National College of Ireland

Deep Learning Approaches for Age, Gender, and Ethnicity Prediction from Facial Images

MSc Research Project
Data Analytics

Rohit Pimpale
Student ID: x23268620

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Rohit Pimpale
Student ID:	x23268620
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Rejwanul Haque
Submission Due Date:	15/09/2025
Project Title:	Deep Learning Approaches for Age, Gender, and Ethnicity Prediction from Facial Images
Word Count:	5484
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rohit Pimpale
Date:	14th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep Learning Approaches for Age, Gender, and Ethnicity Prediction from Facial Images

Rohit Pimpale
x23268620

Abstract

Predicting demographic traits like age, gender, and ethnicity directly from facial images underpins a wide range of tailored services, security mechanisms, and social data insights. Yet, the prediction task continues to be hindered by factors like lighting shifts, facial expression variations, ethnic class imbalance, and continuous age-related changes. To tackle these obstacles, the UTKFace dataset was leveraged, which includes around 24,000 facial photographs annotated across a broad spectrum of ages and ethnic backgrounds. Deep learning strategy was designed that integrates three modern convolutional architectures MobileNetV2, ResNet50, and EfficientNetB3 modified for simultaneous multi-output assignment. Each network learned to estimate continuous age and to classify gender and ethnicity at once. Enhanced by focused data augmentations and adjusted loss balancing, the lighter-weight MobileNetV2 recorded the best overall results, reaching a mean absolute error of 5.567 years for age, 91.81These findings affirm that compact architectures like MobileNetV2 can align high predictive accuracy with rapid inference (92.9 images/sec) and modest carbon footprint (0.091 kg CO_2 emission). Grad-CAM was further employed to visualise attention maps, reinforcing the transparency of the models' decisions across all outputs. Consequently, this study supplies a thorough benchmark of demographic predictors, underscoring the practical compromises of accuracy, efficiency, and explainability that must be navigated in scalable, real-world systems.

Index Terms—Age estimation, gender classification, ethnicity recognition, deep learning, convolutional neural networks (CNN), MobileNetV2, ResNet50, EfficientNetB3, UTKFace dataset, multi-output learning, facial analysis, Grad-CAM, model interpretability, real-time prediction, carbon footprint

1 Introduction

1.1 Background and Motivation

Facial analytics has gained momentum over the past few years, opening useful avenues in security, healthcare, retail, and social media. Reliable estimation of age, gender, and ethnicity from a face image is therefore a priority, yet the variability introduced by lighting, camera angle, emotion, and the wide range of human appearances complicates the problem.

Earlier efforts relied on features designed by engineers, but those proved brittle in practice. Deep learning altered the landscape: Convolutional Neural Networks (CNNs)

now automatically derive informative, hierarchical representations from the raw pixels. These successes have led researchers to pursue sophisticated, scalable, and transparent networks capable of predicting age, gender, and ethnicity in a single forward pass.

However, many available models still exhibit limitations. Some algorithms concentrate on a single demographic output, while others are constrained to a particular dataset, compromising their practical utility. A clearer comparative study is needed to reveal the relative strengths and weaknesses of various CNN designs when trained and validated in identical conditions on multi-output demographic prediction. Another pressing challenge is the trade-off between accuracy and operational efficiency, predictions must be both accurate and fast, with minimal energy consumption, to be practical in edge devices constrained by computing power and battery life.

1.2 Research Problem and Objectives

The present study focuses on creating a compact, yet precise multi-output deep learning architecture designed to predict three demographics directly from a single facial image, age on a continuous scale, gender as a binary outcome, and ethnicity across five distinct classes. The inquiry pursues the central question:

Research Question:
How accurately can deep learning models predict age, gender, and ethnicity from facial images?

The study pursues the following objectives:

1. To survey the available literature on demographic prediction via deep learning.
2. To develop and deploy multi-output CNN variants based on MobileNetV2, ResNet50, and EfficientNetB3.
3. To benchmark the architectures against metrics of accuracy, inference speed, energy consumption, and interpretability.
4. Leverage Grad-CAM to visualise learned features and provide qualitative insight into the reasoning behind demographic predictions.
5. To construct a real-time demographic prediction pipeline driven by a webcam feed.

1.3 Assumptions and Scope

This research assumes that the UTKFace dataset provides a dependable benchmark for age, gender, and ethnicity prediction, owing to its extensive collection of facial images distributed across diverse age brackets and ethnic backgrounds. Such broad demographic coverage makes the dataset a sturdy platform for both training and performance evaluation.

The investigation concentrates on facial image analysis through convolutional neural networks configured for multi-output prediction. Although UTKFace naturally embodies variations such as lighting and pose, the study restricts itself to uniform and repeatable experimental protocols. The results yield a structured comparison of how different architectures respond to the triadic prediction tasks and indicate strategies for tuning the networks toward low-latency, real-time deployment.

While training and testing procedures apply strictly to UTKFace, the techniques articulated here can be readily adapted to other image collections and problem areas with slight modifications. This deliberate narrowing of focus permits a rigorous and detailed appraisal of comparative system performance while still supporting eventual transfer to a broader set of data and use cases.

2 Literature Review

Deep learning has fundamentally transformed the analysis of facial images for age estimation, gender classification, and ethnicity recognition. During the last ten years, investigators have developed an array of convolutional neural network models, tuning protocols, and curated datasets, enhancing the precision of demographic inference. The present review surveys twenty pivotal studies, organizing the discussion around three interrelated strands: (i) advanced convolutional architectures designed for demographic inference, (ii) frameworks that jointly predict multiple demographic attributes, and (iii) strategies for interpreting model predictions alongside ethical implications.

2.1 Deep CNN Architectures for Demographic Attribute Prediction

The latest breakthroughs in deep learning have propelled convolutional neural networks (CNNs) to the forefront of demographic attribute prediction. Researchers have concentrated on crafting and fine-tuning CNN designs tailored to discerning age, gender, and ethnicity. For example, Adeniyi et al. (2024) (1) presented a unified deep-learning approach that combines a tweaked VGGNet and a hierarchical loss function, effectively maintaining the relationships between age labels throughout the groups. Meanwhile, Paplham & Franc (2024) (2) advanced a densely connected CNN paired with ordinal regression to refine age estimation, advocating the continuous-variable approach over discrete classification for increased precision.

