Explainable Deep Learning for Age and Gender Prediction from Facial Images: A Comparative Study of VGG16, Resnet50, and Efficientnet with Grad-CAM and SHAP

Yassir Elhaj

School of Computer Science Nanjing University of Information Science and Technology Nanjing, China

DOI: https://doi.org/10.47772/IJRISS.2026.100400063

Received: 30 March 2026; 06 April 2026; Published: 27 April 2026

ABSTRACT

Automatic age estimation and gender classification from facial images represent two of the most intensively studied problems in computer vision, with wide-ranging applications in human-computer interaction, biometric surveillance, targeted marketing, healthcare monitoring, and forensic analysis. Despite remarkable advances in convolutional neural network architectures over the past decade, the black-box nature of deep learning models continues to pose significant challenges in terms of interpretability, trustworthiness, and accountability, particularly in sensitive deployment contexts. This paper presents a comprehensive comparative study of three state-of-the-art deep learning architectures—VGG16, ResNet50, and EfficientNet-B3—for simultaneous age and gender prediction from facial images, with a strong emphasis on model explainability. Our framework employs the UTKFace dataset, comprising over 20,000 face images spanning ages from 1 to 116 across multiple ethnicities. We describe a rigorous preprocessing pipeline incorporating Multitask Cascaded Convolutional Networks (MTCNN) for face detection and alignment, followed by standardized normalization and extensive data augmentation strategies. Both Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) are integrated into the evaluation workflow to provide visual and quantitative insight into the regions and features that drive model decisions. Experimental results demonstrate that EfficientNet-B3 achieves superior performance with a Mean Absolute Error (MAE) of 4.37 years for age estimation and a gender classification accuracy of 96.8%, while maintaining a significantly reduced computational footprint compared to the other architectures under evaluation. ResNet50 offers a strong middle ground between accuracy and training efficiency, whereas VGG16, though interpretable, lags behind in both performance and computational cost. Our explainability analysis reveals that all three models predominantly attend to periocular regions, nasolabial folds, and frontal skull geometry for age estimation, while gender classification relies more heavily on jaw contour, brow ridge prominence, and lip morphology. These findings underscore the importance of integrating explainability tools into the facial analysis pipeline and provide practical guidance for practitioners deploying deep learning systems in real-world, ethically sensitive environments.

Index Terms— Age Estimation, Gender Classification, Deep Learning, Explainable AI, Grad-CAM, SHAP, VGG16, ResNet50, EfficientNet, UTKFace, Facial Analysis, Convolutional Neural Networks

INTRODUCTION

The human face is a highly information-rich visual structure that conveys essential demographic attributes such as age and gender. Humans can interpret these attributes rapidly and reliably, even under varying conditions. Replicating this capability in automated systems has been a longstanding objective in computer vision and pattern recognition. In particular, age and gender estimation from facial images have attracted significant attention due to their wide range of practical applications, including surveillance systems, personalized recommendation platforms, healthcare monitoring, and forensic analysis [1]–[3].

Traditional approaches to facial demographic estimation relied on hand-crafted features combined with classical machine learning techniques. While these methods achieved reasonable performance under controlled

conditions, they struggled to generalize to real-world environments characterized by variations in pose, illumination, occlusion, and demographic diversity. The emergence of deep learning has fundamentally transformed this field, enabling models to learn hierarchical feature representations directly from data [4]. In particular, deep convolutional neural networks (CNNs) such as ResNet have demonstrated strong performance in visual recognition tasks by addressing optimization challenges in very deep architectures [5]. More recent architectures, including EfficientNet, further improve the trade-off between accuracy and computational efficiency, making them suitable for scalable and real-time applications [6], [7].

Despite these advances, the deployment of deep learning models in real-world and socially sensitive contexts has raised important concerns regarding interpretability, fairness, and accountability [8], [9]. High-performing models often behave as black-box systems, making it difficult to understand the reasoning behind their predictions. This lack of transparency is particularly problematic in applications where biased or unfair decisions can have significant societal implications. Recent studies highlight that demographic estimation systems may exhibit performance disparities across different population groups, emphasizing the need for more transparent and reliable models [10], [11].

To address these challenges, Explainable Artificial Intelligence (XAI) has emerged as a key research direction aimed at improving the transparency and interpretability of machine learning models [12], [13]. Among the most widely used techniques, Gradient-weighted Class Activation Mapping (Grad-CAM) provides visual explanations by highlighting image regions that contribute most to model predictions [14], while SHapley Additive exPlanations (SHAP) offer a theoretically grounded approach for quantifying feature contributions based on cooperative game theory [15]. Recent advances emphasize that combining multiple explainability techniques can yield more robust and reliable insights into model behavior [9], [16].

In parallel, recent surveys confirm the rapid evolution of deep learning-based facial analysis systems, highlighting ongoing challenges related to robustness, scalability, and fairness [17], [18]. Furthermore, the integration of explainability with efficient model architectures and real-time deployment capabilities is increasingly recognized as essential for practical AI systems operating in real-world environments.

Motivated by these challenges, this work proposes a unified and explainable deep learning framework for joint age and gender estimation. The proposed approach combines modern convolutional architectures with complementary explainability techniques, enabling both high predictive performance and interpretable decision-making. In addition, the framework is designed with deployment considerations in mind, bridging the gap between theoretical model performance and practical applicability.

This paper makes the following principal contributions:

- We propose a unified and rigorous multi-task deep learning framework for simultaneous age estimation and gender classification, leveraging well-established convolutional architectures—VGG16 [19], ResNet50 [5], and EfficientNet-B3 [6]—to systematically analyze performance trade-offs between accuracy, efficiency, and model complexity.
- We design a comprehensive preprocessing pipeline based on MTCNN [20] for robust face detection, alignment, and normalization. This pipeline is complemented by an extensive data augmentation strategy to enhance generalization under unconstrained, real-world imaging conditions, improving robustness to variations in pose, illumination, and occlusion.
- We integrate explainability directly into the evaluation process by combining Grad-CAM [14] and SHAP [15], enabling both spatial and feature-level interpretation of model predictions. This allows for a systematic comparison of attention patterns across architectures, providing deeper insights into model behavior.
- We conduct a comprehensive comparative analysis across multiple dimensions, including predictive performance (MAE, accuracy, F1-score), computational efficiency (training time, parameter count, FLOPs), and explanation quality. This multi-dimensional evaluation framework goes beyond conventional accuracy-based comparisons and provides a more holistic assessment of model suitability.

- We investigate the ethical and practical implications of deploying facial demographic estimation systems, with a particular focus on fairness, bias, and model transparency. We further demonstrate how explainability techniques can be leveraged to identify potential biases and support more accountable AI systems [9], [16].
- We extend the proposed framework into a real-time web-based system using FastAPI and Django, demonstrating its practical applicability with interactive inference and integrated explainability visualization, thereby bridging the gap between research and real-world deployment.

