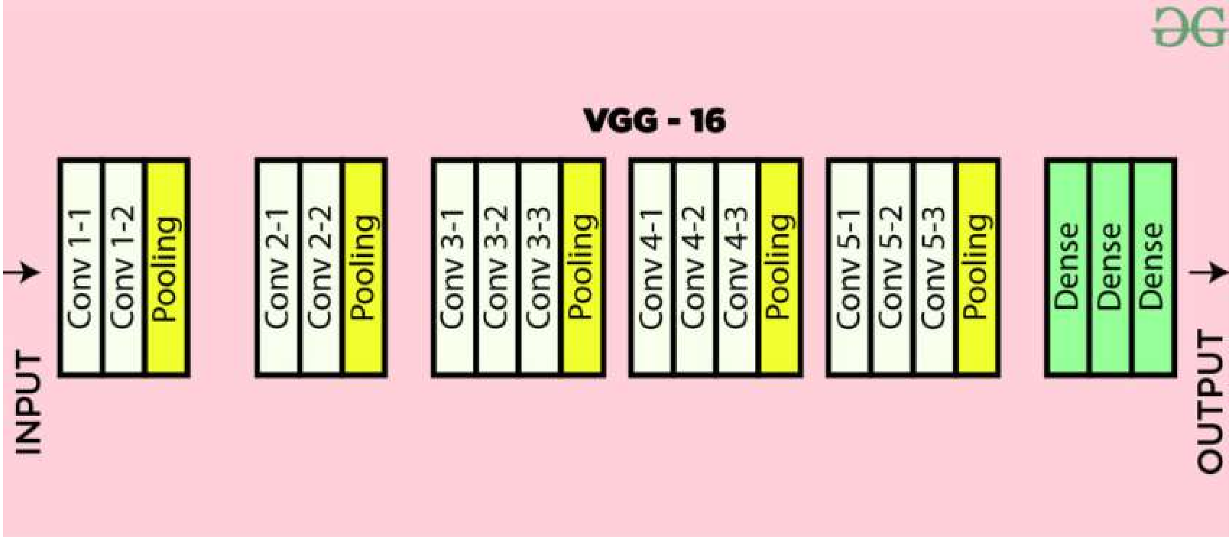


Figure 4.25: VGG-16 architecture map [32].

The following describes the complete architecture based on the provided specifications:

Input dimensions: $(224, 224, 3)$.

- Convolutional block 1

- Two consecutive convolutional layers with 64 filters each.
- Filter size: 3×3.
- Padding: Same padding (to maintain spatial dimensions).
- Activation: ReLU.
- Max Pooling Layer: Pool size 2×2, stride $= 2$.

- Convolutional block 2

- Two consecutive convolutional layers with 128 filters each.
- Filter size: 3×3.
- Padding: Same padding.
- Activation: ReLU.
- Max pooling layer: Pool size 2×2, stride $= 2$.

- Convolutional block 3

- Three consecutive convolutional layers with 256 filters. each
- Filter size: 3 × 3.
- Padding: Same padding.
- Activation: ReLU.
- Max pooling layer: Pool size 2 × 2, stride = 2.

- Convolutional block 4

- Three consecutive convolutional layers with 512 filters each.
- Filter size: 3 × 3.
- Padding: Same padding.
- Activation: ReLU.
- Max pooling layer: Pool size 2 × 2, stride = 2.

- Convolutional block 5

- Three consecutive convolutional layers with 512 filters each.
- Filter size: 3 × 3.
- Padding: Same padding.
- Activation: ReLU.
- Max pooling layer: Pool size 2 × 2, stride = 2.

- Feature map flattening Flatten the output feature map (7 × 7 × 512) into a vector of size 25088.

- Fully connected layers

0 **First FC layer:**

- Input size: 25088.
- Output size: 4096.
- Activation: ReLU.

0 **Second FC layer:**

- Input size: 4096.
- Output size: 4096.
- Activation: ReLU.

0 **Third FC layer (Output layer):**

- Input size: 4096.
- Output size: 1000 (corresponding to 1000 classes in ILSVRC challenge)
- Activation: Softmax (for classification probabilities).

– This architecture follows the specifications provided, including the use of ReLU activation function and the final fully connected layer outputting probabilities for **1000 classes** using softmax activation.

- To adapt these models for the regression task of BMI prediction, we have removed the original classification head of both pre-trained models (which is represented as FC in the figure of ResNet50 model architecture for example), and we used Multi-Task Learning technique (replacing the classification head of the pre-trained models with these new Heads) for simultaneous prediction of:

- BMI (regression), - Age (regression), - Gender (classification).

Each task was handled by dedicated fully connected layers, that's mean we have used **02 custom regression heads** (ends with Relu activation), and **01 custom classification head** (ends with sigmoid activation), consisting of:

03 layers with decreasing dimensionalities: extracted features size → 256 → 128 → 1. These are represented in the Figure 4.26 and 4.27 :

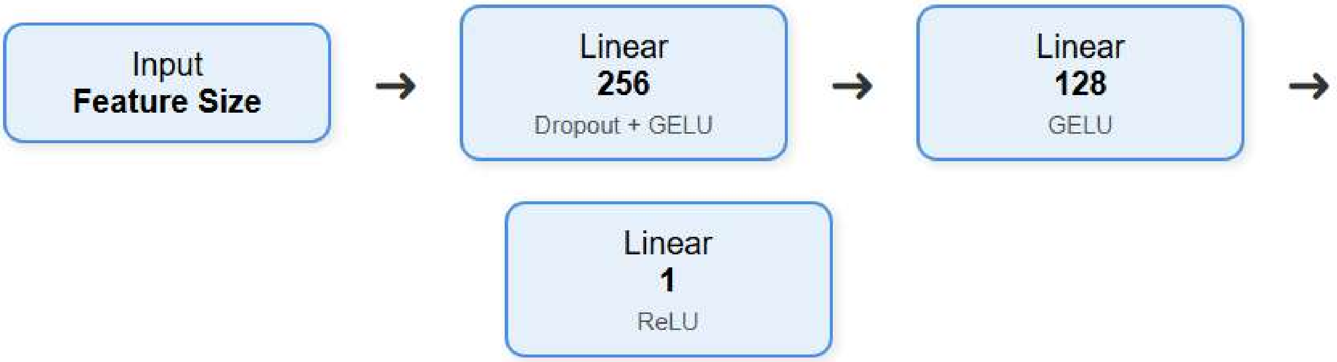


Figure 4.26: Regression head.