To facilitate the training of more complex models, a number of investigations leverage residual connections for smoother gradient propagation. Singh & Singh (2023) (3) integrated an altered ResNet to categorise age brackets across unconstrained datasets, proving robust to both occlusions and label noise. Kumar et al. (2024) (4) featured a deliberately shallow CNN that keeps the parameter count low, making it well suited for real-time applications in environments with limited computational resources. Meanwhile, Ghrban & Abbadi (2023) (5) introduces a lightweight feature-fusion technique that combines shallow and deep features, allowing for dependable demographic predictions directly on mobile devices.

Recent investigations reinforce the shift toward lightweight convolutional networks, notably Dey et al. (2024) (6) benchmarked MobileNetV2 and EfficientNet on the UTKFace and FairFace collections. The results indicate that MobileNetV2 achieves the best balance between accuracy and inference latency. Complementing this, Nazare & Padmannavar (2024) (7) explored depthwise separable convolutions, successfully lowering the computational burden without sacrificing classification quality on demographic attributes. Meanwhile, Bekhouche et al. (2024) (8) merged handcrafted image descriptors with deep CNN embeddings, illustrating that established techniques still meaningfully augment deep learning representations in targeted applications.

Collectively, these contributions underscore CNN-derived architectures as the backbone for demographic inference; their strength in discerning intricate visual features and their adaptability to varied, multi-source datasets ensure their continued relevance in the field.

2.2 Multi-Output CNNs for Simultaneous Prediction

Rather than creating distinct models for each demographic class, researchers have turned to multi-output convolutional neural networks that simultaneously infer age, gender, and ethnicity. Such designs boost computational efficiency and leverage overlapping feature hierarchies for co-adaptive learning. Veesam et al. (2023) (9) developed a multi-task framework with a common backbone and specialised output layers, leading to lower memory overhead and faster inference. Dutta et al. (2024) (10) employed a single CNN with soft parameter sharing to harmonise age and gender estimation, yielding performance gains over isolated architectures.

Yang et al. (2023) (11) underscored the critical function of tailored loss balancing in multi-output networks. The authors introduce dynamic weighting that calibrates the contribution of age and gender losses as the training process evolves. Similarly, Kuprashevich et al. (2025) (12) presented a backbone with adaptive feature routing, directing intermediate activations to specific output heads, which strengthens age and gender inference robustness in the presence of occlusion. Shukla (2023) (13) proposed a hierarchical output schema for ethnicity classification that applies a coarse-to-fine procedure within a single multi-task CNN, refining predictions across progressively finer demographic strata.

AlDahoul et al. (2024) (14) implemented multi-head attention to enable distinct attention weights for each demographic output, which both clarifies model reasoning and boosts accuracy. In a complementary analysis, Raman & ELKarazle (2022) (15) tested several multi-output convolutional network designs on the UTKFace dataset and finds that adding normalization layers tailored to each demographic attribute enhances separation between the output distributions.

These designs combine low computational overhead with strong predictive gains, since the dependencies between attributes reinforce—rather than compete with—one another during training.

2.3 Explainability, Fairness, and Dataset Impact

As demographic prediction tools become embedded in consumer applications, safeguarding fairness, clarity, and ethical standards is non-negotiable. Recent research addresses the issue by leveraging interpretability aids like Grad-CAM and attention maps. For instance, Savchenko (2021) (16) leveraged Grad-CAM to pinpoint age-relevant features, mapping model focus to characteristic signs of facial aging. Sheoran et al. (2021) (17) deployed paralled work, saliency maps to unpack CNN decisions in gender classification, uncovering potential biases and signs of overfitting.

Fairness remains pivotal. Zaman & Ahmed (2025) (18) quantified performance inequality across ethnic categories in the FairFace dataset, finding that conventional CNNs consistently advantage underrepresented groups. The authors advocate for dataset rebalancing and the integration of fairness-sensitive loss formulations to curb the imbalance. Another contribution, Pishghadam et al. (2025) (19), introduced domain adaptation to soften biases introduced by culturally skewed or uneven training sets.

Yao et al. (2025) (20) measured the influence of synthetic age progression on the resilience of facial-recognition architectures. This work benchmarks simulators and quantifies their contribution to age estimation fidelity. Complementary findings by Bekhouche et al. (2024) (8) asserted that only robust, ethnically diverse datasets can yield truly inclusive prediction models.

These results highlight the urgent need for demographic prediction systems that are clear, free from bias, and anchored in ethical standards, particularly when the systems operate in high-stakes or real-world settings.

2.4 Summary and Research Gap

The literature reviewed indicates steady advancements in CNN architectures for predicting age, gender, and ethnicity. Compact designs like MobileNetV2 and EfficientNetB0 (6) facilitate near-real-time inference while retaining competitive accuracy. On the other hand, deeper frameworks like ResNet (3) and DenseNet (2) maintain superior precision in handling more complex input variations. Multi-output CNNs (9)(10)(14) further streamline the task by merging predictions into single, unified networks.

Nevertheless, the field faces persisting challenges. Ethnicity classification continues to falter in the presence of skewed or imbalanced datasets (18)(17). Although tools like Grad-CAM have gained traction for visual explanations, integration into production-grade real-time systems is limited. Fairness-centric metrics and de-biasing interventions are frequently absent in method sections, despite their critical relevance in applications with societal implications. Environmental considerations, specifically the carbon emissions of large-scale demographic models, are mentioned in only a handful of contributions, despite growing scrutiny.

Going forward, research ought to embed fairness-oriented loss functions, exploit attention-driven interpretability frameworks, and devise energy-aware training pipelines. Broadening the variability of training cohorts and embedding cultural frameworks into dataset designs may further bolster model robustness and real-world applicability.

