
Facial Image-Based Gender Identification Using MobileNetV2

Nilam Tri Astuti ^{1*}, Sunardi ², Abdul Fadlil ³

¹ Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

^{2,3} Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

e-mail: ¹ 2408048024@webmail.uad.ac.id, ² sunardi@mti.uad.ac.id, ³ fadlil@mti.uad.ac.id

*Corresponding author

Abstract

Face-based gender classification has become an important component in security, demographic analysis, and personalized digital services, where rapid and automated facial analysis is required. This study proposes a lightweight gender classification framework based on MobileNetV2 with ImageNet transfer learning for deployment in resource-constrained environments. Experiments were conducted using a public Kaggle dataset containing 15,934 facial images (7,884 male and 8,050 female). The preprocessing pipeline consisted of image resizing, Haar Cascade face detection, pixel normalization, and brightness-contrast augmentation. To prevent data leakage, the dataset was split into training, validation, and testing subsets (70:20:10) before augmentation, ensuring that augmented versions of the same image were confined to the training subset. The model was trained using a two-stage transfer learning strategy in which the classification head was trained first, followed by fine-tuning of the last 30 layers of MobileNetV2. Performance was evaluated on a held-out test set comprising 707 unaugmented facial images using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The proposed model achieved a test accuracy of 94.91%, with precision, recall, and F1-score of 96.42%, 94.49%, and 95.44% for the female class, and 93.04%, 95.45%, and 94.23% for the male class, demonstrating balanced classification performance across both classes. A qualitative evaluation using ten independently sourced facial images correctly classified seven samples, suggesting that further investigation is required to improve generalization beyond the training data distribution. The results indicate that MobileNetV2 provides an effective balance between classification accuracy and computational efficiency, making it suitable for deployment on smartphones, embedded systems, and other resource-constrained computer vision applications.

Keywords: Gender Classification, Facial Image, Deep Learning, MobileNetV2, Transfer Learning

Abstrak

Klasifikasi jenis kelamin berbasis citra wajah merupakan salah satu komponen penting dalam sistem keamanan, analisis demografi, dan layanan digital yang dipersonalisasi karena memungkinkan analisis wajah secara otomatis dan cepat. Penelitian ini mengusulkan kerangka klasifikasi jenis kelamin berbasis MobileNetV2 dengan pendekatan transfer learning dari ImageNet yang dirancang untuk lingkungan dengan keterbatasan sumber daya komputasi. Eksperimen menggunakan dataset publik Kaggle yang terdiri atas 15.934 citra wajah (7.884 laki-laki dan 8.050 perempuan). Tahap prapemrosesan meliputi perubahan ukuran citra, deteksi wajah menggunakan Haar Cascade, normalisasi piksel, serta augmentasi brightness-contrast. Untuk mencegah data leakage, pembagian dataset menjadi data latih, validasi, dan uji (70:20:10) dilakukan sebelum augmentasi sehingga seluruh citra hasil augmentasi hanya berada pada data latih. Model dilatih menggunakan strategi transfer learning dua tahap, yaitu pelatihan classifier head diikuti fine-tuning pada 30 layer terakhir MobileNetV2. Evaluasi dilakukan pada 707 citra uji tanpa augmentasi menggunakan metrik accuracy, precision, recall, F1-score, confusion matrix, dan ROC-AUC. Model yang diusulkan memperoleh akurasi 94,91%, dengan nilai precision, recall, dan F1-score masing-masing sebesar 96,42%, 94,49%, dan 95,44% untuk kelas perempuan serta 93,04%, 95,45%, dan 94,23% untuk kelas laki-laki, yang menunjukkan performa klasifikasi yang seimbang pada kedua kelas. Pengujian kualitatif menggunakan sepuluh citra wajah independen berhasil mengklasifikasikan tujuh citra dengan benar, sehingga diperlukan penelitian lebih lanjut untuk meningkatkan kemampuan generalisasi pada distribusi data yang berbeda. Hasil penelitian menunjukkan bahwa MobileNetV2 mampu memberikan keseimbangan yang baik antara

akurasi klasifikasi dan efisiensi komputasi sehingga sesuai untuk diterapkan pada smartphone, sistem embedded, dan aplikasi computer vision dengan sumber daya terbatas.

Kata Kunci: Klasifikasi Jenis Kelamin, Citra Wajah, Deep Learning, MobileNetV2, Transfer Learning

1. Introduction

Face is one of the most widely used biometric traits for recognizing human identity because facial appearance provides rich, distinctive information that can be captured non-invasively in everyday settings. Facial recognition systems have therefore become an important technological component in security, access control, and large-scale user analytics, enabling automatic identification in scenarios where manual inspection is impractical. Within this broader field, facial gender classification constitutes a key subtask with broad practical relevance, supporting applications such as surveillance, targeted advertising, content personalization, customer demographic profiling, human-computer interaction, and social media analytics[1]. Traditionally, gender assessment has relied on human visual inspection of facial cues such as craniofacial structure, eyebrow morphology, lip shape, and skin texture[2], [3]. However, manual assessment is inherently limited in speed, consistency, and scalability, particularly when dealing with continuous video streams or massive image collections.

To overcome these limitations, computational image analysis techniques have been widely adopted for automatic facial recognition and classification tasks. In recent years, the development of deep learning has led to significant breakthroughs in computer vision. Convolutional Neural Networks (CNNs) have proven highly effective for image classification tasks, including face recognition and facial attribute identification[4]. Several CNN architectures, including VGGNet, ResNet, InceptionV3, EfficientNetB0, and MobileNetV2, have demonstrated strong performance in image classification tasks. Recent transfer-learning studies further highlight that InceptionV3 can achieve test accuracies around 90% with macro-F1 scores close to 0.90 for four-class facial skin-type classification, while EfficientNetB0 and MobileNetV2 also show strong performance on multi-class waste-type classification tasks with accuracies above 85%[5], [6]. However, these architectures generally involve high complexity, with millions of parameters requiring substantial computational resources. This limitation becomes critical when deploying models on devices with constrained computational power, such as smartphones, embedded systems, or real-time environments.

Previous studies have applied various feature extraction and classification methods for face recognition, including binary descriptor-based approaches and deep learning methods such as CNNs, which have shown strong performance[4]. However, these methods often face challenges related to computational complexity and reduced accuracy under varying image conditions. To address these limitations, more efficient CNN architectures, such as MobileNet, have been developed. MobileNetV2, introduced by Google in 2018, is an improved version of the original MobileNet architecture[7]. It is specifically designed for mobile and embedded vision applications by utilizing depthwise separable convolutions and inverted residual structures. MobileNetV2 is a lightweight architecture that has demonstrated competitive performance in classification and segmentation tasks, sometimes even outperforming larger models such as VGG, ResNet, and Inception, while using significantly fewer parameters and enabling faster inference on resource-limited devices[8], [9]. Despite these advantages, further investigation is still required to better understand the effectiveness of MobileNetV2 for facial gender classification in practical deployment scenarios.