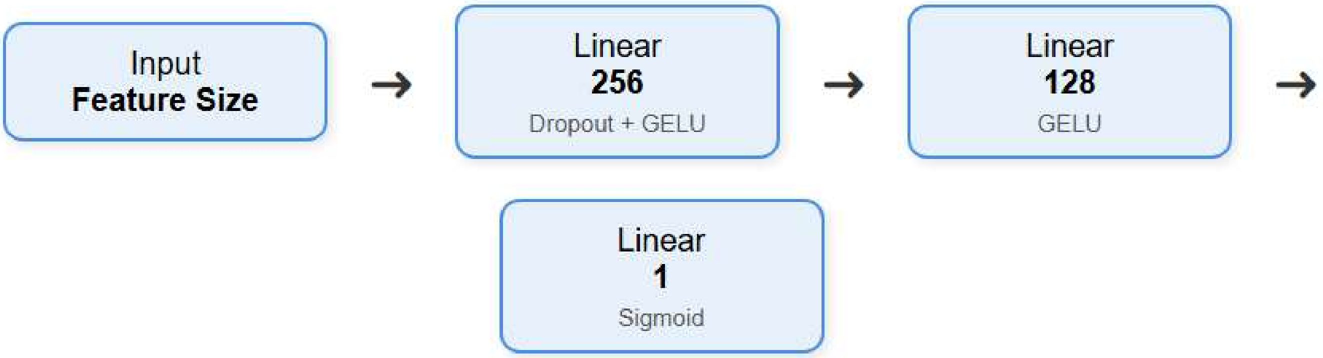


Figure 4.27: Classification head.

- **GELU** activation function after each layer, except the last one **RELU** activation function for the regression head of BMI and AGE, because both BMI and AGE have a positive range values, and **Sigmoid** activation function for the classification head

- To prevent overfitting, we added one dropout layer, with a dropout of 50% into our model architecture, which applied only after in first linear layer. It's usually more effective to regularize at the beginning of deep networks (the largest part of the network, where overfitting is most likely) than in later narrow layers (where underfitting is most likely).

- The final output is a single scalar value representing the predicted BMI, AGE, SEX.

- We have leveraged the pretrained knowledge, all parameters (26 M parameters for ResNet50 / 138 M parameters for VGG16) are one time frozen and other time are trainable during training.

4.4.4 Training and optimization

This model is designed to predict three different outputs:

- **bmi**: a regression output, optimized using the Mean Squared Error (MSE) loss.
- **age**: another regression output, optimized using the Mean Absolute Error (MAE) loss.
- **sex**: a binary classification output, optimized using the Binary Crossentropy loss.

The total loss function $\mathcal{L}_{\text{total}}$ is given by:

$$\mathcal{L}_{\text{total}} = w_{\text{bmi}} \cdot \mathcal{L}_{\text{bmi}} + w_{\text{age}} \cdot \mathcal{L}_{\text{age}} + w_{\text{sex}} \cdot \mathcal{L}_{\text{sex}} + + \underbrace{\lambda \sum_j \|w_j\|_2^2}_{\text{Weight Decay (L2)}} \tag{4.8}$$

Where:

- $(w_{\text{bmi}}, w_{\text{age}}, w_{\text{sex}})$ are the weights assigned to each task's loss (0.8,0.1,0.1).

- $\mathcal{L}_{\text{bmi}}$ is the MSE loss between predicted and true BMI values.

$$\mathcal{L}_{\text{bmi}} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^{\text{bmi}} - \hat{y}_i^{\text{bmi}} \right)^2 \tag{4.9}$$

where:

  - ◦ y_i^{bmi} is the true BMI value for the i-th sample.
  - ◦ $\hat{y}_i^{\text{bmi}}$ is the predicted BMI value.
  - ◦ N is the total number of samples.

- $\mathcal{L}_{\text{age}}$ is the MAE loss between predicted and true AGE values.

$$\mathcal{L}_{\text{age}} = \frac{1}{N} \sum_{i=1}^{N} |y_i^{\text{age}} - \hat{y}_i^{\text{age}}| \tag{4.10}$$

where:

- ○ y_i^{age} is the true age value.
- ○ $\hat{y}_i^{\text{age}}$ is the predicted age.

- $\mathcal{L}_{\text{sex}}$ is the binary crossentropy loss for the SEX classification task.

$$\mathcal{L}_{\text{sex}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i^{\text{sex}}\log\left(\hat{y}_i^{\text{sex}}\right) + \left(1 - y_i^{\text{sex}}\right)\log\left(1 - \hat{y}_i^{\text{sex}}\right)\right] \tag{4.11}$$

where:

- ○ $y_i^{\text{sex}} \in \{0, 1\}$ is the true label (e.g., 0 for female, 1 for male).
- ○ $\hat{y}_i^{\text{sex}}$ is the predicted probability that the input corresponds to class 1 (male).

.

- Adjusting these weights (0.8 for BMI / 0.1 for AGE / 0.1 for SEX) allows us to prioritize one task over the others during training. In our case, setting a higher value for w_{bmi} emphasizes minimizing the BMI prediction error more than the others

- The models are trained for **60 epochs**.

- We have fixed the **batch size = 32**.

- The optimization is handled using the **Adam optimizer** with a learning rate of **1e-4** and **1e-5** in case of fine-tuning, and to reduce or penalize the complexity of the model weights which can reduce the overfitting, we have used a **weight decay** (Regularization Parameter L2) of λ = **1e-4**.

- Input images and labels are transferred to the appropriate device CPU of the PC.

4.4.5 Experimental results, interpretations

After training models for 60 epochs, with the Hyper-Parameters montioned before, we get the results represented by the Table 4.2 :

- Interpretations:

- **Effect of the pretrained model:** ResNet50 outperformed the VGG 16 model even if the VGG16 model has 138 M parameters and the the ResNet50 has 25 M parameters (5.52 times lesser), gaining in both time factor and the accuracy.

Table 4.2: Obtained Results on **Arrest Records**: Existing work (VGGFace) vs. Our model (with different Hyper-Parameters).