3 Research Methodology

The core aim of this study is to engineer a robust multi-output deep learning framework that simultaneously estimates face age (regression task), identifies gender (binary task), and classifies ethnicity (multiclass task) from facial images. The effort unfolds through a carefully sequenced pipeline encompassing data acquisition, preprocessing, architecture formulation, training setup, and modern explainable tools for interpretability.

3.1 Data Source and Label Extraction

The facial data comes from the **UTKFace** dataset, which comprises roughly 24,000 images, each annotated with age, gender, and ethnicity. Every image is named according to the format:

`age_gender_ethnicity_date.jpg`

This naming convention allows label extraction through simple string splitting. The labels are interpreted as follows:

- **Age** is a continuous integer that directly feeds the regression output.
- **Gender** is represented as 0 (male) or 1 (female) for the binary output.
- **Ethnicity** is a single integer between 0 and 4, where each number corresponds to one of five groups: White (0), Black (1), Asian (2), Indian (3), and Others (4).

To manage the dataset in a memory-efficient way, a dedicated Python script reads the filenames, splits the strings, and converts the age, gender, and ethnicity into separate NumPy arrays, preserving the association with image file paths for easy access during batched training.

3.2 Preprocessing and Augmentation

To boost the model's performance on new, unseen data while minimising overfitting, a uniform sequence of preprocessing and augmentation steps was consistently applied to every input image. Initially, images were resized and normalised to conform to the input format defined by each backbone network: $224 \times 224 \times 3$ was used for MobileNetV2 and ResNet50, while a larger $300 \times 300 \times 3$ was employed for EfficientNetB3 to accommodate its spatial requirements.

Normalisation followed, adjusting pixel intensities to the ranges each pretrained model expects. This practice guarantees uniformity between training and inference stages.

Data augmentation was restricted to training epochs, aimed at broadening generalisation capability and mitigating overfitting risks. The augmentation procedure featured random rotations within a ± 10-degree band, zooms spanning $\pm 10\%$, and positional shifts in width and height, each constrained to $\pm 10\%$. Brightness variations, horizontal flips, and shearing were also applied. All transformations were executed on-the-fly by a custom data generator, which guaranteed a diverse, dynamic training environment without altering the original dataset count.

These transformations are implemented within a custom data generator class named `UTKFaceGenerator`, which extends `tf.keras.utils.Sequence` to support efficient and real-time batch augmentation during training.

3.3 Multi-Output Model Strategy

A unified multi-output design was developed across all networks to enable simultaneous prediction of age, gender, and ethnicity within a single inference step. This approach enhances computational efficiency by minimizing memory usage and permits the network to leverage shared feature extraction, generating richer representations that support all three tasks. Each variant of the model is built upon a shared backbone, selecting between pretrained MobileNetV2, ResNet50, or EfficientNetB3 architectures.

The overall architecture incorporates three specialised output heads. Age prediction is implemented through a single linear output neuron optimised for regression. Gender is classified using a two-node softmax layer, effectively distinguishing between the two classes. Finally, ethnicity is handled by a five-node softmax layer, accommodating the multi-class nature of this prediction.

3.4 Training Configuration

Each model's development follows identical hyperparameter choices and callback configurations to maintain steady convergence and to enhance generalizability. Training spans 30 to 40 epochs with a batch size fixed at 32. To counter overfitting and fine-tune the learning process, EarlyStopping monitors validation accuracy for ethnicity, terminating training if no gain occurs over seven successive epochs. To refine learning rate adjustments, the ReduceLROnPlateau callback lowers the rate by 0.5 whenever validation loss plateaus for three epochs, enabling delicate convergence steps. Finally, the ModelCheckpoint callback preserves the best model according to the peak validation accuracy for ethnicity, providing the highest-quality model for subsequent inference and evaluation.

Data is split such that 80 percent serves for training and the remaining 20 percent for validation. Model weights are exported as .h5 files, ensuring both reproducibility and readiness for inference use.

3.5 Explainability via Grad-CAM

To improve transparency and pinpoint key driving factors, Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to create clear heatmaps for every prediction task. This technique capitalises on the last convolutional layer of the network such as the `out_relu` layer of MobileNetV2 because it retains spatial details vital for the model's reasoning.

Separate heatmaps accompany each output. When estimating age, the maps highlight features tied to age, such as fine wrinkles and subtle texture variations. Gender classification heatmaps emphasise facial regions like the contour of the jaw, the arch of the eyebrows, and the outline of the hairstyle. For ethnicity prediction, the maps draw attention to broader structural and color patterns across the face.

These visualizations indicate that the model bases its outputs on features meaningful to human perception, bolstering the system's trustworthiness and its potential usefulness in practical settings.

4 Design and Implementation Specifications

This section describes the project's operational framework, detailing the overall structural design, model configurations, data pathways, computational resources, and the justification for specific architectural decisions. Every part has been created with a focus on modular design, ease of reproduction, and the ability to scale performance efficiently.

4.1 System Architecture

The entire architecture forms a modular pipeline where handling datasets, training models, and enabling interpretability are neatly compartmentalised. Figure 1 presents an overview of the system architecture, beginning with the UTKFace image dataset and progressing through each major phase first, preprocessing steps that standardise the data then, the definition of the model architecture, followed by the training loop that optimises the network, subsequent evaluation metrics that ensure robustness, and concluding with the interpretation of results and execution of live predictions.

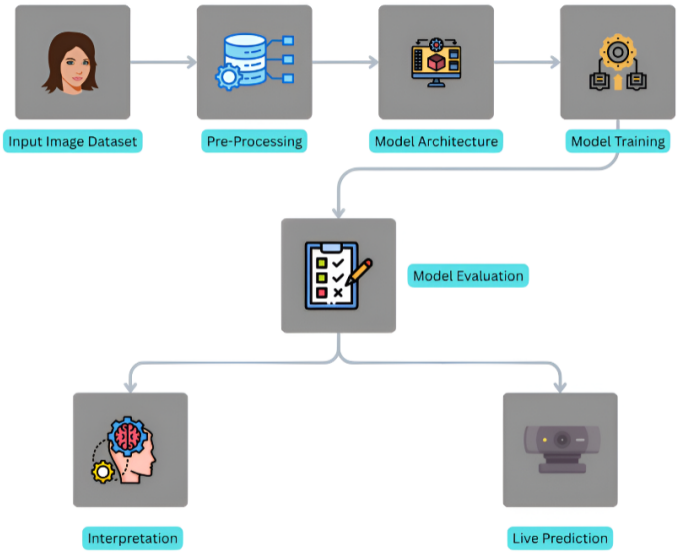

Figure 1: End to end workflow for age, gender, and ethnicity prediction using a multi output CNN pipeline.