Several previous studies have investigated the application of MobileNetV2 for gender classification and facial image analysis. MobileNetV2 combined with transfer learning has demonstrated competitive performance for binary gender classification [10], while comparisons with Inception-based models have confirmed the effectiveness of lightweight CNN architectures for gender prediction [7]. Other CNN-based approaches have also reported satisfactory results for facial gender recognition [11]. However, these studies mainly emphasized classification accuracy and paid limited attention to reproducible experimental protocols, prevention of data leakage, deployment-oriented preprocessing pipelines, and comprehensive evaluation under practical deployment scenarios. Although MobileNetV2 offers an attractive trade-off between computational efficiency and accuracy, its performance on large-scale heterogeneous datasets using reproducible evaluation protocols remains insufficiently investigated. Therefore, this study evaluates MobileNetV2 on a diverse public Kaggle dataset while emphasizing deployment efficiency and generalization performance.

Building upon these previous studies, the present work advances the existing literature by emphasizing not only predictive performance but also methodological rigor and deployment feasibility. Specifically, this study implements a leakage-free evaluation protocol by performing dataset splitting before data augmentation, integrates a structured preprocessing pipeline consisting of Haar Cascade face detection, normalization, and illumination-based augmentation, and adopts a two-stage MobileNetV2 transfer learning strategy designed for computationally constrained environments. Consequently, the contribution of this work extends beyond merely applying MobileNetV2. The novelty of this study lies not in the use of MobileNetV2 itself, but in the integration of a leakage-free evaluation protocol, a structured preprocessing pipeline, and a deployment-oriented transfer learning framework for lightweight facial gender classification, providing a reproducible methodology suitable for resource-constrained environments.

Beyond computational complexity, another critical challenge in facial gender classification is dataset bias arising from demographic imbalance and variations in illumination, pose, and facial appearance, which may reduce model generalization. Therefore, robust preprocessing and class-wise evaluation are required to improve model reliability under diverse real-world conditions. This study aims to evaluate the effectiveness of a MobileNetV2 transfer learning model for facial gender classification on a large public dataset while maintaining computational efficiency suitable for deployment on mobile and embedded devices.

This study makes three main contributions. First, it proposes a reproducible leakage-free experimental protocol for MobileNetV2-based facial gender classification. Second, it develops an efficient preprocessing and transfer learning framework integrating Haar Cascade face detection, illumination-based augmentation, and lightweight feature learning for deployment on resource-constrained devices. Third, it provides a comprehensive evaluation using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC to assess both predictive performance and model reliability. This manuscript presents original research conducted by the authors. To the best of the authors' knowledge, the proposed methodology, experimental design, analysis, and findings have not been published previously and are not currently under consideration for publication elsewhere.

2. Research Methodology

This section describes the main aspects related to the implementation of the research, including the research data used, the image pre-processing stages, the model training process, and the testing methods applied to evaluate model performance.

2.1 Data Collection

The dataset used in this study contains images of human faces with two gender categories, namely male and female[12]. The data were compiled and organized into separate folder structures for each category. The images were collected from a public Kaggle dataset that generally represents adult faces, but detailed age annotations (exact age ranges) are not explicitly provided by the dataset owner. Visual inspection suggests that the dataset includes a mixture of young adult to middle-aged subjects, with noticeable variation in facial appearance such as hairstyle, presence or absence of facial hair, and use of accessories (e.g., glasses or head coverings) [13]. However, precise age labels are not available, so age is treated as an uncontrolled factor in this study. Similarly, information about ethnicity was not formally labeled, based on qualitative observation, the dataset appears to contain faces from multiple ethnic backgrounds (for example, Asian and non-Asian features), but the exact proportions of each group cannot be determined reliably. Consequently, ethnicity is also considered an uncontrolled source of variation rather than an explicit variable in the analysis. In terms of image diversity, the dataset includes variations in pose (mostly frontal with slight head rotations), illumination (from relatively well-lit indoor environments to images with shadowed regions), background complexity, and image resolution. This natural variability helps approximate real-world conditions and provides a more challenging benchmark for evaluating the robustness and fairness of the proposed gender classification model[14].

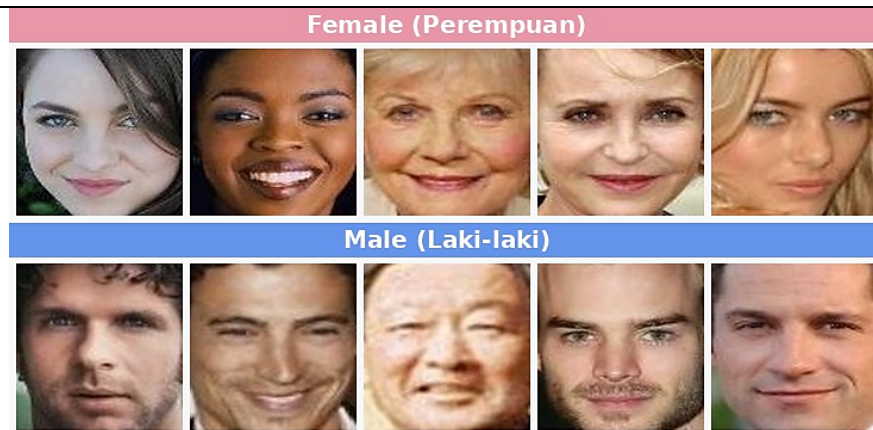


Figure 1. Sample images from the Kaggle dataset used in this study: female samples (top row) and male samples (bottom row), showing natural variation in facial appearance, skin tone, age, and illumination conditions.

The initial dataset structure consists of two main folders, namely the male folder containing 7,884 male facial images and the female folder containing 8,050 female facial images, with a total of 15,934 images. Each folder contains a number of image files in JPG format with varying dimensions in width and height. The distribution of data between the two categories is kept relatively balanced to avoid bias in the model learning process, with a ratio not exceeding 2:1 between the majority and minority categories[15]. In this particular dataset, the male and female classes are almost equal in size (7,884 vs. 8,050 samples), which substantially reduces the risk of strong class imbalance effects during training and evaluation. Given this near-balanced distribution, no additional resampling strategies (such as oversampling or undersampling) were applied. Instead, the model was trained using the original class proportions, and standard performance metrics (accuracy, precision, recall, and F1-score for each class) were reported to monitor potential residual bias. This approach allows a more realistic assessment of how the model behaves under the natural class distribution of the dataset while still providing class-wise metrics to reveal any imbalance in predictive performance, such as differences between male and female recall. Figure 1 illustrates sample images from the dataset used in this study, which was obtained from Kaggle and includes 10 example photos representing the two classes. The figure also shows variations in facial appearance, skin tone, age, and illumination, all of which the model must learn to generalize across.