Works	**Previous work** [7]		**Our work**			
Models	**VGG16**	**ResNet50**	**VGG16**	**ResNet50**	**VGG16**	**ResNet50**
MTCNN Preprocessing	✓	✓	✓	✓	✓	✓
Fine-Tuning	✗	✗	✗	✗	✓	✓
BMI MAE	**6.13**	**5.02**	**3.77**	**3.56**	**4.18**	**4.04**

That may be due to: It introduces residual connections (skip connections), which solve the problem of vanishing gradients and allow training much deeper models effectively.

- **Effect of fully connected layers:**
Our custom deep FC layers very likely contributed to the improved preformances:
- A more deeper FC can give the model more capacity to learn complex relationships.
- Providing better adaptation to regression-specific tasks, unlike (may be) the simpler FC structures used in previous work.

- **Regularization Hyper-Parameter:**
In the previous work, they did not montioned the usage of this hyper-parameter of L2 Regularization, we have used this hyper-parameter preventing the model from learning very large weight values, which are often a sign of overfitting.

- **Effect of fine-tuning:**
Our work included fine-tuning the models, which was not applied in the previous work, but not likely contributing more to the performance gains, because the fine-tuning in this case leads the model to lose performance comparing with not applying the fine-tuning. This confirms that this dataset is relatively small for fine-tuning a deep learning model like VGG16. Fine-tuning a large model requires a significant amount of labeled data to avoid overfitting. Since the dataset is small, fine-tuning the entire VGG16 network might lead to overfit-

ting.
- Why ? Fine-tuning the entire pre-trained model is risky with a small dataset like 1544 images because the model can forget the useful knowledge it learned before (that can lead to a catastrophic forgetting problem, specially when the model have a very big size).

The following Table 4.3 represent BMI MAE results when using MSE loss function and when using the MAE loss function :

Table 4.3: Obtained results with **MSE vs. MAE** loss functions.

Models	**VGG16**	**VGG16**
MTCNN Preprocessing	✓	✓
Fine-Tuning	✗	✗
Loss Function	MSE	MAE
BMI MAE	**3.77**	**3.99**

- Training with **MSE loss function** gives better results, possibly because MSE penalizes larger errors more, encouraging more accurate predictions in outlier cases, That's our reason to choose the MSE loss function.

4.5 Experiment 03 : Inception-V3 + Illinois DOC Faces

4.5.1 Model description

The system first processes input images from the Illinois DOC Faces dataset through MTCNN (Multi-task Convolutional Neural Network) for face detection and alignment. The detected and aligned faces are then fed into Inception-v3 model for deep feature extraction, capturing high-level facial characteristics. Finally, a regression layers maps these extracted features to BMI predictions, enabling the estimation of body mass index directly from facial appearance, This is shown in the Following Figure:

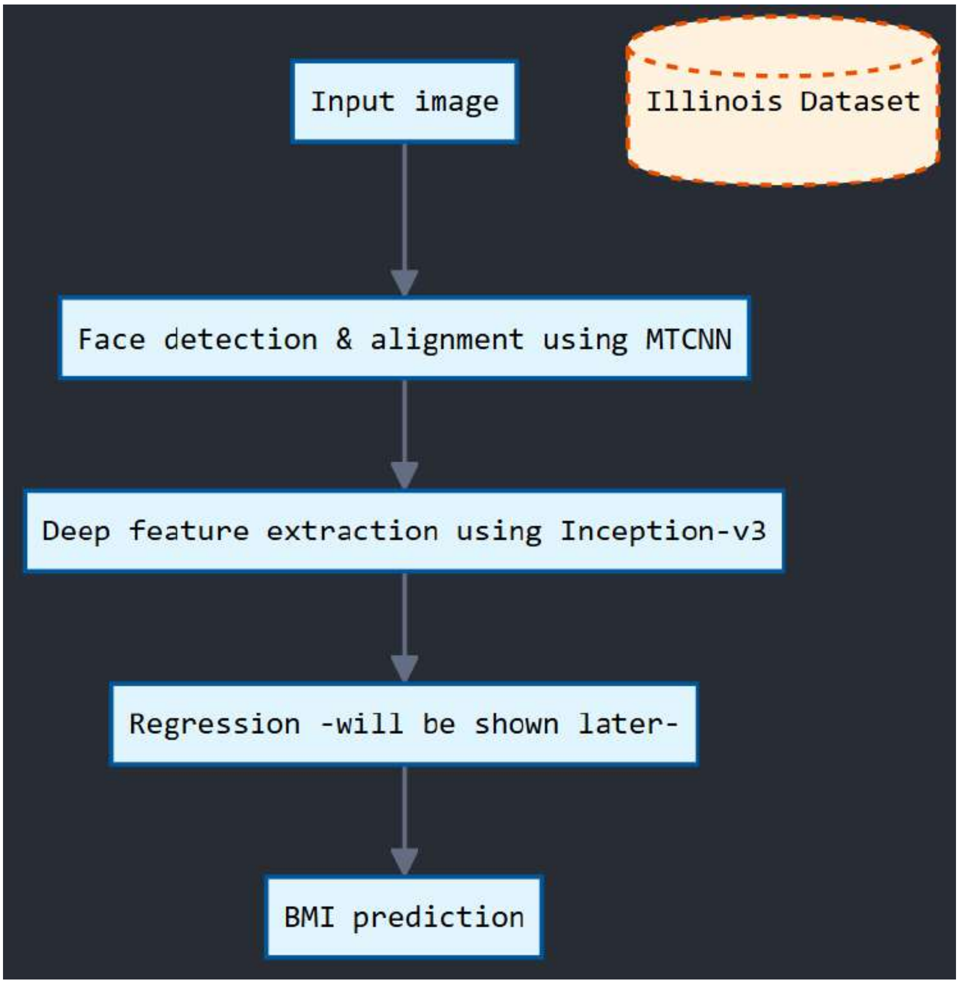

Input image
Illinois Dataset
Face detection & alignment using MTCNN
Deep feature extraction using Inception-v3
Regression -will be shown later-
BMI prediction

Figure 4.28: Model description.

4.5.2 Data preprocessing

- Cleaning malformed or missing entries.
- Converting height to meters and weight to kilograms.
- Calculating BMI: BMI $= \frac{\text{weight (kg)}}{\text{height (m)}^2}$.
- Dataset splitted (70% , 10% , 20%) for (Training , Validation , Testing).
- Face detection using **MTCNN detector**.