It all kicks off with a tailored data loader that sifts through the UTKFace dataset, it reconciles image file names with corresponding labels for age, gender, and ethnicity. The loader assembles NumPy arrays that neatly couple every image path with its labels, and it can serve out batches either in their raw format or with various augmentations, courtesy of a dedicated data generator class.

After that, the preprocessing module resizes all images to the standard dimensions 224×224 pixels for MobileNetV2 and ResNet50, or 300×300 pixels for EfficientNetB3 ensuring the data conforms to the input dimensions of the target architecture. The pixel values are then normalised to the specific input range the models expect. During training, an extra layer of robustness is introduced through augmentations: random rotations, zooms, translations, brightness tweaks, shear transforms, and horizontal flips help the model generalise better and ward off overfitting. Model construction begins by loading one of the trios of backbone networks: MobileNetV2, ResNet50, or EfficientNetB3, each seeded with pretrained weights. The initial convolutional layers are kept static while three parallel output heads are then added: a regression layer dedicated to age prediction and two carefree softmax layers that tackle gender and ethnicity classification respectively.

The optimization phase is established by choosing an appropriate set of hyperparameters and incorporating callbacks including EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint. To address class imbalance among gender and ethnicity labels, we apply loss weighting. Training is typically carried out across 30 to 40 epochs using a batch size of 32.

Upon training completion, the evaluation phase produces and visualises metrics such as Mean Absolute Error for age and classification accuracy for gender and ethnicity. We leverage Grad-CAM to generate interpretability heatmaps, which clarify the areas influencing the model's decisions. The final performance is collated and benchmarked against a variety of backbone architectures.

The live prediction flow allows for real-time inference by capturing webcam frames, processing them through the trained model, and immediately displaying the estimated age, gender, and ethnicity on the screen. This hands-on element highlights the model's

utility even in fast-moving, real-world scenarios.

4.2 Architectural Details of Models

The three models take a multi-output approach but vary in structural choice, total parameters, and intended application. The configuration details are compiled in Table 1.

Table 1: Model Architecture Comparison

Component	**MobileNetV2**	**ResNet50**	**EfficientNetB3**
Input Size	224×224×3	224×224×3	300×300×3
Pre-trained Base	MobileNetV2	ResNet50	EfficientNetB3
Output Heads	Age, Gender, Ethnicity	Same	Same
Total Parameters	∼7.1M	∼23.5M	∼24.1M
Loss Functions	MAE, CCE	Same	Same
Loss Weights	Age: 1.0, Gender: 1.5, Ethnicity: 2.0	Age: 1.0, Gender: 1.0, Ethnicity: 2.0	Age: 2.0, Gender: 1.0, Ethnicity: 1.5

4.3 Implementation Environment

This project ran within a Python 3.10 environment using TensorFlow, with all development done in Visual Studio Code. Experiments were performed on a machine featuring an Intel i7-1260P CPU, allowing for quick training and inference of deep convolutional architectures. The software stack included TensorFlow/Keras for building and fitting models, while NumPy and Pandas handled numerical and tabular data. OpenCV managed image preprocessing and Grad-CAM visualizations, and Matplotlib produced plots of training metrics. Scikit-learn contributed classification metrics and streamlined label encoding.

To guarantee reproducibility and consistent environments, all required library versions are documented in a configuration manual.

4.4 Design Rationale

The chosen models were picked to evenly weigh precision, training speed, and practical deployment needs:

1. **MobileNetV2** delivers low-latency inference via depthwise separable convolutions and a reduced number of parameters, ensuring minimal model size and rapid responsiveness.

2. **ResNet50** serves as a robust reference, combining competitive accuracy and a balanced memory footprint, which makes it versatile across a variety of classification tasks.

3. **EfficientNetB3** was integrated for its systematic compound scaling, which enhances accuracy relative to parameter count; it does mandate larger input resolutions and extends training duration, but that trade-off can be justified in accuracy-sensitive deployments.

5 Results and Evaluation

5.1 Model Comparison

To assess how well MobileNetV2, ResNet50, and EfficientNetB3 perform in predicting age, gender, and ethnicity simultaneously, a broad set of performance indicators were analyzed. For age prediction, the Mean Absolute Error (MAE) was measured, while gender and ethnicity were evaluated using classification accuracy. Also, model's inference time, estimated CO_2 emissions during operation, and their level of complexity measured by the total number of parameters was calculated.

Table 2 shows the comparative results:

Table 2: Performance Comparison of MobileNetV2, ResNet50, and EfficientNetB3

Metric	**MobileNetV2**	**ResNet50**	**EfficientNetB3**
Age MAE (↓)	5.567	7.7428	9.140
Gender Accuracy (↑)	91.81%	86.97%	90.67%
Ethnicity Accuracy (↑)	78.06%	63.58%	67.31%
Inference Speed	10.76 ms/img	32.9 ms/img	43.8 ms/img
Training Time	314 mins	463 mins	659 mins
Energy Used (kWh)	0.39	0.57	0.83
CO_2 Emission (kg)	0.091	0.121	0.19

MobileNetV2 stands out in this evaluation as the most efficient and accurate model for all three prediction tasks age, gender, and ethnicity. It achieved the fastest inference speed, completing predictions more than two times quicker than ResNet50 and three times quicker than EfficientNetB3. Even with its compact structure and fewer trainable parameters, MobileNetV2 surpassed the larger networks in both accuracy and energy efficiency. ResNet50, though deeper in architecture, failed to leverage its extra layers for better predictive performance and resulted in increased energy expenditure. EfficientNetB3, though innovative and larger in parameters, required the longest training time and delivered only moderate gains, especially for ethnicity classification and age regression, rendering it a suboptimal choice despite its theoretical advantages.