The remainder of this paper is organized as follows. Section II reviews the relevant literature on age and gender estimation and explainable AI for facial analysis. Section III describes the proposed methodology, including dataset, preprocessing pipeline, and model architectures. Section IV details the experimental configuration. Section V presents and analyzes the experimental results. Section VI provides a structured cross-model comparison. Section VII discusses the broader implications of the findings. Finally, Section VIII concludes the paper and outlines directions for future research.

Related Work

Age Estimation from Facial Images

Age estimation from facial images has undergone significant advancements with the rapid development of deep learning techniques. Recent surveys highlight substantial progress achieved through data-driven approaches and deep neural networks, enabling more accurate and robust age prediction systems [2], [3], [21]. The availability of large-scale datasets, combined with increasingly powerful architectures, has allowed models to learn complex and non-linear facial aging patterns under unconstrained conditions, making them suitable for real-world applications.

Deep convolutional neural networks (CNNs) have demonstrated strong performance in age estimation tasks by learning hierarchical feature representations that capture both local and global facial characteristics [4], [5]. More advanced architectures such as EfficientNet further improve the trade-off between predictive accuracy and computational efficiency, making them particularly suitable for scalable and deployment-oriented systems [6], [7]. In addition, lightweight CNN variants have been proposed to handle noisy or unconstrained facial data while maintaining robust performance, especially in real-world scenarios [22].

Attention-based mechanisms have also been widely explored to focus on age-relevant facial regions, such as wrinkles, skin texture, and facial contours, thereby improving both interpretability and predictive performance [23]. These mechanisms enable models to selectively emphasize informative regions while suppressing irrelevant background features, which is particularly important in unconstrained imaging environments.

More recently, transformer-based architectures have been introduced for visual recognition tasks, offering competitive performance and enhanced interpretability through self-attention mechanisms [24]. Compared to traditional CNNs, transformers are capable of modeling long-range dependencies across facial regions, which is beneficial for capturing global aging patterns and contextual relationships. Recent studies further explore hybrid CNN-transformer models for improved facial analysis performance [21].

Furthermore, recent research emphasizes the importance of combining multiple learning strategies—including data augmentation, transfer learning, and hybrid architectures—to improve robustness and generalization in age estimation systems [25]. These approaches contribute to building more reliable models that can effectively handle diverse and real-world data distributions.

Gender Classification from Facial Images

Gender classification from facial images has achieved high accuracy with the adoption of deep learning methods, particularly when trained on large-scale datasets [1]. Modern approaches leverage deep convolutional neural

networks to extract discriminative facial features, enabling robust performance under challenging conditions such as variations in pose, illumination, and occlusion.

Multi-task learning frameworks have been widely explored to jointly predict gender alongside other facial attributes such as age and expression, improving overall performance through shared representations [1], [18]. These approaches enable models to exploit correlations between related tasks while maintaining task-specific learning capabilities, leading to improved generalization and reduced overfitting.

Despite their high accuracy, gender classification systems have been shown to suffer from bias across demographic groups. Recent studies highlight disparities in performance across gender and ethnicity, emphasizing the need for fairness-aware approaches in facial analysis systems [10], [11], [26]. These biases are often attributed to imbalanced training data and limitations in model generalization across diverse populations.

Recent research has focused on mitigating these biases through various strategies, including improved data sampling, fairness-aware loss functions, and regularization techniques designed to promote equitable performance across demographic groups [11]. In addition, the integration of explainability methods has emerged as an effective tool for identifying and diagnosing biased model behavior, enabling more transparent and accountable AI systems [9], [16].

Multi-task Learning for Joint Age and Gender Estimation

Joint estimation of age and gender using multi-task learning frameworks has been shown to outperform independent single-task models by leveraging shared representations across related tasks [1], [18]. In such architectures, a shared convolutional backbone captures common low-level features—such as edges, textures, and spatial patterns—while task-specific branches learn higher-level discriminative representations tailored to each prediction objective.

This shared learning paradigm improves generalization by exploiting the inherent correlation between age and gender attributes, reducing overfitting and enhancing model efficiency. By learning multiple tasks simultaneously, the model benefits from additional supervisory signals, leading to more robust and transferable feature representations compared to isolated learning approaches.

Recent studies further demonstrate that multi-task learning frameworks can effectively balance performance across tasks while maintaining computational efficiency, making them particularly suitable for real-world applications where multiple facial attributes must be predicted simultaneously under resource constraints [1]. In addition, these approaches contribute to reducing model redundancy by sharing parameters across tasks, thereby improving scalability.

More recently, transformer-based architectures have been introduced into multi-task learning settings, enabling models to capture complex relationships between different facial attributes through self-attention mechanisms [24]. These models are particularly effective in modeling long-range dependencies and disentangling task-relevant features in the latent space, leading to improved predictive performance and interpretability. Recent work also explores hybrid CNN-transformer architectures to further enhance multi-task learning performance [21].

In addition, the integration of explainability techniques within multi-task frameworks has gained increasing attention, allowing researchers to analyze how different tasks influence shared representations and to identify potential biases or conflicts between tasks [9], [16]. This combination of multi-task learning and explainable AI represents a promising direction for developing transparent, efficient, and reliable facial analysis systems.

Explainable AI for Facial Analysis

The application of Explainable Artificial Intelligence (XAI) techniques to facial analysis models has grown substantially in recent years. Gradient-weighted Class Activation Mapping (Grad-CAM) [14] was originally

proposed as a generalization of Class Activation Mapping (CAM) [27], enabling architecture-independent visual explanations for convolutional neural networks. Recent variants, such as Score-CAM [28], further improve the quality and reliability of visual explanations by reducing gradient-related noise.

SHapley Additive exPlanations (SHAP) [15] provide a unified framework for interpreting model predictions based on cooperative game theory. Recent advances have improved its scalability and applicability in deep learning contexts [12], while recent surveys highlight the growing importance of explainability in computer vision systems [8], [9]. These methods offer complementary perspectives: Grad-CAM provides spatial localization of important regions, while SHAP quantifies feature-level contributions.

In addition to these approaches, other attribution methods such as Integrated Gradients have been proposed to provide theoretically grounded explanations that satisfy desirable axiomatic properties. Recent studies emphasize that different explainability techniques often highlight partially overlapping but not identical regions, suggesting that combining multiple methods can yield more robust and reliable insights into model behavior [9], [29].

Recent work has also explored the application of XAI techniques in biometric and facial analysis systems. For instance, Huber et al. [30] analyzed the relationship between human-perceived facial attributes and model explanations, highlighting both the potential and limitations of current XAI methods. Furthermore, recent research suggests that higher-quality explanations may be associated with improved model robustness and generalization, reinforcing the importance of integrating explainability into model design and evaluation [9].