4.5.3 Face detection and alignment method selection

A very recent study discovered that significant regions for a weight estimation model from face images such as the face-contour (forehead, cheeks and jaw) are usually excluded from face cropping algorithms since eyes contain the most meaningful information for face recognition tasks [2].

Indeed, the VIP attribute dataset is distributed in an already cropped version, narrow bounding box and excluding in most cases parts of the face contour (forehead, cheeks and jaw).

Those who did this study have contacted the authors of the **VIP attribute dataset** which provided us with the original version. Therefore, they evaluated whether different croppings, specially the ones considering larger face areas, will lead to a more accurate prediction as suggested by the explainability approaches.

In their experiment, they have trained and tested their network for 4 different face cropping methods: **Viola-Jones** [34], **Multi Cascade CNN (MTCNN)** [15], the **dlib python package** and a **customized cropping**.

They defined our face cropping by considering the highest, lowest, and furthest at the left and furthest at the right facial landmarks computed by the dlib landmark detector.

The results presented in the following Figure 4.29 represent the MAE for different cropping margins.

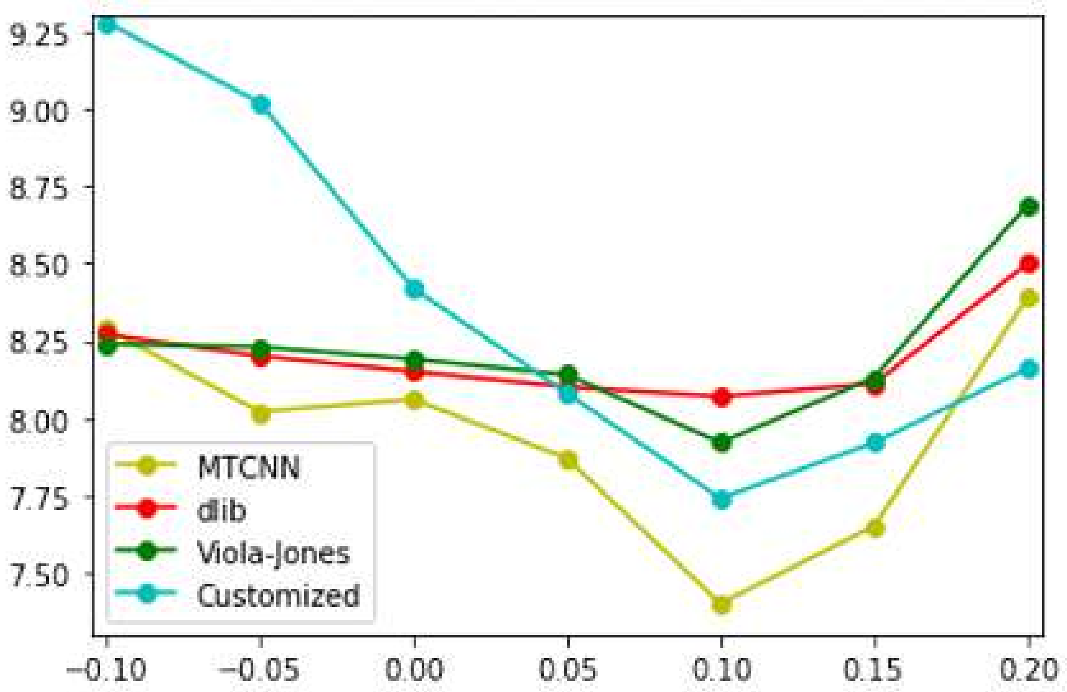


Figure 4.29: MAE in kg of the VIP attribute test set for various face detectors and cropping margins [2].

This Figure 4.1 highlights the benefit of an increased margin, specifically of 0.1 (10% increase of the original bounding box), specially for the rectangular face croppings (MTCNN) whose output is more adapted to the face shape. Nevertheless, large croppings include a high amount of hair and background regions increasing the network's MAE [2].

4.5.4 Model architecture

- Our method for predicting Body Mass Index (BMI) from Visual BMI dataset images utilizes the Inception-V3 model as a feature extractor from the torchvision library, extracting a feature vector of **size = 2048**. This backbone is chosen for its powerful performances as it's shown in the part of **state-of-the-art**, and because of it's small size = **103.9 MB**, with **27 M** parameters.

Inception-v3 model architecture:

Rethinking the Inception Architecture shown in the Figure 4.27, for Computer Vision has been one of the most complicated research papers I have read. The paper's language is very complicated along with its content. It was also difficult to find a detail breakdown of paper on the internet. So, the goal of this article is to breakdown **Inception-v3** research paper in great detail.

- The Figure 4.30 represent an overall architecture of the Inception-v3 model. :

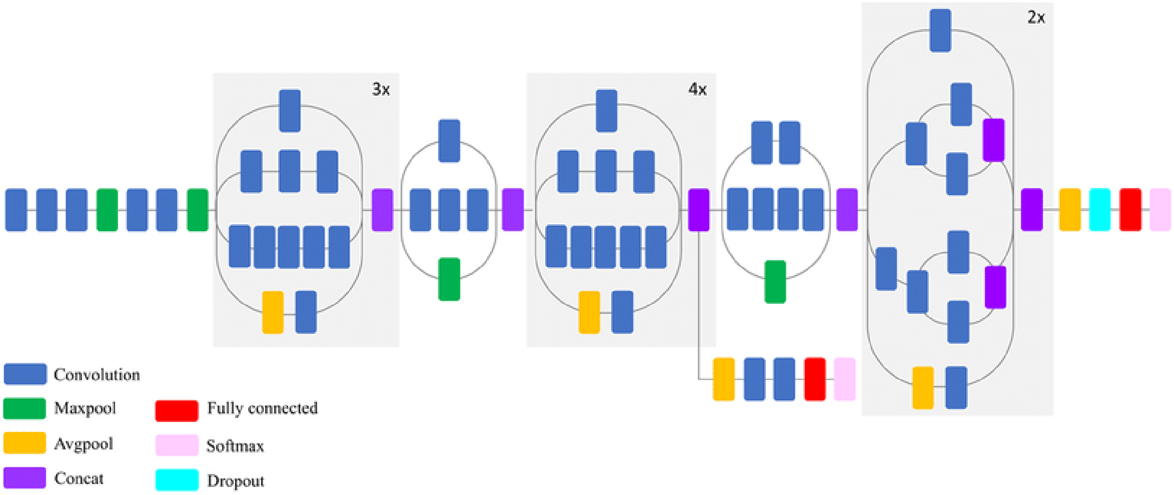


Figure 4.30: Inception-V3 model architecture [35].