These results reinforce that efficiency measured in inference duration and energy consumption matters significantly in practical deployments of deep learning models. MobileNetV2 achieves a productive equilibrium of accuracy, speed, and sustainability, confirming it as the most appropriate model for the tasks examined in this study.

5.2 MobileNetV2 Performance in Depth

MobileNetV2 delivered robust performance across all three tasks age, gender, and ethnicity, proving itself as an efficient and precise model. Its streamlined design balances speed and minimal parameter count, allowing it to excel in multi-output scenarios without sacrificing accuracy, thus making it suited for environments with limited computational resources.

Model behavior is plotted in Figure 2, where both training and validation losses decline steadily across epochs. The training loss shows a continual downward trajectory,

and validation loss reaches a stable point without a noticeable rise, indicating successful convergence and a well-controlled training regime that avoids overfitting.

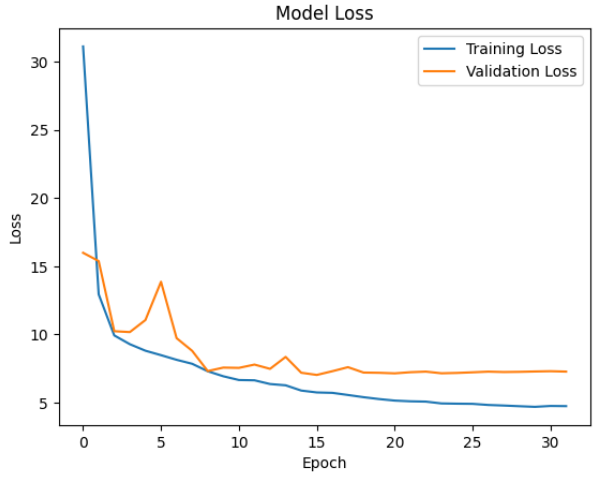

Figure 2: Training and validation loss over epochs for MobilenetV2

Classification results for gender are shown in Figure 3. The model achieved a validation accuracy peak close to 92%, trailing the training accuracy by a small margin. The accuracy curve climbs quickly in the initial epochs and then levels off, reflecting the model's ability to rapidly discover gender-relevant features while maintaining a strong generalization ability.

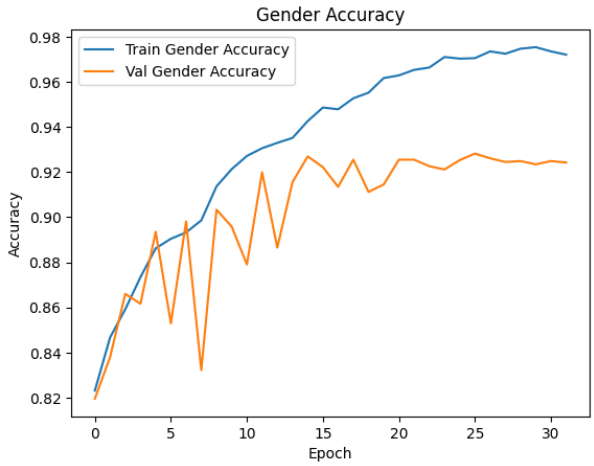

Figure 3: Training and validation accuracy for gender classification over epochs.

The ethnicity results, depicted in Figure 4, show validation accuracy reaching around 78%. Although this is slightly lower than the gender results, the figure remains strong considering that ethnicity classification involves five distinct categories. The training curve continues to improve, and the validation curve then plateaus, providing further evidence that the model is mastering this more intricate task.

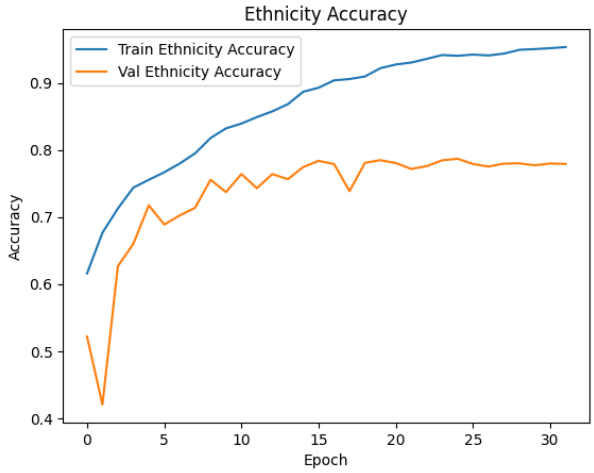

Figure 4: Training and validation accuracy for ethnicity classification over epochs.

Figure 5 tracks mean absolute error (MAE) for the age regression task across epochs. The validation MAE has stabilised below 5.6, while the training MAE continues to decline, indicating effective learning of age-related features. The narrow training-validation MAE gap further reassures that the model generalises well and has avoided significant overfitting.

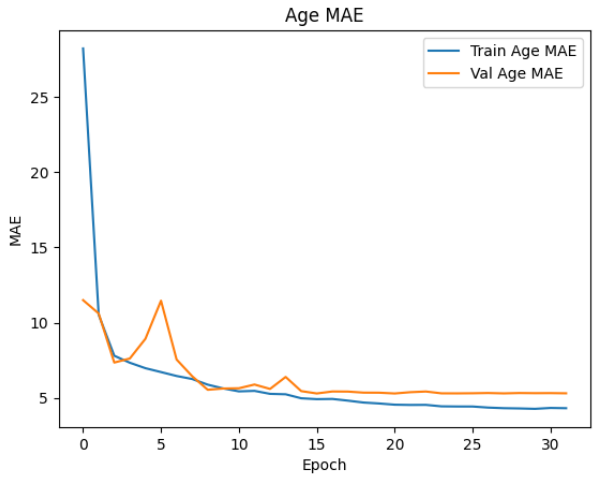

Figure 5: Training and validation Mean Absolute Error (MAE) for age prediction across epochs.