Transfer Learning in Facial Analysis

Transfer learning from large-scale visual recognition datasets, particularly ImageNet, has become a standard approach in facial analysis tasks [4], [5]. The hierarchical feature representations learned during large-scale pre-training—from low-level edge and texture detectors to high-level semantic features—provide an effective initialization for domain-specific fine-tuning, significantly improving performance and convergence speed.

Comparative studies of different backbone architectures have consistently highlighted a trade-off between model capacity and computational efficiency. Modern architectures such as EfficientNet achieve a favorable balance between accuracy and efficiency, making them particularly suitable for transfer learning in real-world and resource-constrained applications [6], [7]. These models enable scalable deployment while maintaining high predictive performance.

More recently, self-supervised and representation learning approaches have gained significant attention, demonstrating that powerful visual features can be learned without relying on large labeled datasets [21]. These methods are particularly promising for facial analysis tasks where labeled data may be limited or costly to obtain. In addition, recent studies emphasize the importance of combining transfer learning with data augmentation and regularization strategies to improve generalization in unconstrained environments [25], [31].

Overall, transfer learning remains a fundamental component of modern facial analysis systems, enabling efficient model training, improved generalization, and practical deployment across diverse application scenarios.

METHODOLOGY

Dataset: UTKFACE

The UTKFace dataset serves as the primary benchmark for our experiments and is widely used in facial analysis research due to its diversity and real-world variability [1], [2]. It is a large-scale, in-the-wild face dataset comprising approximately 23,705 images spanning a wide age range from 1 to 116 years. Each image is annotated with three attributes: age (integer years), gender (binary: male/female), and ethnicity (five categories: White, Black, Asian, Indian, and Others). The images are collected from unconstrained sources such as social

media platforms and public repositories, resulting in substantial variation in pose, illumination, facial expression, background clutter, image resolution, and sensor characteristics.

For our experiments, the dataset is partitioned using a stratified random split that preserves both age and gender distributions across subsets: 70% for training (16,593 images), 15% for validation (3,556 images), and 15% for testing (3,556 images). Stratification is applied jointly on binned age groups (with intervals of 10 years) and gender, ensuring that underrepresented age ranges are consistently included in all subsets. Additionally, images with ambiguous or corrupted labels (approximately 180 samples) are removed to improve data quality. All images are resized to a uniform resolution of 224 × 224 pixels to ensure compatibility with the selected deep learning architectures.

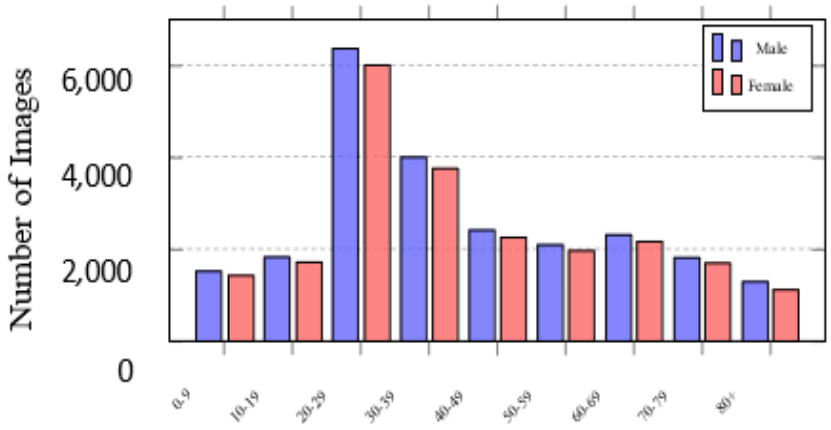

Fig. 1. Age and gender distribution of the UTKFace dataset. The dataset exhibits significant class imbalance, with a concentration *in the 20–29 age range, motivating the use of sample weighting during training.*

A notable characteristic of UTKFace is its pronounced class imbalance: the 20–29 age group constitutes approximately 27% of the dataset, while the extreme age ranges (e.g., 0–4 and 90+) are each represented by fewer than 200 images. This imbalance, illustrated in Fig. 1, can lead to biased models that perform well on dominant age groups but poorly on underrepresented ones.

To mitigate this issue, we adopt an inverse-frequency sample weighting strategy during training, where each sample is assigned a weight inversely proportional to the frequency of its corresponding age group. This approach encourages the model to pay greater attention to minority classes and improves overall generalization performance across the full age spectrum.

Preprocessing Pipeline

Face Detection and Alignment with MTCNN: Although images in UTKFace generally contain a single face, variations in background, occlusion, and pose can negatively impact feature extraction. To address this, we employ the Multitask Cascaded Convolutional Network (MTCNN) [20] to detect facial bounding boxes and five key landmarks (left eye, right eye, nose tip, left mouth corner, and right mouth corner).

MTCNN consists of a three-stage cascade—Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net)—operating at multiple image scales to ensure robust face detection under diverse conditions. Detected faces are geometrically aligned using landmark coordinates to enforce a canonical pose. Specifically, a similarity transformation (rotation, scaling, and translation) is applied to map the eye centers to predefined positions in the output space.

Following alignment, faces are cropped and resized to 224 × 224 pixels. For approximately 3.2% of images where face detection fails, a fallback strategy is applied using center-cropping (90% of the shorter image dimension) followed by resizing. This ensures consistency across all input samples while minimizing data loss.

Pixel Normalization: Aligned face images are normalized using ImageNet channel-wise statistics, with mean $\mu = (0.485, 0.456, 0.406)$ and standard deviation $\sigma = (0.229, 0.224, 0.225)$. This normalization is consistent with the pre-training conditions of VGG16, ResNet50, and EfficientNet-B3, ensuring that input distributions during fine-tuning remain compatible with the learned feature representations [6], [7]. This step improves training stability and accelerates convergence.

Data Augmentation: To improve model robustness and mitigate overfitting, we apply an online data augmentation pipeline during training. Augmentations are applied stochastically on a per-sample basis as follows:

- **Random horizontal flip:** applied with probability 0.5, leveraging the approximate bilateral symmetry of facial structures.
- **Random rotation:** uniformly sampled from $[-15°, +15°]$ with probability 0.6, simulating natural head pose variations.
- **Random brightness and contrast jitter:** brightness and contrast factors sampled from [0.8, 1.2], applied jointly with probability 0.5 to simulate lighting variability.
- **Random Gaussian blur:** kernel size 3×3 with $\sigma \in [0.1, 2.0]$, applied with probability 0.3 to mimic optical defocus and sensor noise.
- **Random erasing:** randomly removes a rectangular region (area ratio in [0.02, 0.33]) and replaces it with noise, improving robustness to occlusions [32].
- **Mixup augmentation:** linearly combines pairs of training samples and labels using a mixing coefficient $\lambda \sim \text{Beta}(0.2, 0.2)$, improving generalization and regularization [33].

No augmentation is applied during validation and testing; only normalization is retained to ensure consistent evaluation.