1/- General design principals [35]:

- **Avoid representational bottlenecks**, especially in the early layers. A bottleneck occurs when there is a sudden, large reduction in input dimension, resulting in a high number of weights. Instead, dimensions should be reduced **gradually** to maintain representational capacity.
- **Higher-dimensional representations** are easier to process locally within a network. They help capture **complex and disentangled features**, whereas lower-dimensional data requires more entangled features to represent the same complexity.
- **Spatial aggregation** (e.g., with 3×3 convolutions) can be effectively performed over lower-dimensional embeddings **without significant loss in rep-**

representational power. This can be achieved by applying **1×1 convolutions** before spatial aggregation to reduce dimensions efficiently.

- **Balance width and depth** of the network. Optimal performance is achieved by tuning both the number of filters per stage (width) and the number of layers (depth). **Deeper networks** are needed to process wider layers effectively. This principle was applied in the design of **EfficientNets**.

2/- Factorizing convolutions with large filter size [35]:

The paper believes that original gains of GoogLeNet network arises from their generous use of dimensional reduction. Here dimensional reduction accounts for using 1x1 convolutions before 3x3 and 5x5 convolutions. This reduced number of parameters per stage and hence allowed to increase the depth of network. Similarly, we will see few techniques that can help to furthur reduce the dimensions [35].

Factorization into smaller convolutions:

Large filters like 7x7 and 5x5 are very expensive in computation. For example, a 5x5 convolution with n filters over a grid is 25/9 = 2.78 times more computationally expensive than a 3x3 convolution with the same number of filters. However, 5x5 filter can capture dependencies between signals between activations of units further away in the earlier layers which is very helpful for capturing spatial invariance. So, can we create 5x5 convolution using 3x3 convolution? Yes, convolving two 3x3 filters can produce same output size as convolving one 5x5 filter. This way, we end up with a net (9+9)/25 × reduction of computation, resulting in a relative gain of 28% by this factorization [35], This shown here in the Figure 4.31 :

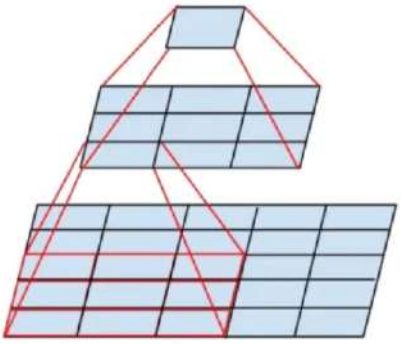


Figure 4.31: Mini-network replacing the 5 × 5 convolutions [35].

Still, this setup raises two general questions: Does this replacement result in any loss of expressiveness? If our main goal is to factorize the linear part of the computation, would it not suggest to keep linear activations in the first layer? According to second question, if we apply ReLU after 5x5 conv then we should apply ReLU after two 3x3 conv, keeping linear activation after first 3x3 conv. They tested both models and found out ReLU activation performs better than linear activation [35].

Spatial factorization into asymmetric convolutions:

- The Figure 4.32 represents the structure of a single Inception module—specifically, a simplified view of how parallel convolution operations are performed and their outputs concatenated :

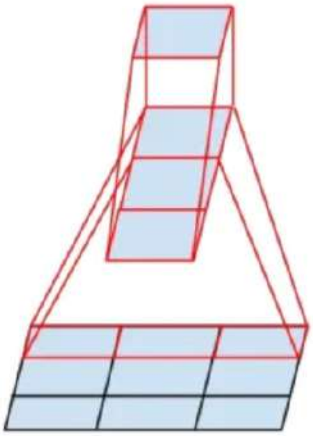


Figure 4.32: Mini-network replacing the 3 × 3 convolutions [35].

A major question arises that should we reduce 3x3 convolutions furthur to 2x2 convolutions? The paper states that reducing 3x3 convolutions into 2x2 convolutions results in 11% saving of computation while reducing 3x3 convolutions into 3x1 convolutions and 1x3 convolutions results in 33% saving of computation. They found that employing this factorization does not work well on early layers, but it gives very good results on medium grid-sizes (On m × m feature maps, where m ranges between 12 and 20). On that level, very good results can be achieved by using 1 × 7 convolutions followed by 7 × 1 convolutions [35].

- The Figure 4.33 represents the improvement made in Inception-v3 using Factorization into smaller convolutions—a technique that makes the model more efficient without sacrificing accuracy :

Efficient grid size reduction:

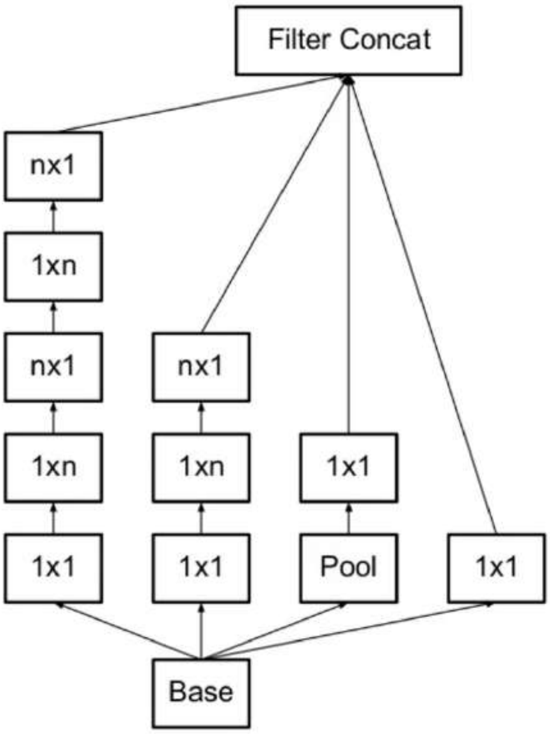


Figure 4.33: Inception modules after the factorization of the n × n convolutions [35].