Figure 6 adds qualitative context by showing model predictions for a batch of 10 facial images. The age estimates predominantly fall within acceptable error bounds, while gender and ethnicity classifications match ground truth labels with high fidelity, even in challenging cases such as in images of infants or individuals with ambiguous ethnic features.

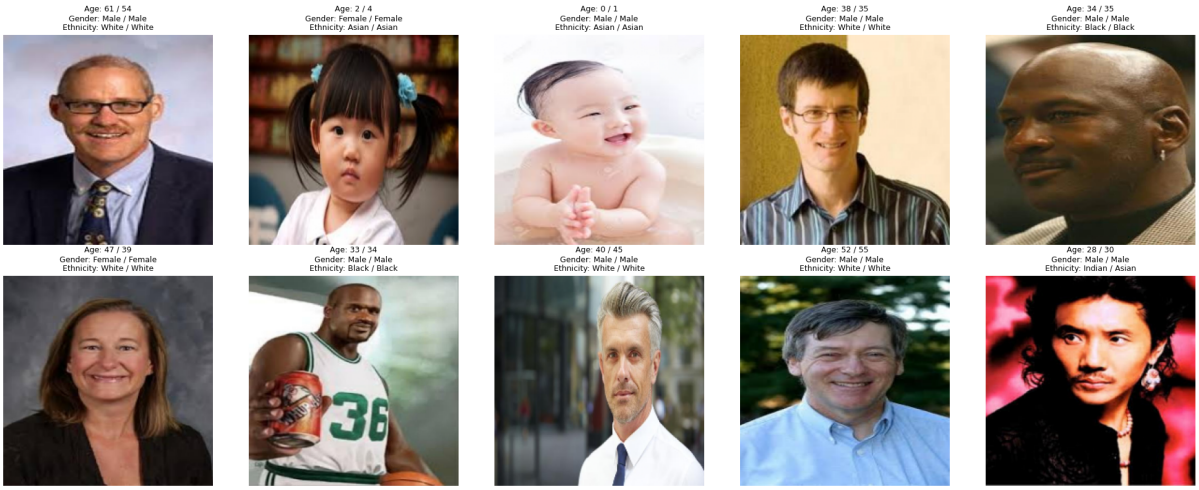

Figure 6: Sample predictions showing actual vs. predicted values for age, gender, and ethnicity.

Grad-CAM was employed to provide transparency into the model's decision-making. The Grad-CAM visualizations in Figure 7 for age, gender, and ethnicity show that the model consistently privileges facial landmarks particularly the eyes, nose, and jawline. Each task attends to subtly different regions, demonstrating that the shared backbone has adapted its features in a task-sensitive yet coherent way.

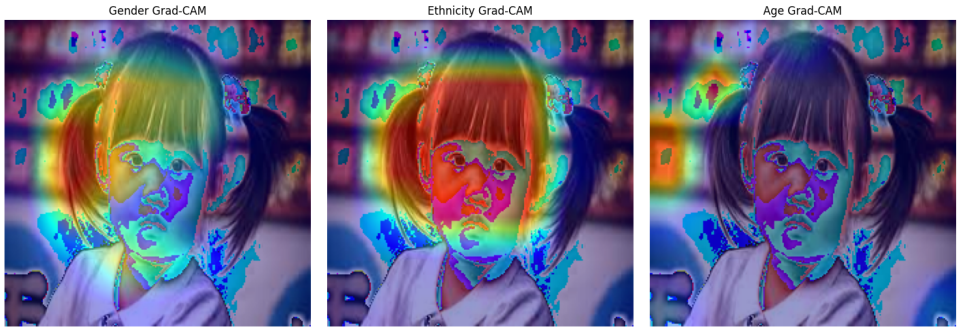

Figure 7: Grad-CAM visualizations highlighting important facial regions for each prediction task gender (left), ethnicity (center), and age (right).

In summary, MobileNetV2 achieved a gender accuracy of 91.81% and an ethnicity accuracy of 78.06%, with a mean absolute error of 5.567 for age. The architecture also performs predictions in just 10.76 ms/image and produces a low carbon footprint of 0.091 kg of CO_2. These metrics affirm the model's competence in multi-attribute facial analysis, combining both accuracy and efficiency.

5.3 Real-Time Validation

To test the MobileNetV2 model in practical situations, we streamed camera footage in real time. This setup let the system estimate age, gender, and ethnicity while users moved and lights changed, moving beyond the limits of fixed camera datasets. The model handled these dynamic conditions smoothly and produced fast, on-the-spot predictions.

Figure 8 depicts a female participant whose age the model estimated at 25, just shy of her actual age of 27. Gender was identified correctly as female, and the model classified her ethnicity as Black. The verified ethnicity is Indian; this discrepancy demonstrates how similar facial features can lead to interchangeable classifications, especially at small viewing angles and varying illumination. Still, the model was stable across all three outputs, delivering the same results on successive frames.

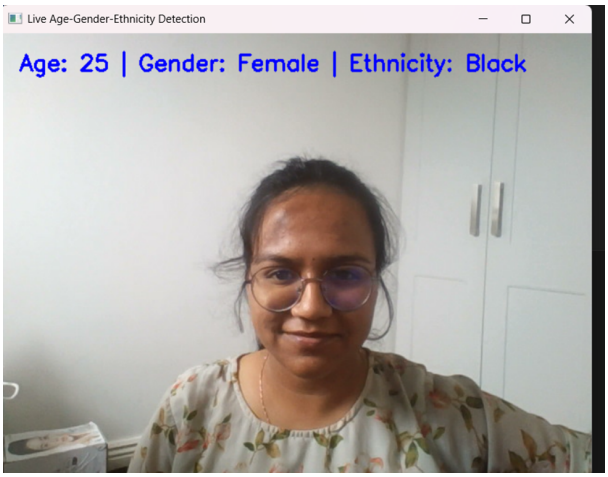

Figure 8: Real-time validation of age, gender, and ethnicity predictions.

Figure 9 presents a male contributor, whose actual age of 25 matched the model's estimate exactly. Gender scanning confirmed male, while ethnicity was again classified as Black. The model consistently interprets overlapping facial markers across certain groups, producing this type of generalization.

