Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

Interpretable and Statistically Validated Comparative Evaluation of EfficientNetB0, MobileNetV2, and ResNet50 for Bold and Natural Makeup Classification on CelebA

Aurelia Chiara Suryabangun*[1], Abdussalam[2]

[1,2]Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Email: [1]111202214757@mhs.dinus.ac.id

Received : Apr 8, 2026; Revised : Apr 29, 2026; Accepted : Apr 29, 2026; Published : Jun 15, 2026

Abstract

Facial makeup classificationplays a critical role in beauty technology, visual style analysis, and intelligent web-based image inference. Distinguishing bold makeup from natural makeup is challenging due to subtle visual overlap, borderline facial appearance, and inconsistent makeup intensity across images. While numerous prior studies have applied deep learning for facial analysis, most focus solely on conventional performance metrics without addressing statistical validation, probability calibration, or interpretability — a critical gap that limits reliable model selection in visually subtle classification tasks. This study presents an interpretable and statistically validated comparative evaluation of three transfer learning architectures — EfficientNetB0, MobileNetV2, and ResNet50 — for binary makeup classification using a curated CelebA-based dataset. The final dataset comprises 12,000 facial images equally divided into natural_makeup and bold_makeup classes, with separate training, validation, and clean test subsets. Models were evaluated using holdout testing, 10-fold cross-validation, McNemar statistical testing, calibration analysis, confidence intervals, ROC and PR curves, and Grad-CAM visualization. Experimental results show that EfficientNetB0 achieved the best overall performance, with 0.7900 Accuracy, 0.7898 Macro-F1, 0.8829 ROC-AUC, and 0.8461 PR-AUC on the clean holdout test set. Across ten-fold cross-validation, EfficientNetB0 further achieved 0.7801 ± 0.0093 Accuracy and 0.8780 ± 0.0090 ROC-AUC. It also demonstrated the strongest calibration performance, with the lowest Expected Calibration Error (ECE = 0.0558) and Brier Score (0.1449) among all compared models. The selected model was further implemented in a FastAPI-based backend system for web-based prediction. From a broader Informatics and Computer Science perspective, this study contributes a rigorous and reproducible evaluation framework that integrates statistical validation, calibration assessment, and interpretability, enabling more reliable model selection in visually subtle facial analysis tasks and supporting practical deployment in intelligent systems.

Keywords: *Bold makeup, Cross-validation, EfficientNetB0, Facial makeup classification, MobileNetV2, Natural makeup, ResNet50*

1. INTRODUCTION

Facial image analysis has become a central topic in computer vision due to its broad applicability in aesthetic computing, beauty technology, recommendation systems, and intelligent visual analysis [1], [2], [3], [4], [11], [13], [19], [21], [22], [24], [25]. Among these applications, makeup style classification is particularly interesting because facial makeup not only represents a cosmetic attribute but also reflects visual patterns that influence perceived appearance, style, and identity [1], [5], [11], [30], [33], [34]. Automatic recognition of makeup style can support practical systems such as beauty assistance platforms, personalized cosmetic recommendation engines, and web-based image analysis tools [2], [3], [4], [21], [44], [48]. The growing demand for intelligent appearance analysis systems in health informatics, e-commerce, and social media platforms has further motivated the development of accurate, interpretable, and statistically reliable classification pipelines for facial attribute analysis [26], [27], [31].

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

However, despite its practical significance, makeup style classification remains underexplored in terms of rigorous evaluation methodology — a gap that this study directly addresses.

Despite its relevance, binary facial makeup classification remains challenging. Distinguishing bold makeup from natural makeup is not straightforward because these two classes often share overlapping visual cues. Variations in lighting, facial pose, image quality, skin tone, eyebrow contrast, lip color, and eye makeup intensity may result in borderline cases that are difficult to classify consistently [5], [11], [12], [30]. In real-world deployment, such ambiguities are further amplified when an inference system is expected to provide stable and interpretable predictions from uploaded facial images [4], [10], [28]. Moreover, the reliability of predictions depends not only on classification accuracy but also on how well model confidence is calibrated, particularly for subtle visual tasks where overconfident errors can mislead downstream decisions [10]. These challenges demand a multi-dimensional evaluation approach that goes beyond single-metric benchmarking.