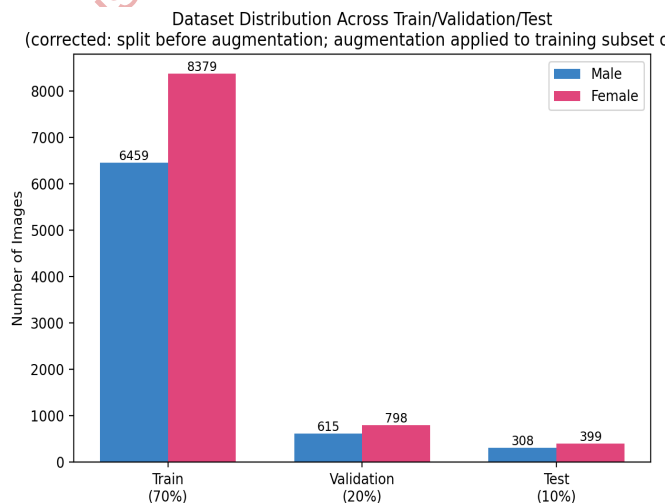


Figure 2. Overview of the dataset distribution of male and female samples across the training, validation, and testing datasets

To avoid data leakage between subsets, the dataset was first divided into training, validation, and testing subsets directly from the 7,066 face-detected, normalized images that is, before any augmentation

was applied using a ratio of 70:20:10[16] The splitting process was carried out randomly using a seed value of 42 to ensure reproducibility of the experimental results, and produced 2,153 male and 2,793 female images for the training subset, 615 male and 798 female images for the validation subset, and 308 male and 399 female images for the testing subset (4,946, 1,413, and 707 images respectively). Brightness–contrast augmentation (Section 2.2) was subsequently applied only to the training subset, expanding it threefold to 6,459 male and 8,379 female images (14,838 images in total, as illustrated in Figure 2), while the validation and testing subsets were left unaugmented so that they continue to represent genuinely unseen photographs during evaluation. Although the raw dataset was nearly balanced between classes (7,884 male vs. 8,050 female, Section 2.1), the Haar Cascade face detector retained a higher proportion of female images (3,990 of 8,050, or 49.6%) than male images (3,076 of 7,884, or 39.0%), an imbalance that is discussed in detail, together with its implications for possible selection bias, in Section 2.2. This class-wise imbalance in detection rate is taken into account in Section 3 through the reporting of class-wise precision, recall, and F1-score, rather than relying solely on overall accuracy, so that any residual asymmetry between male and female predictions remains visible. The training subset was utilized to train the model and optimize network parameters. Meanwhile, the validation subset was employed to monitor model performance during training and to select the best model checkpoint (Section 2.3), without directly influencing the learned weights and without being used to report final performance. The remaining 10% of the (unaugmented) dataset was reserved as the testing subset, which was withheld from both training and model-checkpoint selection and was used exclusively for evaluating and reporting the final performance of the trained model on previously unseen data (Section 3).

2.2 Image Pre-processing

Pre-processing the dataset is a crucial step in this research and is carried out through several sequential stages, starting from uploading the dataset, resizing images, face detection, normalization, train/validation/test splitting, and data augmentation (applied only to the training subset, as detailed below). The first stage is uploading the dataset containing facial images in ZIP format with separate folder structures for male and female categories. The dataset in ZIP format is uploaded to the Google Colab environment and extracted to produce a dataset folder containing male and female subfolders. The upload and extraction processes run smoothly, with folder structure verification showing that all image files are successfully extracted. Scanning the initial dataset produces statistics on the number of images per category and shows a relatively balanced distribution between the two classes.

The RGB images obtained from the initial dataset vary in size from hundreds to thousands of pixels in both width and height. The resizing process uses an adaptive interpolation method, namely INTER_AREA for downscaling when the image is larger than the target size and INTER_CUBIC for upscaling when the image is smaller than the target size, in order to maintain the visual quality of the image. The results of the resizing stage show that all images are successfully converted to a size of 224×224 pixels with a success rate of 100%. The output is stored in the resized_dataset folder with the same structure as the original dataset.

The third stage is face detection using the Haar Cascade Classifier to detect the face area in each resized image. Haar Cascade uses a pre-trained file provided by OpenCV with parameters scaleFactor 1.1, minNeighbors 5, and minSize 30×30 pixels[11]. The face detection process successfully detected a valid facial region in 7,066 of the 15,934 images in the dataset (3,076 male and 3,990 female images), corresponding to an overall detection rate of approximately 44.3%. This comparatively low detection rate is mainly attributable to the strict acceptance criteria applied by the Haar Cascade classifier, which rejected images with extreme head poses, partial occlusion, low resolution, or insufficient contrast between the face and background. This low and class-imbalanced detection rate (39.0% for male images vs. 49.6% for female images) carries an important methodological implication: because more than half of the male images and roughly half of the female images were discarded prior to training, the 3,076 retained male faces and 3,990 retained female faces are not necessarily representative random samples of the original 7,884 male and 8,050 female images. If the images that the Haar Cascade classifier fails to detect differ systematically from those it detects successfully (for instance, due to head pose, occlusion, image resolution, or accessories that may correlate with gender presentation), the resulting dataset may be affected by selection bias, and the reported performance may not generalize to the substantial portion of real-world facial images that this detector would fail to process. This risk is compounded by the roughly ten-percentage-point difference in detection rate between the two classes, meaning that any such bias would not necessarily be symmetric

across genders. As discussed further in Section 3.3, more robust deep-learning-based face detectors, such as Multi-task Cascaded Convolutional Networks (MTCNN), have been shown to be more resilient than classical Haar Cascade detectors to variation in pose, occlusion, and illumination; adopting MTCNN, or at minimum benchmarking Haar Cascade against it, is recommended as a way to both raise the effective detection rate and reduce the risk of class-imbalanced selection bias in future iterations of this pipeline. Images with successfully detected faces are then cropped by taking the largest face bounding box and adding a margin of 20% of the width and height of the bounding box to ensure that the surrounding area, such as hair and ears, is also captured. The cropped results are then resized back to 224×224 pixels and saved in the `face_cropped_dataset` folder. Images for which no face could be reliably detected were excluded from subsequent processing rather than retained as fallback, in order to ensure that only verified facial regions were used for training and evaluation.

The fourth step is normalizing the pixel values from the range 0–255 (uint8) to the range 0–1 (float32) by dividing each pixel value by 255.0. This normalization is important to speed up convergence and improve training stability by avoiding gradient issues that often occur with large input values. The normalization results are stored in .npy (NumPy array) file format in the `normalized_dataset` folder for efficient data loading during training. The .npy format is chosen because it allows fast storage and loading of multi-dimensional arrays without the need to decode image formats such as JPEG or PNG. Verification of the normalization results shows that all images have pixel values in the range of 0–1 with a consistent shape of $224 \times 224 \times 3$.

The fifth stage is splitting the 7,066 normalized, face-detected images into training, validation, and testing subsets, performed before any augmentation is applied (Section 2.1). Performing the split at this stage, directly on the pool of unique original images, ensures that no image derived from the same original photograph can appear in more than one subset. Using a 70:20:10 ratio and a fixed random seed (42), this produced 2,153 male and 2,793 female images for training, 615 male and 798 female images for validation, and 308 male and 399 female images for testing.

The sixth and final stage is data augmentation, applied exclusively to the training subset in order to increase the amount of training data and dataset variation without introducing any risk of near-duplicate images being shared across subsets. Given that the cropped face dataset tends to have frontal and centered poses, the augmentation techniques used are limited to brightness and contrast adjustments. Brightness adjustment is performed by multiplying pixel values by a random factor in the range of 0.8 to 1.2 with a probability of 60% to simulate variations in lighting conditions from dark to bright. Contrast adjustment is performed by scaling pixel differences relative to the mean pixel value using a random factor in the range of 0.8 to 1.2 with a probability of 50% to simulate variations in image quality. Each training image produces two augmented variants, resulting in a training subset that is three times larger than its pre-augmentation size, consisting of one original image and two augmented images per source photograph; this expands the training subset to 6,459 male and 8,379 female images (14,838 images in total). The validation and testing subsets are not augmented and retain only unaugmented original images (1,413 and 707 images respectively), so that both subsets continue to represent genuinely unseen photographs during evaluation. The augmentation results for the training subset are stored in the `augmented_dataset` folder in .npy format, with file names that distinguish between original and augmented images. The results show sufficient variation in lighting and contrast without altering the fundamental characteristics of the face. In this study, data augmentation was intentionally limited to brightness and contrast adjustments because the primary objective was to improve model robustness against illumination variations while preserving the original facial geometry. Since facial images had already been detected, cropped, and aligned during the preprocessing stage, additional geometric transformations such as rotation, horizontal flipping, scaling, and affine transformations were not applied. These transformations may alter discriminative facial characteristics or introduce unrealistic facial orientations that could affect gender classification performance. Therefore, the augmentation strategy was deliberately restricted to illumination-related variations, while the investigation of additional augmentation techniques is considered a direction for future research.