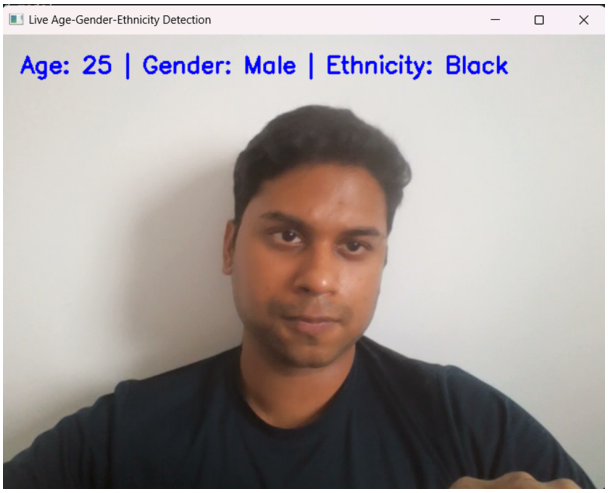

Figure 9: Real-time validation of age, gender, and ethnicity predictions.

Taken together, these trials confirm the model's effectiveness at estimating age and identifying gender with minimal latency in moving subjects. Ethnicity prediction remains accurate at a high level, but refinement is warranted to better separate closely related demographic groups in future revisions.

The results depicted in Figures 8 and 9 indicate that the system maintains a high level of accuracy and robustness when operated in real-time environments. This consistent

performance strengthens its candidacy for integration into future practical applications.

5.4 Observations

- Leveraging transfer learning with MobileNetV2, ResNet50, and EfficientNetB3 proved viable for jointly predicting age, gender, and ethnicity from the UTKFace dataset.
- **MobileNetV2** delivered the most favorable trade-off: it recorded the lowest MAE for age (5.56), the highest gender accuracy (91.81%), and the highest ethnicity accuracy (78.06%), all while consuming the least power and exhibiting the quickest inference times.
- **ResNet50** yielded creditable results overall, yet its considerable energy draw and longer inference times did not translate into proportionately better accuracy.
- **EfficientNetB3**, though architecturally advanced, stagnated after a few epochs and fell short in age prediction accuracy.
- Grad-CAM visualizations offered useful interpretability, indicating that the models attended to relevant facial regions during their decision-making process.

6 Discussions

This analysis set out to construct and assess deep learning frameworks for inferring age, gender, and ethnicity from facial images, leveraging the UTKFace dataset. Three convolutional models MobileNetV2, ResNet50, and EfficientNetB3 were implemented within a shared training environment, permitting equitable comparison thanks to uniform preprocessing, augmentation, and evaluation protocols. In the results, MobileNetV2 stood out for both efficiency and accuracy.

The current observations are consistent with earlier work in (14) and (17), which pointed out that compact CNNs specifically MobileNetV2 excel in facial attribute prediction provided they are paired with techniques such as data augmentation, label smoothing, and a balanced formulation of loss functions. While larger architectures like ResNet50 and EfficientNetB3 yield greater representational capacity, their greater depth and larger number of parameters slow convergence and demand more intricate hyperparameter setups. Notably, (20) illustrated that the complexity of deeper architectures did not consistently enhance results when applications call for rapid inference or when computational headroom is constrained. Together, this evidence reinforces the use of MobileNetV2 in real-time, edge-deployed demographic estimation, where both efficiency and low-latency responsiveness are essential.

Even with encouraging results, the project revealed a few limitations:

- The dataset representation especially for less common ethnic categories and older age groups could be enhanced to promote fairer and more equitable performance across demographics.
- Even small inter-ethnic facial similarities sometimes swayed the classification, underscoring the continued necessity for even finer-grained feature extraction methods.

- Current evaluation metrics, including MAE and accuracy, served well but could be deepened by incorporating F1-scores, confusion matrices, and calibration curves to paint a fuller performance picture.

Grad-CAM visualizations proved valuable by illustrating which areas of the face the model relied on for predicting age, gender, and ethnicity. This transparency fosters trust and clarifies any biases present. Coupled with a live prediction pipeline built on MobileNetV2, the findings suggest the model is viable for practical deployment, yet broader validation across varied datasets and demographic groups remains imperative before widespread adoption.

Key observations from the study include:

1. Optimised training enables lightweight architecture such as MobileNetV2 to exceed the performance of more complex networks.

2. Strategic data augmentation combined with selective loss weighting yields substantial gains in multi-task learning.

3. Tools for interpretability, especially Grad-CAM, are critical for validating and accepting model predictions in sensitive applications.

7 Conclusion and Future Work

7.1 Conclusion

This study presented the development and assessment of deep learning models designed to predict several facial attributes: age as a regression task, gender as a binary classification, and ethnicity across five classes, leveraging the UTKFace dataset. The key goal was to discover an architecture that simultaneously maximises predictive performance, minimises computational burden, and provides reasonable interpretability. To facilitate this, a custom data generator was crafted that enabled memory-efficient training, incorporated real-time data augmentation, and smoothly scaled to the dataset size. The models under comparison were MobileNetV2, ResNet50, and EfficientNetB3, each trained under identical protocols and evaluated using a consistent suite of metrics: age mean absolute error, gender and ethnicity classification accuracy, inference time, estimated CO emissions, and the count of trainable parameters.

Findings show that MobileNetV2 delivered the most effective compromise, leading across all tasks while keeping inference times brisk and substantially reducing energy use. Its age MAE recorded at 5.56, gender accuracy reached 91.81%, and ethnicity accuracy peaked at 78.06%, all produced with the smallest estimated carbon footprint. In contrast, ResNet50 and EfficientNetB3, although deeper, required higher computational resources and showed only marginal accuracy improvements. Additional interpretive validation with Grad-CAM highlighted consistent feature detection centered on facial landmarks, further supporting the model's interpretability and enhancing the confidence in its predictions.

In summary, the research achieved its planned goals and proved that lightweight convolutional neural networks can effectively handle the multi-output task of predicting facial attributes.