To efficiently perform max pooling or average pooling, they present another variant that reduces the computational cost even further. We can use two parallel stride 2 blocks: P and C. P is a pooling layer (either average or maximum pooling) the activation, both of them are stride 2 the filter banks of which are concatenated [35], as it's shown in the Figure 4.34:

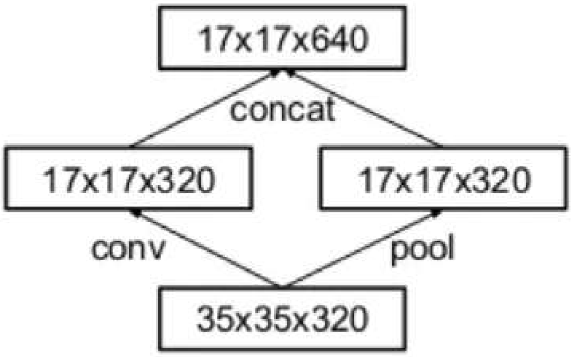


Figure 4.34: Inception module that reduces the grid-size while expands the filter banks [35].

- To adapt the Inception-V3 model for the regression task of BMI prediction, we replaced its classification head with a custom regression head, which is a deep fully connected layers consisting of:

04 layers with decreasing dimensionalities: 2048 → 1024 → 512 → 64 → 1.

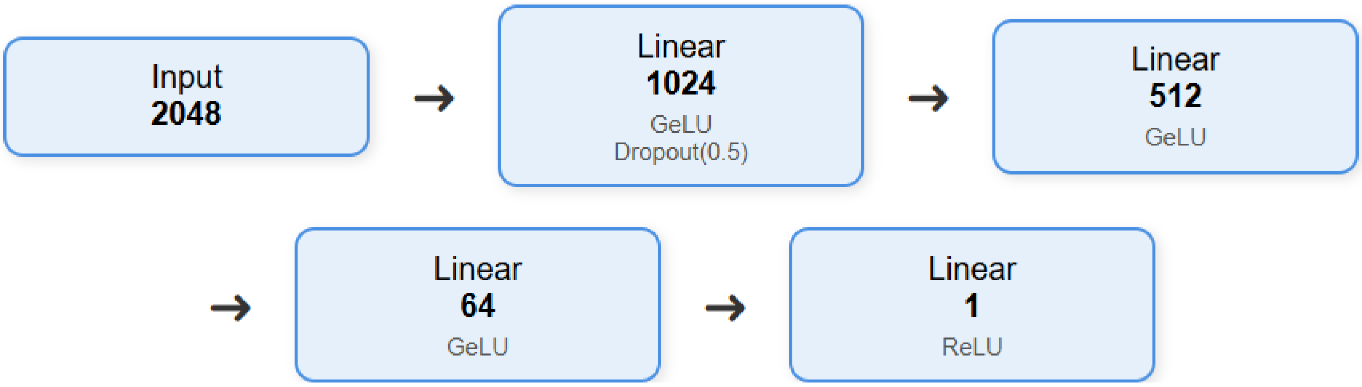


Figure 4.35: Regression model architecture

- **GELU** activation function after each layer, except the last one **RELU**, because the BMI has a positive range values.

- To prevent overfitting, we added one dropout layer, with a dropout of 50% into our model architecture, which applied only after in first linear layer. It's usually more effective to regularize at the beginning of deep networks (the largest part of the network, where overfitting is most likely) than in later narrow layers (where underfitting is most likely).

- The final output is a single scalar value representing the predicted BMI.

- We have leveraged the pretrained knowledge, all parameters (27 M parameters) of the Inception-v3 model are one time frozen and other time are trainable -Fine-Tuning- during training.

4.5.5 Data augmentation option

- The dataset is splitted on 03 different sets (70% , 10% , 20%) for training, validation, testing respectively.

- To improve the performances more, we have applied a Data augmentation technique for each training image (1x5) for each model (because there are 03 types of models and we will speak about that later), the applied Data Augmentation technique is exactly the same as the previous one (in our 1st method).

4.5.6 Gender-based model

- Because we have a comparatively large dataset than the previous ones that contain **56200 males** and **3649 females**, and because there are differences between the general face caracteristics between **man face image** and **female face image**, they have different bone mineral and muscle density, so their facial appearance differs even when they are of the same BMI; we have applied the idea of **gender separation**.

- So we want to demonstrate that BMI estimation benefits from gender perception, therefore we implement a **gender-mixed model** and **02 gender-based models**.

4.5.7 Training and optimization for (Gender-mixed model / Gender-based model)

- The models are trained using the **Mean Squared Error (MSE)** loss function, appropriate for regression tasks, for **50 epochs** in case of **Female-based Model** and for **30 epochs** in case of **Male-based Model** and **Gender-mixed model**.

- We have fixed the **batch size = 16**.

- The optimization is handled using the **Adam optimizer** with a learning rate of **1e-4** and **1e-5** in case of fine-tuning the Inception-V3 model, and to reduce or penalize the complexity of the model weights which can reduce the overfitting, we have used a **weight decay** (Regularization parameter L2) of λ = **1e-4**.

$$\mathcal{L}_{\text{total}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}_{\text{MSE Loss}} + \underbrace{\lambda \sum_j \|w_j\|_2^2}_{\text{Weight Decay (L2)}} \tag{4.12}$$

- y_i: Ground truth BMI.

- $\hat{y}_i$: Predicted BMI.

- N: Number of samples.

- w_j: Trainable weight parameters (e.g., in your regression head).

- $\lambda = 0.0001$: Weight decay factor.

- Input images and labels are transferred to the appropriate device CUDA, to use the GPU of the PC, reducing time consuming of the training process, and increasing the memory capacity.

4.5.8 Experimental results, interpretations

After building and training models, we have tested the final models on 03 different datasets:

- 1st one: the same as the training dataset **Illinois DOC Faces dataset**. - 2nd one: the **Visual BMI dataset**. - 3rd one: the **Arrest Records dataset**.

Here are the obtained results of female-based model represented by the Table 4.4 :

Table 4.4: Obtained results on our **03 datasets**: Female-based model (with different hyper-parameters).