Model Architectures

VGG16: VGG16 [19] is a foundational deep convolutional architecture characterized by its simple and uniform design. It consists of 13 convolutional layers organized into five blocks with increasing filter depths (64, 128, 256, 512, and 512 channels), each followed by ReLU activation and max-pooling layers for spatial downsampling. The convolutional backbone is followed by three fully connected layers (4096, 4096, and N units respectively, where N denotes the number of output classes). The model contains approximately 138 million parameters, with the majority concentrated in the fully connected layers.

Despite its age, VGG16 remains widely used as a strong baseline and feature extractor due to the high transferability of its learned representations. For our multi-task framework, we replace the original classification head with two parallel branches: (1) an age regression head composed of two fully connected layers (512 units with ReLU activation followed by a single linear output), and (2) a gender classification head consisting of two fully connected layers (256 units with ReLU followed by a 2-unit softmax output). A dropout rate of 0.5 is applied prior to each task head to reduce overfitting. The convolutional backbone is initialized with ImageNet pre-trained weights, and the network is fine-tuned end-to-end using differential learning rates.

ResNet50: ResNet50 [5] introduced the residual learning framework, which addresses the vanishing gradient problem in deep networks by incorporating identity shortcut connections that facilitate gradient flow. The architecture consists of four stages of bottleneck residual blocks with filter dimensions of 64, 128, 256, and 512, respectively. Each block includes a 1×1 convolution for dimensionality reduction, a 3×3 convolution for feature extraction, and a 1×1 convolution for dimensionality expansion, with a residual connection bypassing these transformations.

ResNet50 contains approximately 25.6 million parameters and provides an effective balance between representational capacity and computational efficiency, making it well-suited for facial analysis tasks [1]. For multi-task adaptation, we replace the original fully connected layer with the same dual-head structure used in VGG16. An additional advantage of ResNet50 is its global average pooling layer, which allows Grad-CAM [14] to be applied directly to the final convolutional features, producing high-quality localization maps without architectural modification.

EfficientNet-B3: EfficientNet [6] is based on a compound scaling method that systematically scales network depth, width, and input resolution to optimize performance. EfficientNet-B3, used in this work, contains approximately 12 million parameters and achieves superior accuracy-to-computation efficiency compared to traditional architectures.

The architecture employs Mobile Inverted Bottleneck Convolution (MBConv) blocks combined with squeeze-and-excitation (SE) mechanisms [34], which adaptively recalibrate channel-wise feature responses to improve representational power. Although EfficientNet-B3 is originally designed for higher input resolutions, we adopt an input size of 224 × 224 to maintain consistency with our preprocessing pipeline.

For multi-task learning, we attach the same dual-head structure to the output of the global average pooling layer, enabling simultaneous age regression and gender classification. This design ensures consistency across architectures and allows for a fair comparative evaluation of performance, efficiency, and interpretability.

Loss Functions and Optimization

The proposed multi-task framework is trained using a joint objective function that combines the losses from both tasks:

$$\mathcal{L} * \text{total} = \alpha \cdot \mathcal{L} * \text{age} + \beta \cdot \mathcal{L} * \text{gender} \qquad (1)$$

where α and β are task weighting hyperparameters that control the relative importance of age estimation and gender classification during training.

For age estimation, we adopt the Mean Absolute Error (MAE) loss:

$$\mathcal{L} * \text{age} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{\text{age}} - \hat{y}_i^{\text{age}}\right| \qquad (2)$$

MAE is preferred over Mean Squared Error (MSE) due to its robustness to outliers and its direct alignment with the evaluation metric commonly used in age estimation tasks [2].

For gender classification, we use the categorical cross-entropy loss:

$$\mathcal{L} * \text{gender} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\{M,F\}} y_{i,c}^{\text{gender}}\, log\hat{p}_{i,c}^{\text{gender}} \qquad (3)$$

This loss function is widely used for classification tasks and provides stable gradient signals for optimizing probabilistic outputs.

We set α = 1.0 and β = 1.5 based on empirical tuning via grid search on the validation set. This configuration slightly prioritizes the gender classification task, resulting in improved convergence of the joint objective while maintaining strong age estimation performance.

All models are optimized using the Adam optimizer [35], which remains a standard and effective choice for deep learning optimization due to its adaptive learning rate mechanism. We employ differential learning rates, setting 1×10^{-4} for the pre-trained backbone and 1×10^{-3} for the task-specific heads to stabilize fine-tuning.

A cosine annealing learning rate schedule with warm restarts is applied to improve convergence and avoid local minima [36], [37]. The initial cycle length is set to 10 epochs with a cycle multiplier of 2. Weight decay of 5×10^{-4} is used as L_2 regularization to reduce overfitting.

Training is performed for a maximum of 80 epochs, with early stopping based on validation MAE and a patience of 15 epochs. This prevents overfitting while ensuring sufficient training for convergence. The overall optimization strategy is designed to balance stability, convergence speed, and generalization performance.