2.3 Model Training

The MobileNetV2 model was trained using a two-stage transfer learning strategy. During the first stage, all layers of the pretrained MobileNetV2 backbone were frozen, and only the newly added classification head was trained. In the second stage, the last 30 layers of the backbone were unfrozen and

fine-tuned to improve adaptation to the facial gender classification task. The hyperparameter configuration used in this study is summarized in Table 1.

Table 1. Parameters used in the MobileNetV2

Parameter	Value/Setting
Backbone Architecture	MobileNetV2(pre-trained ImageNet, includetop = False)
Input Size	224 x 224 x 3
Pooling layer	Global Average Pooling
Regularization layer	Dropout (rate = 0.3)
Output layer	Dense (2 neurons, softmax activation)
Transfer learning strategy	Stage 1: Freeze all MobileNetV2 layers (20 epochs); Stage 2: Unfreeze last 30 layers for fine-tuning (30 epochs)
Optimizer	Adam
Learning rate	0.00001
Loss function	Categorical cross-entropy
Batch size	16
Number of epochs	50 total (Stage 1: 20 epochs; Stage 2: 30 epochs)
Training data	Training set (70% of dataset)
Validation data	Validation set (20% of dataset)
Computing platform	Google Colab with GPU support

The hyperparameter configuration was determined based on established transfer learning practices reported in the literature for MobileNetV2. During the first stage, only the newly added classifier layers were optimized using the Adam optimizer with a learning rate of 0.001, enabling rapid convergence while preserving the pretrained ImageNet feature representations. In the second stage, the learning rate was reduced to 0.00001, and the last 30 layers of MobileNetV2 were fine-tuned to adapt high-level features to the facial gender classification task while minimizing the risk of catastrophic forgetting. A batch size of 16 was selected to balance GPU memory utilization and training stability, whereas a dropout rate of 0.3 was applied before the final classification layer to reduce overfitting. The training process also employed EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint callbacks to improve convergence, prevent overfitting, and retain the model with the highest validation accuracy.

Figure 3 illustrates the proposed MobileNetV2 architecture. Facial images that had undergone face detection, cropping, normalization, and brightness-contrast augmentation were resized to $224 \times 224 \times 3$ and used as inputs to the pretrained MobileNetV2 backbone with include_top=False. The extracted feature maps were aggregated using a Global Average Pooling layer, followed by a Dropout layer with a rate of 0.3. Finally, a Dense layer with two neurons and Softmax activation generated the probabilities for the male and female classes.

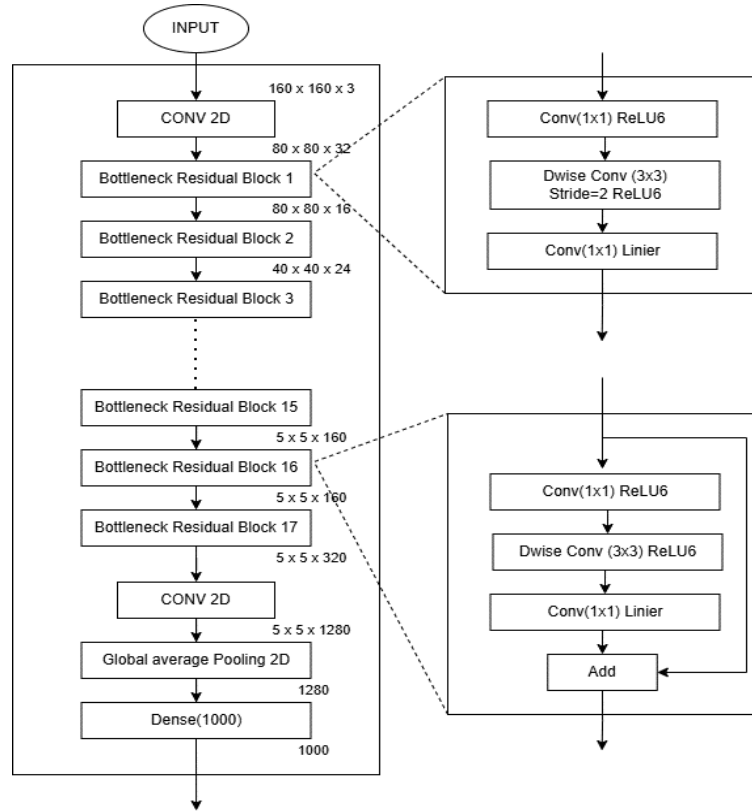


Figure 3. Architecture of MobileNetV2

Training was performed in two stages. During Stage 1, only the classification head was trained for 20 epochs while the MobileNetV2 backbone remained frozen. During Stage 2, the last 30 layers of the backbone were unfrozen and fine-tuned for up to 30 additional epochs using a lower learning rate. Model performance was monitored using the validation set, with EarlyStopping terminating training when no further improvement in validation loss was observed, ModelCheckpoint preserving the best-performing model according to validation accuracy, and ReduceLROnPlateau automatically reducing the learning rate when validation performance plateaued. The final model was subsequently evaluated on an independent test set that was excluded from both training and model selection, ensuring an unbiased assessment of generalization performance using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

2.4 Model Evaluation

Final model evaluation in this study is carried out on the test set described in Section 2.1, comprising 707 unaugmented images (308 male and 399 female) that were withheld from both model training and model-checkpoint selection. This is the split used for the final performance evaluation reported in Section 3 (confusion matrix, classification report, and ROC-AUC analysis). The validation set (1,413 images: 615 male and 798 female), by contrast, is used only during training to monitor performance across epochs and to select the best-performing checkpoint via ModelCheckpoint and EarlyStopping (Section 2.3); because it directly informs model selection, it is not used to report final performance, in order to avoid conflating model-selection data with the generalization estimate. Both the validation and test splits consist of unaugmented images of male and female faces taken from the Kaggle dataset, with a balanced class distribution for each category so that the evaluation is free from class bias. The Kaggle dataset is chosen because it is an international standard data source used by thousands of researchers globally, thus ensuring the reproducibility of the methodology and providing representative real-world conditions for testing the robustness of the system[17].

The evaluation process involves applying the trained MobileNetV2 model to all images in the test split in batches. Each evaluated image goes through the same pre-processing stages as the training data, namely resizing to 224×224 pixels, normalizing pixel values to the range of 0–1, and inputting them into the model to obtain class predictions and confidence scores. The model predictions are interpreted based

on the softmax probabilities of the final dense layer, where the class with the highest probability is selected as the gender classification result.