While numerous studies have applied deep learning for facial analysis, most focus solely on conventional performance metrics, such as accuracy and F1-score, without addressing statistical validation, probability calibration, or interpretability [6], [7], [8], [19], [20], [23], [37], [40], [43]. Specifically, Boutros et al. [5] applied VGG16-based transfer learning for automatic makeup detection on multiple public datasets but evaluated model performance solely using accuracy, without incorporating statistical significance testing to validate differences between classifiers. Similarly, Wiles et al. [16] demonstrated that evaluation of facial attribute prediction models on the CelebA dataset commonly relies on accuracy-based metrics alone [6], [7], [8], [19], [20], [23], [45], which is insufficient for determining which approach is truly superior, while also lacking probability calibration analysis essential for confidence reliability in deployed systems. Liu et al. [30] applied graph convolutional networks for facial attribute classification and reported accuracy-based results without incorporating McNemar statistical testing or calibration assessment. Furthermore, studies applying deep CNN architectures for facial beauty and appearance analysis [1], [22], [32] have consistently reported performance using threshold-dependent metrics without incorporating visual interpretability methods such as Grad-CAM, SHAP, or LIME [18] to identify which facial regions drive model decisions. Bobba [14] compared ResNet50 and EfficientNet for transfer learning image classification but did not extend evaluation to include statistical significance testing or probability calibration analysis. Şener et al. [26] applied McNemar's test alongside EfficientNetB0 for medical image classification, demonstrating the value of statistical validation beyond raw accuracy, yet their work did not address calibration quality or visual interpretability in a unified framework.

Collectively, these prior works reveal a critical and recurring gap in the literature: no study on facial makeup or facial attribute classification has simultaneously integrated (1) cross-validation, (2) pairwise statistical significance testing via McNemar's test, (3) probability calibration analysis using Expected Calibration Error (ECE) and Brier Score, and (4) visual interpretability via Grad-CAM into a single, reproducible evaluation pipeline. This omission is consequential — selecting a model based on holdout accuracy alone, without validating statistical significance or calibration quality, risks deploying an unreliable system in practice [10], [17], [27], [28], [29]. The present study is the first to directly address this gap specifically within the domain of binary facial makeup classification on a curated CelebA-based dataset, thereby contributing a new methodological benchmark for reliable model evaluation in visually subtle facial analysis tasks.

Recent advances in deep learning and transfer learning have significantly improved image classification performance across diverse visual tasks. Architectures such as EfficientNetB0, MobileNetV2, and ResNet50 are widely adopted due to their strong feature extraction capabilities and competitive performance across many domains [6], [7], [8], [15], [31], [41], [46]. Patrício et al. [27] highlighted that interpretability methods such as Grad-CAM remain underutilized in classification

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

evaluation pipelines, despite their demonstrated value in revealing model decision mechanisms [9]. Bradshaw et al. [28] further emphasized that cross-validation strategies are critical for unbiased performance estimation in AI classification systems, particularly when comparing multiple architectures [29]. However, model comparison alone is insufficient for building reliable applied systems; evaluation must also consider statistical significance, calibration quality, confidence reliability, and interpretability — especially for tasks with subtle visual differences [10], [13], [15], [16].

To address these gaps, this study presents an interpretable and statistically validated comparative evaluation of three transfer learning architectures — EfficientNetB0, MobileNetV2, and ResNet50 — for bold and natural makeup classification using a curated CelebA-based dataset. The dataset was refined into two balanced classes and split into training, validation, and clean test subsets to support fair benchmarking [16], [17]. The models were evaluated using holdout testing, 10-fold cross-validation, McNemar statistical testing, calibration analysis, confidence intervals, ROC and PR curves, confusion matrix analysis, and Grad-CAM visualization [9], [10], [17], [18], [26], [28]. The best-performing model was further implemented in a FastAPI-based backend system for web-based prediction [4]. The primary objective of this study is to demonstrate that a comprehensive and statistically rigorous evaluation framework produces more reliable model selection outcomes than conventional accuracy-based comparison alone, thereby advancing the standard of evaluation practice in visually subtle facial analysis tasks within the field of Informatics and Computer Science.
The main contributions of this study are as follows:

1. Construction of a balanced binary facial makeup classification dataset based on curated CelebA images consisting of natural_makeup and bold_makeup classes.
2. Comprehensive comparison of EfficientNetB0, MobileNetV2, and ResNet50 under a unified and rigorous evaluation protocol incorporating holdout testing and 10-fold cross-validation.
3. Extension of the evaluation beyond conventional metrics by incorporating statistical validation via McNemar's test, calibration assessment via ECE and Brier Score, and visual interpretability via Grad-CAM — the first study to unify all four evaluation dimensions in facial makeup classification [17], [18], [36], [37].
4. Deployment of the selected model in a web-oriented inference pipeline, demonstrating practical applicability and deployment readiness.
5. Provision of a reproducible evaluation framework for reliable model selection in visually subtle facial analysis tasks, directly contributing to the field of Informatics and Computer Science by guiding both research and applied system development.