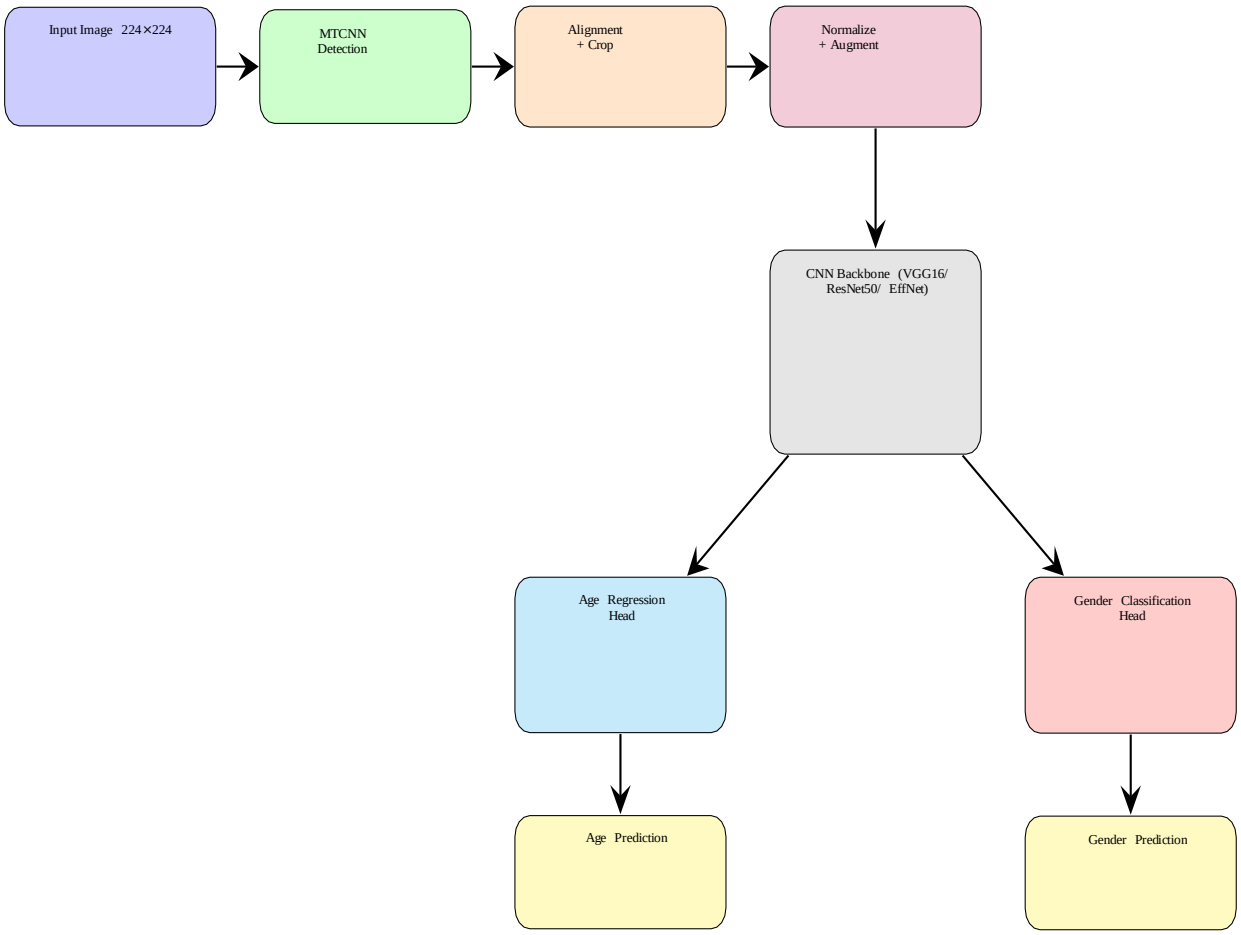

Fig. 2. End-to-end training pipeline. Raw images pass through MTCNN-based face detection and alignment, pixel normalization, and stochastic augmentation before being fed to the shared CNN backbone. Task-specific heads produce simultaneous age regression and gender classification outputs.

Explainability Methods

1) Gradient-weighted Class Activation Mapping (Grad- CAM): Grad-CAM [14] generates class-discriminative localization maps by computing the gradient of a target output with respect to the feature maps of the final convolutional layer. The importance weights for each feature map A_k are computed as :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}} \qquad (4)$$

where Z = H × W denotes the spatial dimensions of the feature map, and (i, j) indexes spatial locations. These weights are used to compute a weighted combination of feature maps, followed by a ReLU activation :

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A_k\right) \qquad (5)$$

The resulting heatmap is upsampled to the input image resolution using bilinear interpolation and overlaid on the original image for visualization.

For age estimation (a regression task), Grad-CAM is computed with respect to the scalar age prediction, treating it analogously to a single-output target. For gender classification, Grad-CAM is computed with respect to the predicted class score. A key advantage of Grad-CAM in our multi-task framework is that it can be applied independently to each task head without modifying the shared backbone, enabling direct comparison of attention regions for age and gender predictions.

2) SHapley Additive Explanations (SHAP): SHAP [15] is based on cooperative game theory and provides a unified framework for interpreting model predictions. The SHAP value φ_i for a feature i represents its average marginal contribution to the model output :

$$\varphi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup i) - f(S)] \qquad (6)$$

where F is the set of all input features and f(S) denotes the model output conditioned on subset S. Due to the computational complexity of exact Shapley value estimation, we employ the DeepSHAP approximation, which propagates feature attributions through the network using a modified backpropagation scheme.

A background dataset of 200 randomly sampled training images is used as the reference distribution for estimating conditional expectations. SHAP values are computed at the pixel level and subsequently aggregated into superpixel regions using SLIC segmentation [35] to improve interpretability and visualization.

The resulting attribution maps provide signed, quantitative explanations: positive SHAP values indicate features that increase the prediction (e.g., older age or male classification), while negative values indicate features that decrease it. This complementary property enhances interpretability when combined with Grad-CAM, which provides spatial but unsigned importance maps.

The integration of Grad-CAM and SHAP enables a multi-perspective explanation framework, combining spatial localization with quantitative attribution, thereby offering deeper insights into model behavior and decision-making processes.

Real-Time Deployment System

To demonstrate the practical applicability of the proposed framework beyond offline experimentation, we develop a real-time web-based system that enables interactive age and gender prediction from facial images. The system integrates trained deep learning models into a scalable client-server architecture using FastAPI for backend inference and Django for frontend interaction. This design reflects recent trends toward deploying deep learning models in real-world, user-facing applications [37].

System Architecture

The deployment architecture follows a modular design composed of three main components: (1) a frontend user interface, (2) a backend inference server, and (3) the trained deep learning models. The frontend, implemented using Django, provides users with an intuitive interface for uploading images and visualizing prediction results. The backend, built with FastAPI, exposes RESTful API endpoints responsible for image processing, model inference, and response generation.

Upon receiving an input image, the backend applies the same preprocessing pipeline described in Section III, including face detection and alignment using MTCNN, resizing, and normalization. The processed image is then passed to the selected model (VGG16, ResNet50, or EfficientNet-B3), which simultaneously produces age estimation and gender classification outputs.

API Workflow

The system exposes a lightweight API endpoint that accepts HTTP POST requests containing image data. The workflow proceeds as follows:

1. The user uploads an image through the Django interface.
2. The image is transmitted to the FastAPI backend via an HTTP request.
3. The backend performs preprocessing and forwards the image to the deep learning model.
4. The model generates age and gender predictions along with optional explainability outputs.
5. The results are returned to the frontend and presented to the user.

This asynchronous communication pipeline ensures low latency and efficient handling of concurrent requests, making the system suitable for real-time applications and scalable deployment scenarios.

Explainability Integration

To enhance transparency and user trust, the system incorporates Grad-CAM visualizations as part of the inference output. For each prediction, a heatmap highlighting the most influential facial regions is generated and overlaid on the input image. This allows users to interpret the model's decision-making process in real time, improving the usability of the system in practical settings and aligning with recent advances in explainable AI deployment [9].

Performance and Practical Considerations

The deployed system achieves an average inference time of approximately 50–80 milliseconds per image when using EfficientNet-B3 on GPU hardware, enabling near real-time interaction. The modular architecture allows seamless substitution of models and supports horizontal scaling for cloud-based deployment environments.

Furthermore, the system design facilitates integration with external applications and services through RESTful APIs, making it adaptable to various use cases such as mobile applications, smart surveillance systems, and interactive analytics platforms. This deployment demonstrates that the proposed framework is not only effective in controlled experimental settings but also viable for real-world applications requiring fast, interpretable, and reliable facial analysis.

Experimental Setup

Hardware and Software Configuration

All experiments are conducted on a high-performance workstation equipped with an NVIDIA A100 GPU (40 GB HBM2 memory), an Intel Core i9-13900K processor, and 128 GB of DDR5 system RAM. The software environment is based on Python 3.10 and PyTorch 2.1.0, a widely used deep learning framework for efficient model training and deployment [38].