Model performance analysis is conducted through a comprehensive evaluation covering accuracy, recall, precision, and F1-score, which are four key metrics calculated based on the following standard formulas:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F-1\ Score = 2 \times \frac{recall \times precision}{recall + precision} \quad (4)$$

Formulas (1) to (4) provide a mathematical basis for measuring overall accuracy, which is a key indicator of the model's reliability in recognizing gender patterns from facial images. Throughout this study, the male class is designated as the positive class and the female class as the negative class (this convention is used consistently in the formulas below and in the ROC-AUC analysis in Section 3.2). Each formula quantifies a different dimension of classification effectiveness, thus forming a complete picture of the system's performance. True positive (TP) represents the model's success in accurately recognizing male samples, while true negative (TN) indicates the accuracy of female sample identification. Conversely, false positive (FP) describes errors when female samples are classified as male, and false negative (FN) occurs when male samples fail to be detected as such. The interaction between these four components reveals the dynamics of classification performance in depth.

Overall accuracy reflects the model's capability in handling the entire dataset, both for positive and negative classes. Meanwhile, recall (or sensitivity) measures the proportion of positive samples that are correctly detected, which is crucial for evaluating the system's performance against actual cases. This approach ensures that performance analysis does not focus solely on average figures but also considers the specific characteristics of image-based gender classification tasks[18].

In addition to accuracy, precision, recall, and F1-score metrics, this study also uses Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) to measure the model's ability to discriminate at various decision thresholds. The ROC curve maps the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR). The AUC value is the area under the ROC curve; AUC = 0.5 indicates performance equivalent to random, while values approaching 1.0 indicate improved class performance. In this study, the ROC-AUC was calculated on the held-out test data using the model's softmax output probability scores for each class. Consistent with the positive-class convention established in Section 2.4, the male class is treated as the positive class, so the TPR and FPR calculations refer to male predictions. The AUC value is used as a complement to the confusion matrix to see whether the classification error (e.g. low recall in the male class) is more influenced by the choice of prediction threshold or due to the limited probability score allowance between classes.

3. Results and Discussion

3.1 Model Performance Evaluation

Final evaluation of the proposed MobileNetV2-based gender classification model was conducted using the held-out test set of 707 unaugmented samples, consisting of 308 male and 399 female facial images that were not used during training or model-checkpoint selection. The evaluation results indicate that the model achieves an overall accuracy of 94.91% on this test set. These results demonstrate that the model is capable of learning highly discriminative facial features for binary gender classification under real-world variations such as lighting conditions, pose differences, and facial appearance diversity. The results of the confusion matrix can be seen in Figure 4.

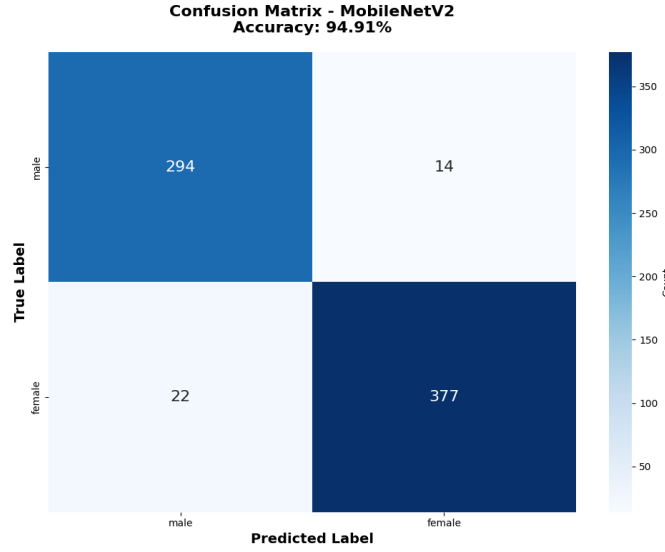


Figure 4. Confusion Matrix (test set) MobileNetV2

The confusion matrix analysis shows that the model correctly classifies the majority of female samples, while a relatively higher number of misclassifications occurs in the male class. Specifically, male images are more frequently predicted as female (false negatives), which indicates a lower sensitivity toward male facial characteristics. Conversely, false positives in the male class remain relatively limited, suggesting that the model maintains a stable decision boundary in avoiding incorrect male predictions.

This balance between classes is further reflected in the classification report on the test set, where the male class achieves a precision of 0.9304, recall of 0.9545, and F1-score of 0.9423, while the female class achieves a precision of 0.9642, recall of 0.9449, and F1-score of 0.9544. Both classes demonstrate consistently high performance with a modest asymmetry male precision is somewhat lower than male recall, mirroring the female class's slightly lower recall relative to its precision indicating that the two-stage fine-tuning strategy yields reasonably robust and balanced, though not perfect, discriminative features for both gender categories under this corrected, leakage-free evaluation.

3.2 ROC-AUC Analysis

To complement threshold-dependent evaluation metrics such as accuracy, precision, recall, and F1-score, this study employs Receiver Operating Characteristic (ROC) curve analysis and Area Under the Curve (AUC) to assess the discriminative capability of the proposed MobileNetV2 model. Unlike traditional classification metrics derived from a fixed decision threshold, ROC-AUC provides a threshold-independent evaluation that captures the model's ability to separate the two classes across all possible classification thresholds.

The ROC curve is constructed by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) using the softmax probability outputs generated by the trained model on the held-out test dataset. In this study, the male class is defined as the positive class, where TPR corresponds to the recall of male predictions, while FPR represents the proportion of female samples incorrectly classified as male. This formulation allows a more comprehensive evaluation of the ranking quality of predicted probabilities rather than relying solely on binary decision outcomes. The resulting ROC curve, as illustrated in Figure 5, shows a strong deviation from the diagonal baseline representing a random classifier. The curve consistently approaches the top-left region of the plot, indicating that the model achieves a high true positive rate while maintaining a relatively low false positive rate across multiple threshold settings.

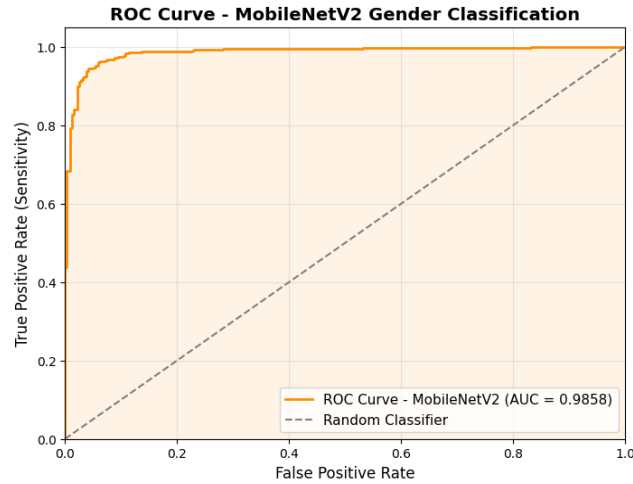


Figure 5. ROC Curve of MobileNetV2 (test set)

3.3 Comparative Analysis and Error Discussion

To provide a more quantitative comparison with previous CNN-based studies, Table 2 summarizes the classification performance, dataset characteristics, classification tasks, and reported computational cost of the compared architectures. The comparison includes InceptionV3, EfficientNetB0, and MobileNetV2 across different image classification tasks. Because the compared studies use different datasets, class distributions, and evaluation protocols, the reported accuracy values are presented as contextual comparisons rather than direct head-to-head benchmarks. The table is intended to illustrate the trade-off between classification performance and computational requirements and to position the proposed MobileNetV2 model relative to previously reported CNN architectures.