Based on these objectives, the remainder of this paper is organized as follows: Section 2 presents the dataset construction process, preprocessing pipeline, transfer learning models, and evaluation protocol. Section 3 reports the experimental results, including comparative performance, statistical analysis, calibration, and interpretability findings. Section 4 discusses the implications of the results, the observed error patterns, and the limitations of the study. Section 5 concludes the paper and outlines possible future work.

2. METHOD

This study employed a comparative deep learning framework to evaluate three transfer learning architectures, namely efficientnetb0, mobilenetv2, and resnet50, for binary facial makeup classification [6]–[10]. The overall workflow consisted of dataset construction, label definition, preprocessing, model evaluation, statistical validation, interpretability analysis, and system implementation [9], [10], [17]–[20]. The complete research pipeline is illustrated in figure 1, while representative examples of the final dataset are shown in figure 2.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

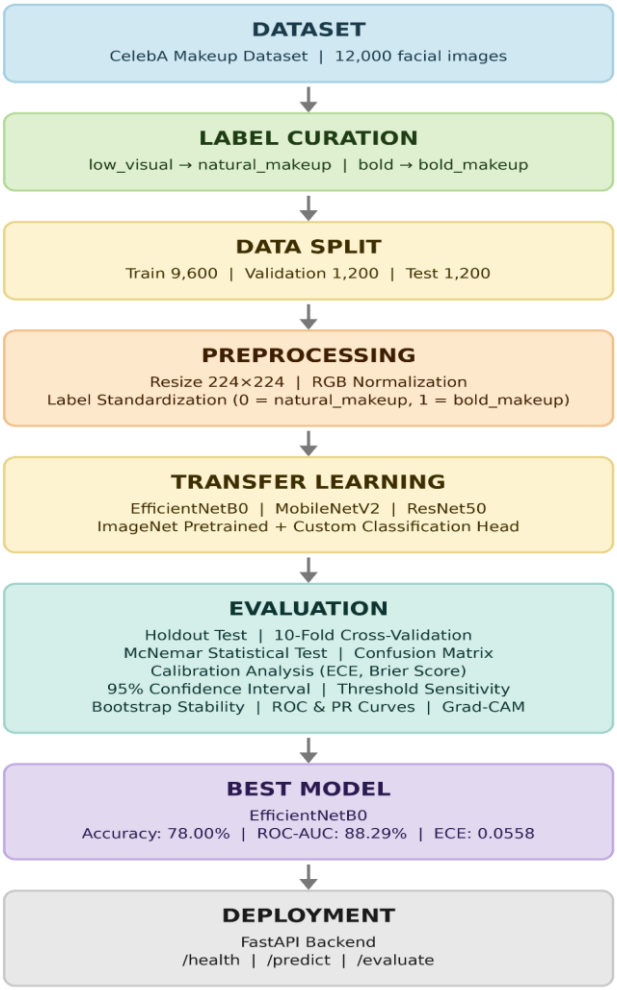

Figure 1. Research Workflow

Figure 2. Sample Dataset Collage

2.1. Dataset Construction and Label Definition

The dataset used in this study was based on CelebA, which provides a large collection of facial images suitable for appearance-based visual analysis [13]. To support the objective of this work, the dataset was curated into a binary classification setting consisting of two classes, namely

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

natural_makeup and **bold_makeup**. The final curated dataset contained **12,000 facial images**, with **6,000 images per class**, in order to maintain class balance and reduce bias during model evaluation.

The final label structure was standardized to ensure consistency throughout the training, evaluation, and deployment pipeline. Images categorized as visually subtle makeup appearance were assigned to the **natural_makeup** class, whereas images with stronger and more visually prominent makeup appearance were assigned to the **bold_makeup** class. This binary setting was selected to simplify the classification task while preserving its practical relevance for web-based inference and comparative deep learning evaluation [3], [4], [13].

The distribution of the final dataset is presented in **Table 1**, which shows the balanced allocation across training, validation, and clean test subsets.

Table 1. Composition of the Final CelebA-Based Dataset

Split	Natural Makeup	Bold Makeup	Total
Train	4,800	4,800	**9,600**
Validation	600	600	**1,200**
Test Clean	600	600	**1,200**
Total	6,000	6,000	**12,000**

2.2. Dataset Split Strategy and Preprocessing

To ensure fair evaluation, the final dataset was divided into three disjoint subsets: training, validation, and clean test data. Each class contained **4,800 images for training**, **600 images for validation**, and **600 images for clean testing**, resulting in a final split of **9,600 training images**, **1,200 validation images**, and **1,200 clean test images**. This split strategy was designed to support robust benchmarking while preventing overlap between model development and final evaluation [17], [18].