Models	Inception_v3	Inception_v3	Inception_v3	Inception_v3
MTCNN preprocessing	✗	✓	✓	✓
Fine-tuning	✗	✗	✓	✓
Data augmentation	✗	✗	✗	✓
BMI MAE (Illinois)	5.11	4.98	3.98	3.85
BMI MAE (VisualBMI)	6.72	7.09	6.68	6.05
BMI MAE (Arrest Records)	6.46	5.81	5.21	4.82

- Interpretations: As shown in the Table 4.4:

- The baseline model had a high error rate, indicating poor generalization from face images, this is absolutely due to the small size of the training dataset.
- Integrating MTCNN helped reduce MAE slightly, by improving facial region alignment.
- Fine-tuning significantly improved accuracy (4.98 down to 3.98 MAE), showing that task-specific weight adjustments were crucial.
- The best results were achieved after applying data augmentation, which further reduced the MAE to 3.85, highlighting its effectiveness in enhancing model

robustness and generalization.

Here are the obtained results of male-based model represented by the Table 4.5 :

Table 4.5: Obtained results on our **03 datasets**: Male-based model (with different hyper-parameters).

Models	Inception_v3	Inception_v3	Inception_v3	Inception_v3
MTCNN preprocessing	✗	✓	✓	✓
Fine-tuning	✗	✗	✓	✓
Data augmentation	✗	✗	✗	✓
BMI MAE (Illinois)	**3.52**	**3.46**	**2.84**	**2.78**
BMI MAE (VisualBMI)	**7.04**	**6.95**	**5.74**	**5.44**
BMI MAE (Arrest Records)	**4.30**	**4.01**	**3.15**	**3.12**

- Interpretations: As shown in the Table 4.5:

- Similar to the female model, male facial features slightly benefited from MTCNN preprocessing.
- Fine-tuning improved MAE considerably, from 3.46 to 2.84.
- After applying data augmentation, the MAE dropped to 2.78, showing that male faces were even more predictable under the model compared to female faces.
- This model showed a very good performance with MAE=2.78, this is absolutely due the large number of the training images (56200).

Here are the obtained results of gender-mixed model represented by the Table 4.6 :

Table 4.6: Obtained results on our **03 datasets**: Gender-mixed model (with different hyper-parameters).

Models	Inception_v3	Inception_v3	Inception_v3	Inception_v3
MTCNN preprocessing	✗	✓	✓	✓
Fine-tuning	✗	✗	✓	✓
Data augmentation	✗	✗	✗	✓
BMI MAE (Illinois)	**3.59**	**3.49**	**2.86**	**2.81**
BMI MAE (VisualBMI)	**7.13**	**6.99**	**5.89**	**5.62**
BMI MAE (Arrest Records)	**4.38**	**4.08**	**3.21**	**3.15**

- Interpretations: As shown in the Table 4.6:

- The mixed model performed slightly worse than the male-only model. This could be attributed to the increased variability in the dataset due to the inclusion of both genders, making the task slightly more challenging, in other word: this is compatible with what we montioned previously (BMI estimation benifits from Gender Separation).
- Nevertheless, the progressive improvement through MTCNN, fine-tuning, and augmentation led to a final MAE of 2.81 (3.59 to 2.81), showing that the model was still able to generalize well across genders.

.

- Overall insights:

- MTCNN preprocessing consistently led to better detection and alignment, and slightly improved results with only about 2.5%, because the model can easly see the face and extract best features related with BMI, due to: Approximatively there is no differences between backgrounds in images (the thing that can complex the feature extraction step).

- Fine-tuning is the most impactful strategy across all configurations because when fine-tuning a pre-trained model, this model adapts its previous learned (knowledge) representations to be more relevant to the target domain.

- Data augmentation can lead to improve the performance more with 3% - 2%, because of the comparatively lesser variabilty in poses and lighting conditions in this dataset.

- Gender-specific models slightly outperform the mixed model in their respective domains, especially the male model, which achieved the lowest MAE, and we cannot talk a lot about the error rate for the female model because of the insuffiscient training images (just: 3649x0.70 = 2555 images) to achieve a reasonable performances. So we want to demonstrate that BMI estimation really benefits a little bit from gender perception.

- The BMI MAE values on VisualBMI are very High in all cases, in all models, because this dataset is known for its very noisy nature. It contains: some of very low-resolution images, many Non-frontal or varied poses, also many of images contain objects such as headphones, phones, glasses, or other accessories, which can cause a poor generalization performances.

4.6 Conclusion

This chapter presented our key contributions to BMI estimation from facial images, validated across three diverse datasets. Our experiments showed that Vision Transformers, especially ViT-H-14, outperformed traditional CNNs in BMI prediction tasks, but the training of this model require a huge computational resources and a lot of time. Data augmentation significantly improved model robustness specifically in case of the large variabilty in poses and lighting conditions in this dataset, fine-tuning is a very impactful option but need a suffiscient size of dataset, and multi-task learning strategies further enhanced accuracy that ensurs that the BMI estimator model need informations on the age and the gender of the person.

Gender-specific modeling is a good idea, because BMI estimation really benefits from gender perception.

These findings emphasize the importance of dataset diversity, attention-based architectures, and tailored training strategies for medical image analysis.

In the conclusion, we conclude this study by summarizing our findings and discussing future research directions in automated health monitoring using facial analysis.

Conclusion

This thesis investigated the estimation of Body Mass Index (BMI) from facial images using deep learning techniques. The goal was to design a non-invasive and scalable solution to facilitate health monitoring, especially where traditional BMI measurements are impractical.

We reviewed the importance of BMI in public health and presented a methodology that combines face detection and alignment (via MTCNN) with feature extraction using state-of-the-art deep networks: ViT-H/14, ResNet50, VGG16, and Inception-V3. These models were trained and fine-tuned using three diverse datasets (VisualBMI, Arrest Records, and Illinois DOC Faces), incorporating data augmentation and evaluating the effect of hyperparameters and loss functions.

Key results:

- ViT-H/14 achieved the best results on VisualBMI.
- ResNet50 and VGG16 performed reliably on Arrest Records.
- Inception-V3 showed good generalization on Illinois DOC Faces.
- All models outperformed prior VGGFace-based approaches.

Challenges encountered include dataset bias, input sensitivity, and BMI's limited medical expressiveness. Ethical and privacy considerations were also highlighted.

In summary, our findings confirm the potential of deep learning models for facial-based BMI estimation, offering promising applications in digital health, wellness, and remote diagnostics—provided ethical and clinical validation is pursued.

Summarized Conclusion

- BMI is a widely used indicator of health status.

- Traditional BMI methods are sometimes invasive or impractical.
- Deep learning enables non-invasive, image-based BMI estimation.
- Our models demonstrated improved accuracy over previous methods.