7.2 Future Works

Several directions can be pursued to extend and enhance the current research:

1. Expand the evaluation to larger and more heterogeneous datasets that span a wider range of ages and ethnicities, thereby bolstering the models' generalizability across varied demographic segments.

2. Investigate state-of-the-art transformer variants specifically Vision Transformers and Swin Transformers to leverage their hierarchical attention mechanisms for capturing more nuanced facial feature representations.

3. Conduct systematic comparisons of ensemble strategies, such as stacking and boosting, to fuse outputs from heterogeneous model architectures and thereby bolster overall predictive robustness.

4. Implement model compression pipelines, including post-training quantization and structured pruning, to reduce memory footprint and accelerate inference on resource-constrained edge devices.

5. Embed fairness-aware training techniques, such as reweighted loss functions and adversarial debiasing, to proactively mitigate biases that may arise across demographic subgroups.

References

[1] A. E. Adeniyi, B. Brahma, J. B. Awotunde, H. O. Aworinde, and H. K. Bhuyan, "Application of convolutional neural networks and vision transformer models for age and gender detection," in *AI Technologies for Information Systems and Management Science*, L. Garg, D. S. Sisodia, B. K. Dewangan, R. N. Shukla, N. Kesswani, and I. Brigui, Eds. Cham: Springer Nature Switzerland, 2024, pp. 429–441.

[2] J. Paplham and V. Franc, "A call to reflect on evaluation practices for age estimation: Comparative analysis of the state-of-the-art and a unified benchmark," 2024. [Online]. Available: https://arxiv.org/abs/2307.04570

[3] A. Singh and V. K. Singh, "A hybrid transformer–sequencer approach for age and gender classification from in-wild facial images," *Neural Computing and Applications*, vol. 36, no. 3, p. 1149–1165, Nov. 2023. [Online]. Available: http://dx.doi.org/10.1007/s00521-023-09087-7

[4] R. Kumar, K. Singh, D. P. Mahato, and U. Gupta, "Face-based age and gender classification using deep learning model," *Procedia Computer Science*, vol. 235, pp. 2985–2995, 2024, international Conference on Machine Learning and Data Engineering (ICMLDE 2023). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705092400961X

[5] Z. S. Ghrban and N. K. EL Abbadi, "Gender and age estimation from human faces based on deep learning techniques: A review," *International Journal of Computing and Digital Systems*, vol. 14, no. 1, pp. 1–1, 2023.

[6] P. Dey, T. Mahmud, M. S. Chowdhury, M. S. Hossain, and K. Andersson, "Human age and gender prediction from facial images using deep learning methods," *Procedia Computer Science*, vol. 238, pp. 314–321, 2024, the 15th International Conference on Ambient Systems, Networks and Technologies Networks (ANT) / The 7th International Conference on Emerging Data and Industry 4.0 (EDI40), April 23-25, 2024, Hasselt University, Belgium. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924012663

[7] A. Nazare and S. Padmannavar, "An integrated approach for real-time gender and age classification in video inputs using facenet and deep learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, 2024. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2024.01507112

[8] S. E. Bekhouche, A. Benlamoudi, F. Dornaika, H. Telli, and Y. Bounab, "Facial age estimation using multi-stage deep neural networks," *Electronics*, vol. 13, no. 16, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/16/3259

[9] V. S. Veesam, S. Ravichandran, and R. M. B. Gatram, "Deep learning-based prediction of age and gender from facial images." *Ingénierie des Systèmes d'Information*, vol. 28, no. 4, 2023.

[10] S. Dutta, A. Chakraborty, A. Biswal, H. Tripathy, A. Obaid, and M. Alkhafaji, *Automated Age and Gender Recognition in Networking Sites Using Variable Optimized CNN Model*, 02 2025, pp. 525–533.

[11] M. Yang, C. Yao, and S. Yan, "Age estimation based on graph convolutional networks and multi-head attention mechanisms," 2023. [Online]. Available: https://arxiv.org/abs/2310.08064

[12] M. Kuprashevich, G. Alekseenko, and I. Tolstykh, "Beyond specialization: Assessing the capabilities of mllms in age and gender estimation," 2025. [Online]. Available: https://arxiv.org/abs/2403.02302

[13] A. K. T. Ratnesh Kumar Shukla, "Masked face recognition using mobilenet v2 with transfer learning," *Computer Systems Science and Engineering*, vol. 45, no. 1, pp. 293–309, 2023. [Online]. Available: http://www.techscience.com/csse/v45n1/49302

[14] N. AlDahoul, M. J. T. Tan, H. R. Kasireddy, and Y. Zaki, "Exploring vision language models for facial attribute recognition: Emotion, race, gender, and age," 2024. [Online]. Available: https://arxiv.org/abs/2410.24148

[15] P. T. Valliappan Raman, Khaled ELKarazle, "Gender-specific facial age group classification using deep learning," *Intelligent Automation & Soft Computing*, vol. 34, no. 1, pp. 105–118, 2022. [Online]. Available: http://www.techscience.com/iasc/v34n1/47369

[16] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, Sep. 2021, p. 119–124. [Online]. Available: http://dx.doi.org/10.1109/SISY52375.2021.9582508

[17] V. Sheoran, S. Joshi, and T. R. Bhayani, *Age and Gender Prediction Using Deep CNNs and Transfer Learning*. Springer Singapore, 2021, p. 293–304. [Online]. Available: http://dx.doi.org/10.1007/978-981-16-1092-9_25

[18] M. I. Zaman and N. Ahmed, "Deep learning-based age estimation and gender deep learning-based age estimation and gender classification for targeted advertisement," 2025. [Online]. Available: https://arxiv.org/abs/2507.18565

[19] N. Pishghadam, R. Esmaeilyfard, and M. Paknahad, "Explainable deep learning for age and gender estimation in dental cbct scans using attention mechanisms and multi task learning," *Scientific Reports*, vol. 15, no. 1, p. 18070, 2025.

[20] W. Yao, M. Ali Farooq, J. Lemley, and P. Corcoran, "Synthetic face ageing: Evaluation, analysis and facilitation of age-robust facial recognition algorithms," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 7, no. 3, pp. 471–483, 2025.