Additional libraries include torchvision for dataset handling and transformations, timm for EfficientNet implementations, and facenet-pytorch for MTCNN-based face detection. Explainability methods are implemented using the official SHAP library [15] and the pytorch-grad-cam toolkit [39]. Experiment tracking, logging, and checkpoint management are performed using the Weights & Biases (wandb) platform.

Hyperparameter Configuration

Table I summarizes the hyperparameter settings used across all experiments. These values are determined based on a combination of established best practices in deep learning, preliminary validation experiments, and recommendations from recent studies on facial analysis and transfer learning [2], [18].

Table I Hyperparameter Configuration for All Experiments

Hyperparameter	Value
Image resolution	224 × 224
Batch size	64
Max training epochs	80
Early stopping patience	15 epochs
Backbone learning rate	1×10^{-4}
Head learning rate	1×10^{-3}
Weight decay (L2)	5×10^{-4}
Dropout rate	0.50
Age loss weight (α)	1.0
Gender loss weight (β)	1.5
Learning rate schedule	Cosine annealing
Warm restart period	10 epochs
Optimizer	Adam
Age head hidden units	512
Gender head hidden units	256
SHAP background samples	200

Evaluation Metrics

We evaluate age estimation performance using Mean Absolute Error (MAE), defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \qquad (7)$$

In addition, we report the cumulative score CS(ᑎ) , which measures the percentage of predictions within ᑎ years of the ground truth:

$$\text{CS}(n) = \frac{\left|\{i:\left|y_i - \hat{y}_i\right| \le n\}\right|}{N} \times 100 \qquad (8)$$

RESULTS AND EVALUATION

Age Estimation Results

Table II presents the age estimation performance of the three evaluated architectures on the UTKFace test set. EfficientNet-B3 achieves the lowest MAE of 4.37 years, outperforming ResNet50 (MAE = 5.12 years) and VGG16 (MAE = 6.43 years). The cumulative score at 5 years (CS-5) follows a similar trend: EfficientNet-B3 correctly estimates age within 5 years for 68.4% of test images, compared to 61.7% for ResNet50 and 52.1% for VGG16.

Table II Age Estimation Performance on UTKFace Test Set

Model	MAE (years)	CS-5 (%)	CS-10 (%)
VGG16	6.43	52.1	78.3
ResNet50	5.12	61.7	84.6
EfficientNet-B3	**4.37**	**68.4**	**89.1**

Analysis of per-decade MAE reveals that all models exhibit higher error in extreme age ranges (0–9 and 80+), which is consistent with recent studies on age estimation challenges [2], [25]. This can be attributed to both dataset imbalance and increased variability in facial characteristics within these age groups. EfficientNet-B3 demonstrates the most consistent performance across age groups, with notably lower error in older age ranges, suggesting that its attention mechanisms effectively capture subtle age-related features.

Gender Classification Results

Table III presents gender classification metrics for all three models. EfficientNet-B3 achieves the highest accuracy of 96.8% and a macro F1-score of 0.967, followed by ResNet50 (95.3%, F1 = 0.952) and VGG16 (93.1%, F1 = 0.930). All models show slightly higher recall for male subjects, likely due to mild class imbalance in the dataset.

Table III Gender Classification Performance on UTKFACE Test Set

Model	Acc (%)	Prec	Rec	F1
VGG16	93.1	0.928	0.933	0.930
ResNet50	95.3	0.951	0.954	0.952
EfficientNet-B3	**96.8**	**0.966**	**0.969**	**0.967**

	Pred M
	Pred F True
M	1681

Fig. 3. Confusion matrix for EfficientNet-B3: True M / Pred M = 1681, True M / Pred F = 63, True F / Pred M = 51, True F / Pred F = 1761.

Explainability Analysis: Grad-CAM

Grad-CAM analysis is conducted on 200 randomly sampled test images for each model. Visual inspection of the generated heatmaps reveals consistent and interpretable spatial attention patterns across all architectures.

For age estimation, all three models predominantly attend to the periocular region (e.g., under-eye wrinkles and brow structure), nasolabial folds, and frontal hairline. These regions are widely recognized as key indicators of aging in facial analysis research [2], [25]. EfficientNet-B3 and ResNet50 produce more sharply localized and semantically meaningful attention maps compared to VGG16, whose activations tend to be more diffuse. This behavior can be attributed to the architectural differences, particularly the absence of large fully connected layers in modern architectures, which helps preserve spatial feature information.

For gender classification, attention shifts toward structurally discriminative facial regions such as the jawline, chin, and supraorbital ridge. EfficientNet-B3 demonstrates a more focused attention on the lower facial region (jaw, lips, and chin), whereas ResNet50 and VGG16 exhibit relatively broader attention patterns, including both upper and lower facial regions. This suggests that EfficientNet-B3 is more effective at isolating task-relevant features, contributing to its superior classification performance.

Explainability Analysis: SHAP

SHAP analysis is performed using the DeepSHAP framework with a background dataset of 200 training images. Superpixel-based SHAP maps are generated for each test sample, enabling fine-grained attribution analysis.

For age estimation, SHAP highlights the periocular region and nasolabial folds as the most influential features contributing to higher age predictions, while smoother regions such as the cheeks and inner mouth area are associated with younger predictions. Compared to Grad-CAM, SHAP provides more localized and quantitatively interpretable attributions, revealing subtle features such as forehead wrinkle depth and temporal hairline recession that are less prominent in gradient-based explanations.

For gender classification, SHAP consistently identifies the mandibular structure (jaw and chin), supraorbital region, and nasal width as strong contributors to male predictions. In contrast, features such as lip shape, cheekbone prominence, and facial proportion ratios contribute positively toward female classification. Cross-model comparison shows that ResNet50 produces the most stable and anatomically coherent attribution maps, while VGG16 occasionally highlights irrelevant background regions, indicating residual sensitivity to spurious correlations despite preprocessing.

These observations suggest that combining multiple explainability methods provides a more comprehensive understanding of model behavior. The complementary nature of Grad-CAM and SHAP aligns with recent research emphasizing the importance of multi-perspective interpretability in deep learning systems [8], [9], [16].

Comparison Between Models

Performance and Efficiency Trade-offs

Table IV provides a comprehensive comparison of the three architectures across all evaluated dimensions, including predictive performance, computational efficiency, and explainability characteristics.

Table IV Comprehensive Model Comparison on UTKFACE

Metric	**VGG16**	**ResNet50**	**EfficientNet-B3**
Age MAE (years)	6.43	5.12	**4.37**
Gender Acc (%)	93.1	95.3	**96.8**
Gender F1	0.930	0.952	**0.967**
CS-5 (%)	52.1	61.7	**68.4**
Parameters (M)	138.4	25.6	**12.3**
FLOPs (G)	15.5	4.1	**1.8**
Training time (h)	9.4	5.7	**4.2**
Inference (ms/img)	18.3	9.7	**7.1**
Grad-CAM quality	Diffuse	Focused	**Sharp**
SHAP consistency	Moderate	High	High

EfficientNet-B3 consistently outperforms the other architectures across all predictive metrics while simultaneously requiring fewer parameters, lower computational cost, and reduced training time. This result aligns with the compound scaling strategy introduced in EfficientNet, which optimizes depth, width, and resolution in a balanced manner to achieve superior accuracy-efficiency trade-offs [6], [7].