Future Work

- Collect larger and more diverse datasets with demographic labels.
- Explore multi-task models combining BMI with age or gender prediction.
- Extend ViT-H/14 to real-world, high-resolution applications.
- Integrate 3D face modeling and robust preprocessing techniques.
- Partner with healthcare institutions for validation and ethical review.

Bibliography

[1] Miss Neeta B Kumbhare et al. *Body Mass Index Inference Using Facial Features and Machine Learning Algorithms*. 2024.

[2] Enes Kocabey et al. *Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media*. 2017.

[3] A. K. Jain, A. Ross, and S. Prabhakar. “An introduction to biometric recognition”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14(1) (2004), pp. 4–20.

[4] A. Dantcheva et al. “A survey on soft biometrics for human identification”. In: *Pattern Recognition Letters* 33(1) (2011), pp. 38–48.

[5] Jucheng Yang et al. *A Survey on Soft Biometrics for Human Identification*. 2018.

[6] Jiten Sidhpura et al. *Face To BMI: A Deep Learning Based Approach for Computing BMI from Face*. 2022.

[7] Dhanamjayulu C et al. *Identification of malnutrition and prediction of BMI from facial images using real-time image processing and machine learning*. 2021.

[8] *World Obesity Atlas 2024*. 2024.

[9] Single Care Team. *Overweight and obesity statistics 2021*. https://www.singlecare.com/blog/news/obesity-statistics/. Accessed: 2021. 2021.

[10] C.M. Hales et al. “Prevalence of obesity among adults and youth: United States, 2015–2016”. In: *NCHS Data Brief* 288 (2017), pp. 1–8.

[11] Y. C. Wang et al. “Severe obesity in adults cost state medicaid programs nearly dollar 8 billion in 2013”. In: *Health Aff* (2015), pp. 1923–1931.

[12] Rohan Soneja et al. *Body Weight Estimation using 2D Body Image*. 2021.

[13] Dr. O. Aruna et al. *Estimating Body Mass Index from Facial Images*. 2022.

[14] S. Huang. *Obesity Prediction Based on Logistic Regression, Random Forest and Support Vector Machine*. 2022.

[15] K. Zhang et al. “Joint face detection and alignment using multitask cascaded convolutional networks”. In: *IEEE Signal Processing Letters* 23(10) (2016), pp. 1499–1503.

[16] K. He et al. “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification”. In: *IEEE Int. Conf. Comput. Vis.* (2015), pp. 1026–1034.

[17] *Guide to Transfer Learning in Deep Learning*. https://medium.com/@davidfagb/guide-to-transfer-learning-in-deep-learning-1f685db1fc94. n.d.

[18] *Understanding Mean Absolute Error (MAE) in Regression: A Practical Guide*. 2023.

[19] Doug Steen. *Understanding the ROC Curve and AUC.* 2020.

[20] Nadeem Yousaf et al. *Estimation of BMI from Facial Images using Semantic Segmentation based Region-Aware Pooling.* 2021.

[21] *Metrics Evaluation: MSE, RMSE, MAE and MAPE.* 2024.

[22] *Show me your face and I will tell you your height, weight and body mass index.* 2018.

[23] *GAUSSIAN ERROR LINEAR UNITS (GELUS).* 2018.

[24] David J. Fisher. *Illinois DOC labeled faces dataset, Version 1.* https://www.kaggle.com/davidjfisher/illinois-doc-labeled-faces-dataset. 2019.

[25] *Reddit: r/progresspics.* https://www.reddit.com/r/progresspics/. n.d.

[26] https://github.com/Vishan007/face_to_bmi_mlops/tree/main/data. n.d.

[27] *Reddit: r/progresspics.* https://www.reddit.com/r/progresspics/. n.d.

[28] *Vision Transformers (ViT) brought recent breakthroughs in Computer Vision achieving state-of-the-art accuracy with better efficiency.* 2023.

[29] A. Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2021).

[30] *ViT Vision Transformer.* https://medium.com/machine-intelligence-and-deep-learning-lab/vit-vision-transformer-cc56c8071a20. n.d.

[31] *Exploring ResNet50: An In-Depth Look at the Model Architecture and Code Implementation.* 2023.

[32] *VGG-16 CNN Model.* https://www.geeksforgeeks.org/computer-vision/vgg-16-cnn-model/. n.d.

[33] *Softmax Function.* https://botpenguin.com/glossary/softmax-function. n.d.

[34] P. Viola and M. J. Jones. "Robust real-time face detection". In: *International Journal of Computer Vision* 57(2) (2004), pp. 137–154.

[35] *Understanding Inception-v3 Rethinking the Inception Architecture for Computer Vision.* 2022.

[36] Nelida Mirabet-Herranz, Khawla Mallat, and -Luc Dugelay. *New Insights on Weight Estimation from Face Images.* 2023.

[37] *How to Track Your Body Weight.* https://www.fitstream.com/articles/how-to-track-your-body-weight-a213. n.d.

[38] Hera Siddiqui et al. *AI-based BMI Inference from Facial Images: An Application to Weight Monitoring.* 2020.

[39] Venkata Rao Maddumala et al. *Body Mass Index Prediction and Classification Based on Facial Morphological Cues Using Multinomial Logistic Regression.* 2021.

[40] Hong Pan et al. *Hierarchical PSO-Adaboost Based Classifiers for Fast and Robust Face Detection.* 2011.

[41] Guodong Guo. *A computational approach to body mass index prediction from face images.* 2013.

[42] A. Efendi and H.W. Ramadhan. "Parameter estimation of multinomial logistic regression model using least absolute shrinkage and selection operator (LASSO)". In: *AIP Conference Proceeding* (2021).

[43] Asma El Kissi et al. *Soft and hard biometrics for the authentication of remote people in front and side views*. 2016.

[44] P. Samangouei, M. Kabkab, and R. Chellappa. “Discriminative representation learning for soft biometrics”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 694–701.

[45] Y. Wen et al. “BMI Prediction from Facial Images Using Machine Learning Techniques”. In: *IEEE Access* 8 (2020), pp. 132142–132150.

[46] Domestic Violence Database. *National Violent Offender & Domestic Violence Registry*. https://domesticviolencedatabase.net/registryia/?co=Polk&abc=R#abc. 2020.