Tabel 2. Quantitative Comparison of CNN Architectures in Previous Studies and This Study

Study	Architecture	Dataset Characteristics	Classification Task	Accuracy
Listio [7]	InceptionV3	2,681 eye images; 1,251 male and 1,430 female	Binary gender classification from eye images	91.82%
Listio [7]	MobileNetV2	2,681 eye images; 1,251 male and 1,430 female	Binary gender classification from eye images	89.10%
Muttaqin and Widodo [6]	InceptionV3	1,733 facial images; 4 skin-type classes	Four-class facial skin-type classification	90.12%
Muttaqin and Widodo [6]	MobileNetV2	1,733 facial images; 4 skin-type classes	Four-class facial skin-type classification	81.15%
Thio and Susilo [5]	MobileNetV2	2,527 waste images; 6 classes; 60:20:20 split	Six-class waste classification	87.31% training; 91.57% validation
Thio and Susilo [5]	EfficientNetB0	2,527 waste images; 6 classes; 60:20:20 split	Six-class waste classification	82.46% training; 88.29% validation
This study	MobileNetV2	15,934 facial images; 7,884 male and 8,050 female; 70:20:10 split	Binary facial gender classification	94.91% test

As presented in Table 2, the proposed MobileNetV2 model achieved an accuracy of 94.91% on the held-out test set, which is higher than the accuracy reported by Listio [7] for InceptionV3 (91.82%) and

MobileNetV2 (89.10%) in gender classification using eye images. The proposed accuracy is also higher than the results reported by Muttaqin and Widodo [6], where InceptionV3 achieved 90.12% and MobileNetV2 achieved 81.15% for four-class facial skin-type classification. However, these accuracy values should not be interpreted as direct evidence of architectural superiority because the compared studies use different datasets, image characteristics, classification tasks, and evaluation protocols. Therefore, the comparison provides contextual evidence regarding the performance of the evaluated architectures rather than a controlled head-to-head benchmark.

A more direct comparison of computational efficiency can be observed in the study by Thio and Susilo [5], which evaluated MobileNetV2 and EfficientNetB0 using the same dataset and experimental setting. MobileNetV2 achieved a validation accuracy of 91.57% with a training time of 406 seconds, whereas EfficientNetB0 achieved a validation accuracy of 88.29% with a training time of 535 seconds. Thus, MobileNetV2 required 129 seconds less training time while achieving higher validation accuracy in that experimental setting. This result provides empirical support for the efficiency advantage of MobileNetV2 over EfficientNetB0 within the reported experimental conditions. However, training time is dependent on hardware, software implementation, and training configuration, and therefore should not be interpreted as a standardized measure of computational efficiency across different studies.

The observed differences in classification performance should also be considered in relation to dataset composition, preprocessing strategies, and task complexity. The studies summarized in Table 2 involve different image characteristics and classification objectives, ranging from eye-image gender classification and facial skin-type classification to waste classification. Consequently, the accuracy values cannot be used to establish that one architecture is universally superior to another. Nevertheless, the results demonstrate that MobileNetV2 can achieve competitive classification performance while retaining characteristics suitable for computationally constrained environments. Its use of depthwise separable convolutions and inverted residual structures contributes to its lightweight design and supports efficient inference on resource-constrained devices.

The error analysis indicates that misclassifications are primarily driven by visual ambiguity between male and female facial features. Male samples with long hair, smooth facial contours, or low-contrast illumination are frequently misclassified as female, while female images with strong shadows or pronounced jawline visibility tend to be misclassified as male. This behavior is consistent with findings in recent facial analysis studies showing that CNN-based models are highly sensitive to variations in illumination, pose, and intra-class appearance diversity, which can lead to unstable feature learning under real-world conditions[19]. These results suggest that the model relies heavily on global texture and low-level visual cues rather than robust semantic facial representations, particularly when discriminative facial attributes overlap between classes.

The ROC-AUC analysis reported in Section 3.2 ($AUC = 0.9858$ on the corrected, leakage-free test set) reflects the model's threshold-independent ranking quality between the two classes, with AUC providing a scalar summary of classification separability across decision thresholds. The corrected AUC of 0.9858 indicates strong class separability while remaining below perfect separation, providing a more realistic assessment of the model's discriminative capability on unseen data.

Additionally, the preprocessing pipeline using Haar Cascade face detection introduces minor inconsistencies in face cropping, which may affect the visibility of discriminative facial regions. Classical Haar Cascade detectors, while computationally efficient, are less robust under variations in pose, occlusion, and illumination compared to modern deep learning-based face detection methods[20]. More recent approaches such as Multi-task Cascaded Convolutional Networks (MTCNN) have demonstrated improved robustness by jointly performing face detection and alignment under unconstrained conditions[19]. Combined with limited augmentation strategies focused only on brightness and contrast variations, this may constrain the model's ability to generalize across broader real-world conditions.

3.4 External Image Inference Evaluation

To assess generalization beyond the Kaggle-derived validation and test splits, the trained MobileNetV2 model was additionally evaluated on a small, class-balanced set of 10 external facial photographs (5 male and 5 female) uploaded independently of the original dataset, as shown in Figure 6. Each external image was processed through the same pipeline as the training data (Haar Cascade face detection and cropping, resizing to 224×224 pixels, and pixel normalization) before being passed to the model for prediction. Face detection succeeded on all 10 external images (10/10, 100%).

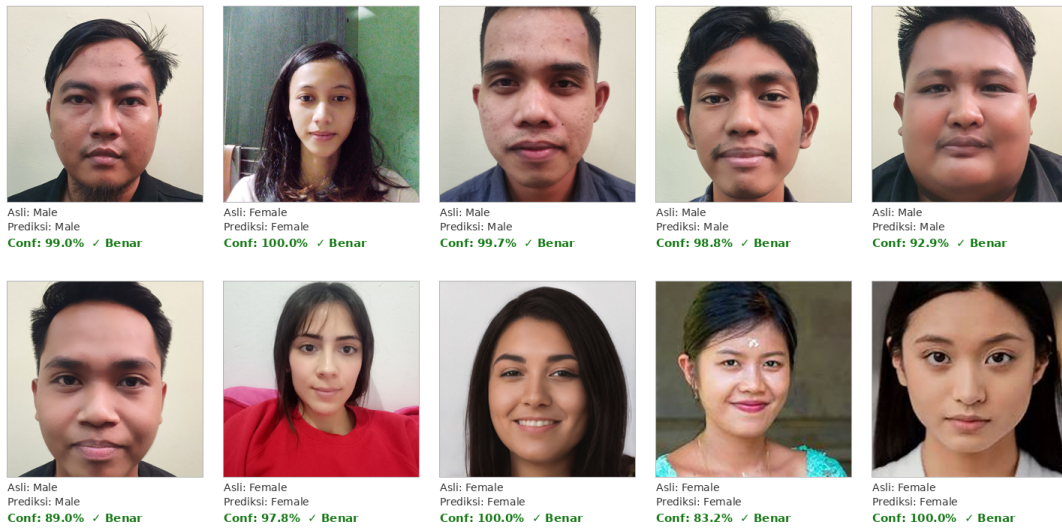


Figure 6. External test photographs with ground-truth label, model prediction, and confidence score