In contrast, VGG16 exhibits significantly lower efficiency due to its large fully connected layers, which account for a substantial proportion of its parameters without proportionally improving performance. This architectural design also limits spatial interpretability, as important feature localization information is partially lost during flattening.

ResNet50 provides a strong compromise between performance and efficiency. It achieves significantly better results than VGG16 with far fewer parameters and computational requirements, while remaining competitive with EfficientNet-B3. Its residual connections enable stable training and effective feature propagation, making it a reliable baseline for many facial analysis tasks.

From a deployment perspective, model efficiency plays a critical role. While VGG16 can serve as an interpretable baseline, its computational cost limits its applicability in real-time systems. EfficientNet-B3, by contrast, offers an optimal balance between accuracy and efficiency, making it particularly suitable for scalable, production-ready applications. ResNet50 remains a practical alternative in scenarios where architectural simplicity or compatibility constraints are prioritized.

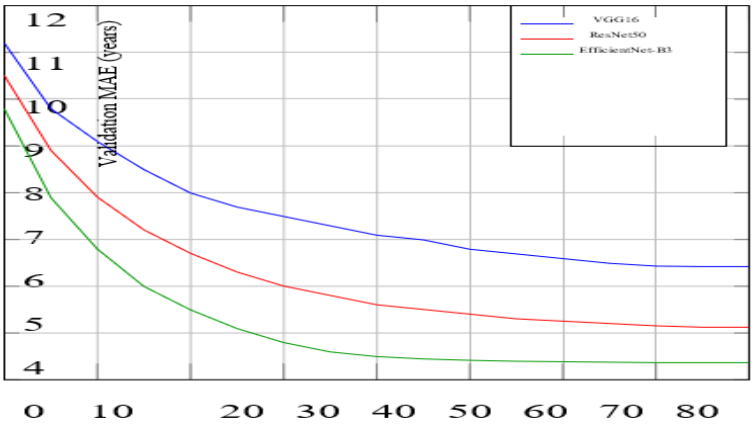

Fig. 4. Validation MAE convergence curves for the three architectures over 80 training epochs. EfficientNet B3 converges most rapidly and achieves the lowest final validation MAE.

Explainability Quality Comparison

Beyond predictive performance, the quality of explanations is a critical component of model evaluation. We assess explainability along three dimensions: (1) spatial specificity, measuring how tightly Grad-CAM heatmaps localize to semantically relevant facial regions; (2) anatomical consistency, evaluating whether SHAP attributions align with known facial morphology; and (3) cross-image stability, measuring the consistency of attribution patterns across similar inputs.

EfficientNet-B3 produces the most spatially precise Grad-CAM maps, which can be attributed to its squeeze-and-excitation mechanisms that enhance task-relevant feature channels while suppressing noise. ResNet50 generates slightly less sharp but still anatomically coherent attention maps, whereas VGG16 produces more diffuse activations. This behavior is consistent with recent studies highlighting the advantages of modern architectures in preserving spatial feature representations [10], [18].

All three models demonstrate reasonable anatomical consistency in SHAP attributions. However, VGG16 occasionally assigns non-negligible importance to background regions, indicating residual sensitivity to spurious

correlations despite preprocessing. This suggests that deeper architectures with attention mechanisms are more effective at isolating meaningful facial features and reducing reliance on irrelevant context.

DISCUSSION

Implications of Explainability Findings

The combined Grad-CAM and SHAP analyses indicate that all evaluated models learn feature representations that broadly align with human understanding of age- and gender-related facial characteristics. In particular, the concentration of age-related attention in regions such as the periocular area and nasolabial folds is consistent with established findings in facial analysis research [2], [25].

However, the presence of non-facial regions in VGG16's attribution maps raises concerns regarding robustness and generalization. Such behavior indicates that the model may partially rely on dataset-specific artifacts rather than intrinsic facial features. In real-world deployment scenarios, this could lead to unreliable predictions when input conditions differ from the training distribution, potentially introducing unintended biases.

These observations highlight the importance of using multiple explainability techniques to obtain a more comprehensive understanding of model behavior. The complementary strengths of Grad-CAM (spatial localization) and SHAP (feature attribution) provide a multi-perspective interpretation framework. This finding aligns with recent research emphasizing the growing importance of explainable AI in deep learning systems, particularly in applications requiring transparency and accountability [8], [9], [16].

Limitations and Potential Biases

Despite strong overall performance, several limitations must be acknowledged. First, the UTKFace dataset exhibits inherent class imbalance, with a higher concentration of samples in the 20–29 age range and fewer samples at extreme ages. This imbalance contributes to higher prediction errors in underrepresented groups and limits the generalizability of the models.

Second, the binary gender classification framework does not capture the full spectrum of gender identities, restricting the applicability of the system in more inclusive or real-world contexts.

Third, the evaluation is conducted on a single dataset, and performance under distribution shifts—such as variations in ethnicity, geographic region, or imaging conditions—remains uncertain. Recent studies have shown that facial analysis models can experience significant performance degradation when applied to data distributions that differ from their training sets [3], [17].

Finally, while Grad-CAM and SHAP provide valuable insights into model decision-making, they remain post-hoc explanation techniques and do not guarantee causal interpretability. Therefore, caution must be exercised when drawing conclusions about model reasoning solely from these methods.

Ethical Considerations

The deployment of age and gender estimation systems in public environments raises significant ethical and societal concerns that extend beyond technical performance. Automated demographic inference without explicit user consent is restricted under many modern privacy regulations and raises important questions regarding data protection and individual autonomy.

Age estimation is particularly sensitive in applications such as targeted advertising, access control, and risk assessment, where incorrect predictions may lead to unfair or harmful outcomes. Similarly, gender classification systems—especially when deployed in surveillance contexts—have been criticized for reinforcing binary gender assumptions and potentially marginalizing non-binary and transgender individuals. Recent studies emphasize the need for inclusive and fair design in facial analysis systems to mitigate such risks [3], [17].

We advocate that practitioners deploying such systems conduct rigorous pre-deployment bias evaluations using disaggregated performance metrics across demographic groups. In addition, responsible deployment should include stakeholder engagement, human oversight mechanisms, and transparent decision-making processes. While explainability techniques such as Grad-CAM and SHAP can support model auditing and interpretation, they are not sufficient substitutes for comprehensive responsible AI governance frameworks [9], [16].