All input images were resized to **224 × 224 pixels** and converted into RGB format before being processed by the classification models [6]–[8]. The input tensors were prepared in floating-point representation to match the expected inference pipeline of the selected transfer learning architectures. In addition, the final label mapping was standardized such that **0 represented natural_makeup** and **1 represented bold_makeup**. This consistent label ordering was maintained across prediction files, statistical analysis, calibration analysis, and backend implementation.

For web-based inference, an additional face validation step was employed before classification. If no valid facial region was detected in the uploaded image, the system returned a **no_face_detected** status instead of forcing a class prediction. This mechanism was introduced to improve system robustness during deployment [4], [10].

2.3. Transfer Learning Architectures

Three widely used transfer learning architectures were selected for comparative evaluation, namely **EfficientNetB0**, **MobileNetV2**, and **ResNet50** [6]–[8], [14]–[16]. These models were chosen because they represent different design trade-offs in deep learning-based image classification.

EfficientNetB0 was included due to its compound scaling strategy and strong efficiency-to-performance balance [14]. MobileNetV2 was selected because of its lightweight architecture and deployment-oriented efficiency, making it relevant for practical inference systems [15]. ResNet50 was included as a widely adopted residual learning model that has demonstrated strong performance in numerous computer vision tasks [16]. The use of these three architectures enabled a balanced comparison between model compactness, discriminative performance, and suitability for final deployment.These three architectures represent a diverse range of design philosophies — from

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

compound scaling [13] and lightweight mobile-oriented design [6] to deep residual learning [7] — and have demonstrated strong performance across numerous visual classification domains including facial analysis and appearance recognition [38], [39], [47].

2.4. Evaluation Protocol

The evaluation protocol in this study was designed to extend beyond conventional accuracy-based comparison. First, all three models were assessed on the clean holdout test set using four primary metrics: Accuracy, Macro-F1, ROC-AUC, and PR-AUC [9], [10], [17]. These metrics were selected to capture both threshold-dependent and ranking-based classification performance.

Second, a 10-fold cross-validation experiment was conducted to estimate model stability and generalization under repeated partitioning [28], [29]. For each model, the mean and standard deviation of Accuracy, Macro-F1, ROC-AUC, and PR-AUC were reported across all ten folds. This procedure provided a more robust view of performance consistency compared with a single split evaluation. The reported 10-fold cross-validation results were as follows:

EfficientNetB0: Accuracy = 0.7801 ± 0.0093, Macro-F1 = 0.7796 ± 0.0094, ROC-AUC = 0.8780 ± 0.0090, PR-AUC = 0.8469 ± 0.0158

MobileNetV2: Accuracy = 0.7819 ± 0.0118, Macro-F1 = 0.7818 ± 0.0118, ROC-AUC = 0.8636 ± 0.0131, PR-AUC = 0.8199 ± 0.0211

ResNet50: Accuracy = 0.7371 ± 0.0126, Macro-F1 = 0.7335 ± 0.0124, ROC-AUC = 0.8568 ± 0.0110, PR-AUC = 0.8224 ± 0.0148

Third, pairwise McNemar statistical testing was applied to compare classifier disagreement patterns on the clean test set [17], [26]. This statistical test was used to determine whether the observed differences between models were statistically significant or merely incidental.

Fourth, calibration analysis was performed using Expected Calibration Error (ECE) and Brier Score [10]. Calibration analysis was included because a strong classification model should not only produce correct predictions, but also provide confidence estimates that are well aligned with actual correctness. This aspect is particularly important for deployed systems where prediction confidence directly influences user trust and downstream decision-making.

Fifth, approximate 95% confidence intervals were computed for holdout accuracy in order to provide an uncertainty-aware interpretation of model performance [17]. In addition, threshold sensitivity analysis was performed by varying the positive-class decision threshold to assess how decision boundaries affected classification outcomes. To further examine performance robustness, a bootstrap-based stability analysis was also conducted to estimate repeated metric variability across sampled test subsets [17], [18]. The formal definitions of all evaluation metrics are provided in Section 2.5.