To further evaluate the model under practical conditions, ten external facial images that were not included in the training, validation, or testing datasets were used for qualitative evaluation. As illustrated in Figure 6, all ten images were correctly classified, resulting in an external testing accuracy of 100% for this sample. The prediction confidence ranged from 83.2% to 100%, indicating that the proposed MobileNetV2 model was able to generalize well to unseen facial images with variations in facial appearance and image quality. Although these results are encouraging, the external evaluation was conducted on a limited number of images and should therefore be interpreted as a qualitative demonstration rather than a comprehensive benchmark of real-world performance. Future studies should employ a larger and more diverse external dataset to further validate the robustness of the proposed approach under different acquisition conditions. The resulting classification report shows a precision, recall, and F1-score of 0.7500, 0.6000, and 0.6667 for the male class, and 0.6667, 0.8000, and 0.7273 for the female class. Predicted-class confidence ranged from 71.3% to 100.0% across all 10 images; notably, all three misclassifications had moderate-to-high confidence (71–80%) rather than being borderline near 50%, indicating that the model was confidently wrong on these images rather than simply uncertain. This result indicates a clear generalization gap between the in-domain, Kaggle-derived test performance (94.91% accuracy, Section 3.1) and performance on independently sourced photographs, reinforcing the conclusion (Section 4) that the corrected, leakage-free evaluation gives a more realistic and considerably more modest picture of the model's real-world generalization than the original, leakage-affected evaluation had suggested.

This result contrasts with the strong, but no longer perfect, class separability observed in the ROC-AUC analysis on the in-domain test set (Section 3.2, AUC = 0.9858), and suggests that while the preprocessing pipeline (Haar Cascade face detection, resizing, and normalization) is technically compatible with photographs captured outside the Kaggle distribution, the classifier's decision boundary does not transfer as reliably to such images. It is important to note, however, that this external evaluation set remains small ($n = 10$, 5 male and 5 female), so the resulting 70.00% accuracy should be interpreted as a preliminary, qualitative indication of a generalization gap rather than a statistically robust performance estimate on its own. A larger external evaluation set, ideally spanning more diverse ethnicities, ages, lighting conditions, and image sources, would be required to characterize the extent and causes of this gap and to draw firm conclusions about real-world deployment performance.

Although brightness and contrast augmentation was employed to improve robustness against illumination variations, the present study did not explicitly evaluate model performance under specific lighting scenarios such as low-light, backlight, normal illumination, or overexposed conditions. Therefore, the reported results should be interpreted as reflecting overall performance on the test dataset rather than robustness under controlled illumination variations. Future work will include dedicated evaluations under diverse lighting environments to further assess the model's reliability in real-world

4. Conclusion

This study implements a lightweight facial gender classification system using the MobileNetV2 architecture with transfer learning from ImageNet, achieving an accuracy of 94.91% on a held-out, unaugmented test set of 707 samples that were excluded from both training and model-checkpoint selection. The main contribution of this study lies in the design and evaluation of an efficient end-to-end pipeline that combines MobileNetV2 as a lightweight feature extractor with a structured preprocessing strategy (face detection, normalization, leakage-free train/validation/test splitting, and training-only augmentation), evaluated through a two-stage fine-tuning approach while maintaining low computational complexity suitable for real-time deployment.

The proposed approach demonstrates that a two-stage transfer learning strategy first training only the top layers on a frozen MobileNetV2 backbone, then partially fine-tuning the last 30 backbone layers can achieve meaningful discriminative performance on a relatively large and heterogeneous facial dataset. This confirms the feasibility of using lightweight transfer learning architectures as an efficient alternative to deeper CNN models for binary facial attribute classification tasks, particularly in environments where computational efficiency is a primary constraint.

In terms of performance, the model achieves a recall of 0.9545 for male and 0.9449 for female on the corrected test set, with the confusion matrix revealing 36 misclassified samples out of 707 (14 male images predicted as female, 22 female images predicted as male). These results, together with the corrected test-set accuracy of 94.91% (Section 3.1) and AUC of 0.9858 (Section 3.2), support the effectiveness of the two-stage fine-tuning strategy in learning discriminative gender features, while being substantially more modest than the 99.76% accuracy and 1.0000 AUC obtained under the original, leakage-affected evaluation consistent with the conclusion that the original figures were inflated by data leakage rather than reflecting the model's true generalization ability. This interpretation is further reinforced, rather than contradicted, by a small, qualitative check on the same 10 independently sourced photographs (5 male and 5 female) used in the original submission included here only as an illustrative signal, not as a quantitative result carrying weight comparable to the 707-image test-set evaluation above: under the corrected, leakage-free model, this set achieves only 70.00% accuracy (7 of 10 correctly classified), a marked drop from the 100% accuracy reported previously, even though face detection still succeeded on all 10 images. Given the dataset-level leakage issue identified in Section 2.1, this small external check being sourced entirely outside the Kaggle dataset—is unaffected by the internal split/augmentation correction itself; the drop in its measured accuracy (100% to 70.00%) most plausibly reflects the corrected model being less overfit to the specific Kaggle-derived training distribution, rather than any change in the external photographs themselves. As a qualitative signal alone, it is consistent with, and further supports, the conclusion that the original evaluation substantially overstated real-world generalization, though a much larger external sample would be needed before drawing quantitative conclusions from it.

Furthermore, the ROC-AUC score of 0.9858 obtained on the corrected test set (Section 3.2) reflects the class separability between male and female facial representations achieved by the partial fine-tuning of MobileNetV2 backbone layers; whether this constitutes near-perfect or more moderate separability should be interpreted directly from the updated value rather than assumed from the previous, leakage-affected result of 1.0000.

Despite these promising results, several limitations remain. First, the dataset does not include demographic annotations such as age and ethnicity, which restricts the ability to perform fairness and bias analysis. Second, the evaluation is limited to standard classification metrics and does not include fairness-oriented measures such as demographic parity or equal opportunity difference. Third, this revision corrects two methodological issues identified in the original submission: (i) augmentation was previously applied before the train/validation/test split, allowing augmented variants of the same photograph to be distributed across different subsets (data leakage), and (ii) the validation set was previously used both to select the best model checkpoint and to report final performance (double dipping). The pipeline has been corrected (Sections 2.1–2.4) so that splitting precedes augmentation, augmentation is applied only to the training subset, and the held-out test set rather than the validation set is used for final reporting; the model has since been retrained and re-evaluated under this corrected pipeline, and all accuracy, confusion-matrix, classification-report, and ROC-AUC values reported in Section 3 and Section 4 reflect this corrected evaluation. As expected under the leakage hypothesis, the corrected test-set accuracy (94.91%) and AUC (0.9858) are noticeably lower than the original, leakage-affected figures (99.76% and 1.0000), and the external evaluation accuracy dropped even more sharply (100% to 70.00%, Section 3.4), together

confirming that the original results substantially overstated the model's generalization ability. Fourth, although Section 2.2 now discusses the possibility of selection bias arising from the low and class-imbalanced Haar Cascade detection rate (44.3% overall; 39.0% male vs. 49.6% female), this study does not directly quantify that bias, for instance by comparing Haar Cascade against a more robust detector such as MTCNN on the same images; doing so is recommended as a priority for future work. Fifth, the small-scale external evaluation on 10 independently sourced photographs (Section 3.4) achieved only 70.00% accuracy under the corrected model, revealing a clear generalization gap relative to the 94.91% accuracy obtained on the in-domain Kaggle-derived test set; because this sample size ($n = 10$) is too small to be considered a statistically robust real-world deployment validation, the extent and root causes of this gap and the model's behavior under a wider range of conditions such as motion blur, camera quality variation, diverse ethnicities, and environmental noise remain largely untested and are a priority for future work.