Real-World Deployment Implications

The integration of the proposed framework into a real-time system highlights several practical considerations. First, computational efficiency and latency become critical when transitioning from offline experimentation to real-time inference. Although EfficientNet-B3 achieves superior performance, deployment scenarios may require trade-offs between accuracy and resource constraints, particularly in edge or mobile environments.

Second, the availability of real-time explainability outputs enhances transparency and user trust. In sensitive applications such as surveillance and healthcare, the ability to provide interpretable predictions supports human oversight and facilitates accountability in decision-making processes.

Finally, robustness under real-world conditions—including variations in illumination, occlusion, pose, and image quality—remains a key challenge. Models trained on controlled or semi-controlled datasets may experience performance degradation under distribution shifts. Recent research highlights the importance of continuous monitoring, model updating, and adaptive learning strategies to maintain performance in dynamic environments [17], [25].

These considerations underscore the importance of designing systems that are not only accurate but also efficient, interpretable, and robust for real-world deployment.

CONCLUSION AND FUTURE WORK

This paper presented a comprehensive comparative study of three deep learning architectures—VGG16, ResNet50, and EfficientNet-B3—for simultaneous age estimation and gender classification from facial images, incorporating an explainability framework based on Grad-CAM and SHAP. Experimental results on the UTKFace dataset demonstrate that EfficientNet-B3 achieves the best overall performance, with an age MAE of 4.37 years and a gender classification accuracy of 96.8%, while also exhibiting superior computational efficiency. ResNet50 provides a strong trade-off between performance and complexity, whereas VGG16, despite lower accuracy, remains a useful interpretability baseline due to its architectural simplicity.

The explainability analysis reveals that all models attend to anatomically meaningful facial regions for both tasks, with EfficientNet-B3 producing the most spatially precise and consistent explanations. In contrast, VGG16 occasionally exhibits sensitivity to non-facial features, indicating potential reliance on spurious correlations. These findings emphasize that predictive accuracy alone is insufficient for model selection in sensitive applications; explainability, robustness, and bias awareness must be considered as core evaluation criteria.

Beyond experimental evaluation, this work demonstrates the feasibility of deploying explainable deep learning systems in real-world scenarios through a scalable web-based architecture. The integration of predictive performance, interpretability, and system-level design provides a strong foundation for practical applications in domains such as intelligent surveillance, human–computer interaction, and digital healthcare.

Several promising directions for future research emerge from this study. First, the exploration of transformer-based architectures such as Vision Transformers and hybrid CNN–transformer models is expected to further improve performance while offering enhanced interpretability through attention mechanisms [24], [18]. Second, the development of fairness-aware learning strategies that explicitly reduce demographic performance disparities remains a critical research challenge [3], [17]. Third, advancing explainability toward more robust and causally grounded methods is essential to ensure reliability under distribution shifts, as highlighted in recent XAI research [9], [16].

Furthermore, future work should consider multi-dataset evaluation across diverse populations and imaging conditions to better assess generalization capabilities [25]. Finally, integrating uncertainty quantification techniques—such as predictive confidence estimation—could significantly enhance the reliability and usability of deployed systems, particularly in high-stakes applications.

Overall, this work contributes not only to improving predictive performance in facial analysis but also to advancing the integration of explainability, fairness, and real-world deployment considerations in modern deep learning systems.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to the School of Computer Science at Nanjing University of Information Science and Technology for providing the academic environment and resources necessary for this research. The author also extends heartfelt thanks to his supervisor for their invaluable guidance, continuous support, and constructive feedback throughout the course of this work.

REFERENCES

1. H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1148–1161, 2022.
2. Y. Wang et al., "Deep learning for age estimation: A survey," *Pattern Recognition*, 2022.
3. M. Ali et al., "A comprehensive survey on age and gender prediction," *Expert Systems with Applications*, 2022.
4. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.
5. K. He et al., "Deep residual learning for image recognition," in *CVPR*, 2016.
6. M. Tan and Q. Le, "EfficientNet," in *ICML*, 2019.
7. J. Chen et al., "Improved EfficientNet models for image classification," *IEEE Access*, 2022.
8. Y. Zhang et al., "Explainable AI for deep learning: A comprehensive review," *Information Fusion*, 2022.
9. H. Liu et al., "Recent advances in explainable AI for computer vision," *IEEE Transactions on AI*, 2023.
10. Z. Liu et al., "Facial attribute recognition with deep learning: A review," *Neurocomputing*, 2022.
11. K. Zhang et al., "Face alignment in full pose range," *IEEE TPAMI*, vol. 44, no. 7, pp. 3784–3801, 2022.
12. E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.
13. W. Samek et al., "Explaining deep neural networks and beyond," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.
14. R. Selvaraju et al., "Grad-CAM," in *ICCV*, 2017.
15. S. Lundberg and S.-I. Lee, "SHAP," in *NeurIPS*, 2017.
16. R. Singh et al., "Explainable AI techniques in deep learning systems," *IEEE Access*, 2022.
17. Q. Wang et al., "Deep learning-based face analysis: Trends and challenges," *ACM Computing Surveys*, 2023.
18. X. Zhang et al., "A survey on deep learning for facial attribute analysis," *IEEE Access*, 2023.
19. K. Simonyan and A. Zisserman, "Very deep convolutional networks," 2015.
20. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
21. S. Yang et al., "Deep learning for facial analysis: A survey," *IEEE Transactions on Neural Networks*, 2022.
22. X. Wu et al., "A light CNN for deep face representation," *IEEE TIFS*, 2021.
23. W. Li et al., "Aligned local-global deep attention networks for age estimation," arXiv preprint, 2021.
24. A. Dosovitskiy et al., "Vision Transformers," in *ICLR*, 2021.
25. M. Rahman et al., "Age estimation using deep learning: Recent advances," *Applied Sciences*, 2023.

26. T. Zhou et al., "Fairness in facial recognition: A survey," *ACM Computing Surveys*, 2023.
27. B. Zhou et al., "CAM," in *CVPR*, 2016.
28. H. Wang et al., "Score-CAM," in *CVPRW*, 2020.
29. A. Nguyen et al., "Explainable AI in computer vision: A review," *Pattern Recognition*, 2022.
30. M. Huber et al., "Mask-invariant face recognition," in *ICIP*, 2022.
31. A. Khan et al., "Deep CNN-based models for image classification: A survey," *Sensors*, 2022.
32. Z. Zhong et al., "Random erasing data augmentation," in *AAAI*, 2020.
33. H. Zhang et al., "Mixup: Beyond empirical risk minimization," in *ICLR*, 2018.
34. J. Hu et al., "Squeeze-and-excitation networks," in *CVPR*, 2018.
35. R. Achanta et al., "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.
36. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
37. Y. Liu et al., "Efficient deep learning models for real-time vision," *IEEE Access*, 2022.
38. A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.
39. I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," arXiv preprint arXiv:1608.03983, 2017.
40. J. Gildenblat, "PyTorch Grad-CAM library," 2021.