Future work should address these limitations by incorporating partial fine-tuning of deeper MobileNetV2 layers and integrating attention mechanisms to improve focus on discriminative facial regions. In addition, future studies should consider expanding evaluation using more comprehensive metrics such as balanced accuracy, Matthews Correlation Coefficient (MCC), and fairness-aware indicators to better assess model robustness. Furthermore, comparisons with other lightweight architectures such as EfficientNet-Lite or ShuffleNet would provide a more comprehensive understanding of the trade-off between accuracy and computational efficiency. As noted above, formally retraining and evaluating the corrected, leakage-free pipeline on the held-out test set described in Section 2.4 is an immediate priority rather than a longer-term direction; beyond that, benchmarking Haar Cascade against a more robust detector such as MTCNN to quantify the selection-bias risk discussed in Section 2.2, and extending the preliminary external evaluation in Section 3.4 to a larger sample with broader diversity in ethnicity, age, and image source, are recommended next steps. Finally, validation on real-world datasets or live camera inputs is recommended to evaluate practical deployment feasibility and robustness under uncontrolled conditions.

References

- [1] Mr. P. Nagesh, "Face Recognition Systems," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 3, pp. 962–964, Mar. 2023, doi: 10.22214/ijraset.2023.49567.
- [2] A. A. Stăncioiu, A. C. Motofealea, A. Popa, R. Nagib, R. B. Lung, and C. Szuhaneck, "Advanced 3D Facial Scanning in Orthodontics: A Correlative Analysis of Craniofacial Anthropometric Parameters," *J. Clin. Med.*, vol. 14, no. 21, Nov. 2025, doi: 10.3390/jcm14217578.
- [3] Z. S. A. Ghrban and N. K. EL Abbadi, "Gender Classification from Face and Eyes Images Using Deep Learning Algorithm," *Journal of Computer Science*, vol. 19, no. 3, pp. 345–362, 2023, doi: 10.3844/JCSP.2023.345.362.
- [4] P. A. Javier Orlando, J. M. Robinson, and J. E. M. Baquero, "Comparison of convolutional neural network models for user's facial recognition," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 192–198, Feb. 2024, doi: 10.11591/ijece.v14i1.pp192-198.
- [5] S. E. Thio and J. Susilo, "Identifikasi Pemilahan Sampah Berbasis Algoritma Transfer Learning CNN Menggunakan MobileNetV2 dan EfficientNetB0," *bit-Tech*, vol. 8, no. 1, pp. 25–32, Aug. 2025, doi: 10.32877/bt.v8i1.1900.
- [6] N. H. Muttaqin and A. M. Widodo, "Evaluation of Transfer Learning-Based Convolutional Neural Networks (InceptionV3 and MobileNetV2) for Facial Skin-Type Classification," *Jurnal Ilmu Komputer dan Informatika*, vol. 5, no. 1, pp. 11–32, Aug. 2025, doi: 10.54082/jiki.264.
- [7] S. W. P. Listio, "Performance of Deep Learning Inception Model and MobileNet Model on Gender Prediction Through Eye Image," *Sinkron*, vol. 7, no. 4, pp. 2593–2601, Nov. 2022, doi: 10.33395/sinkron.v7i4.11887.
- [8] Y. Gulzar, "Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique," *Sustainability (Switzerland)*, vol. 15, no. 3, Feb. 2023, doi: 10.3390/su15031906.
- [9] Z. Wu, Y. Tang, B. Hong, B. Liang, and Y. Liu, "Enhanced Precision in Dam Crack Width Measurement: Leveraging Advanced Lightweight Network Identification for Pixel-Level Accuracy," *International Journal of Intelligent Systems*, vol. 2023, 2023, doi: 10.1155/2023/9940881.
- [10] N. R. Hamza, "Gender Classification from Human Face Images Using Deep Learning Based on MobileNetV2 Architecture," *Journal of Al-Qadisiyah for Computer Science and Mathematics*, vol. 17, no. 1, pp. 145–156, Mar. 2025, doi: 10.29304/jqcs.2025.17.11970.
- [11] N. Chalista Imanuela Natun, M. Angelica Santhia, and Y. R. Kaesmetan, "Identifikasi Pengenalan Wajah Berdasarkan Jenis Kelamin Menggunakan Metode Convolutional Neural Network (CNN)," *Journal of Technology and Informatics (JoTI)*, vol. 6, no. 1, pp. 50–57, Oct. 2024, doi: 10.37802/joti.v6i1.694.

-
- [12] Y. F. Riti, R. P. Kristianto, and D. R. A. Setiawan, "PERFORM COMPARATION OF DEEP LEARNING METHODS IN GENDER CLASSIFICATION FROM FACIAL IMAGES," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 4, pp. 926–936, Jun. 2025, doi: 10.33480/jitk.v10i4.4717.
- [13] K. K. Yadav and A. Gajanan, "Transfer Learning for Image Processing: A Review and Practical Considerations," *Asian Journal of Research in Computer Science*, vol. 19, no. 2, pp. 192–203, Mar. 2026, doi: 10.9734/ajrcos/2026/v19i2831.
- [14] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," Mar. 01, 2023, *MDPI*. doi: 10.3390/computation11030052.
- [15] P. Jyothi, D. Haritha, and K. Arava, "Issue 27s (2025) Ponukumati Jyothi, Dr. Dasari Haritha, Dr. Karuna Arava, (2025) Gender Classification Using Efficientnetb0 And Mobilenetv2." [Online]. Available: <https://www.jneonatsurg.com>
- [16] D. Dhimas, P. Putra, G. Kurnia Anaga, and W. T. Fitriyana, "Implementasi Algoritma Convolutional Neural Network Arsitektur Mobilenetv2 Untuk Klasifikasi Ekspresi Wajah Pada Dataset FER," 2024.
- [17] P. Wu *et al.*, "A Review on Research and Application of AI-Based Image Analysis in the Field of Computer Vision," 2025, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2025.3565300.
- [18] S. Mandasari, D. Irfan, W. Wanayumini, and R. Rosnelly, "Comparison of SGD, Adadelta, Adam Optimization In Gender Classification Using CNN," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 9, no. 3, pp. 345–354, Jun. 2023, doi: 10.33330/jurteks.v9i3.2067.
- [19] S. P. Samadhi and E. Izquierdo, "Deep-learned faces: a survey," Dec. 01, 2020, *Springer*. doi: 10.1186/s13640-020-00510-w.
- [20] I. Rodriguez-Conde, C. Campos, and F. Fdez-Riverola, "Optimized convolutional neural network architectures for efficient on-device vision-based object detection," *Neural Comput. Appl.*, vol. 34, no. 13, pp. 10469–10501, Jul. 2022, doi: 10.1007/s00521-021-06830-w.