2.5. Evaluation Metrics Formulation

The evaluation metrics used in this study are formally defined as follows. The selection of these metrics was guided by the need to assess not only raw predictive accuracy but also statistical reliability, confidence calibration, and discriminative capability across decision thresholds [10], [36], [40].

Accuracy measures the proportion of correctly classified samples over the total number of samples, as defined in Equation (1):

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

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. This metric was selected to provide a general overview of classification performance across the balanced test set.

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

Macro-F1 score is computed as the unweighted mean of per-class F1 scores, as defined in Equation (2):

$$F1\text{-}Score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (2)$$

This metric was selected to ensure balanced evaluation across both classes, as subtle misclassification in either direction carries practical consequences in makeup style recognition [9], [10].

The McNemar test statistic is defined in Equation (3) [26]:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \quad (3)$$

where b and c represent the number of samples misclassified differently by two classifiers on the same test set. This test was used to determine whether performance differences between models were statistically significant rather than incidental [26], [27].

Expected Calibration Error (ECE) is defined in Equation (4) [10]:

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{n} |\mathrm{acc}(B_k) - \mathrm{conf}(B_k)| \quad (4)$$

where B_k denotes the set of samples in bin k, acc(B_k) is the empirical accuracy within that bin, conf(B_k) is the mean predicted confidence, and n is the total number of samples. This metric was selected to evaluate how well predicted probabilities reflect actual correctness — a property especially critical for deployed systems [10], [40], [43].

The inclusion of these four metrics — together with Brier Score, ROC-AUC, PR-AUC, confidence intervals, and bootstrap-based stability analysis — ensures comprehensive coverage of predictive performance, statistical validity, confidence reliability, and model interpretability [27], [28], [29].

Data augmentation strategies [17] such as random flipping, rotation, and brightness adjustment were also applied during training to improve model robustness and reduce overfitting in borderline visual classification cases.

2.6. Interpretability Analysis

To improve transparency in model behavior, this study incorporated Grad-CAM as a visual interpretability method [9], [18], [27]. Several studies have demonstrated the complementary use of Grad-CAM alongside SHAP and LIME for visual explanation of CNN decisions [18], [36], [37]. Alternative interpretability techniques such as SHAP [36] or LIME were considered for future extension but were not included in the primary evaluation scope of this study. Grad-CAM was applied to the final selected model in order to highlight facial regions that most strongly influenced the predicted class. This analysis was particularly important because makeup classification involves subtle visual features, such as eye makeup intensity, eyebrow contrast, and lip appearance, which may overlap between classes [5], [9].

In addition to Grad-CAM, a **confusion matrix** was analyzed to identify dominant error patterns in the clean test set [10], [17]. An **error summary table** and a **representative failure case figure** were also prepared to support qualitative interpretation of false positive and false negative behavior.

2.7. Backend and Web-Based Implementation

The best-performing model was deployed within a FastAPI-based backend system to demonstrate the practical applicability of the proposed framework [4], [10], [35]. Face validation before inference

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

was applied to ensure that only valid facial inputs were processed by the classifier, following established face feature extraction practices [35], [38]. The deployed backend provided three main endpoints: /health, /predict, and /evaluate. The /predict endpoint accepted uploaded facial images and returned the predicted class, confidence score, and class probabilities when a valid face was detected. If no face was detected, the system returned a no_face_detected response instead of a forced class assignment.

This deployment stage was included to demonstrate that the selected comparative model was not only suitable for offline experimentation, but also feasible for web-based inference. The system-level implementation further strengthened the applied contribution of this study by bridging experimental benchmarking and practical usage [3], [4], [10].

3. RESULT

This section presents the experimental results of the proposed comparative evaluation framework for binary facial makeup classification. The reported findings include holdout evaluation, ten-fold cross-validation, statistical significance testing, calibration analysis, confidence interval estimation, confusion matrix interpretation, Grad-CAM visualization, threshold sensitivity analysis, bootstrap-based stability analysis, and web-based system output. The overall project was conducted on a curated CelebA-only dataset containing 12,000 facial images, equally distributed into natural_makeup and bold_makeup classes.

3.1. Holdout Evaluation Results

As shown in Table 2, the primary comparative results on the clean holdout test set indicate that EfficientNetB0 achieved the best overall performance, obtaining 0.7900 Accuracy, 0.7898 Macro-F1, 0.8829 ROC-AUC, and 0.8461 PR-AUC. MobileNetV2 followed very closely with 0.7892 Accuracy, 0.7892 Macro-F1, 0.8654 ROC-AUC, and 0.8167 PR-AUC, while ResNet50 produced the lowest overall performance with 0.7383 Accuracy, 0.7356 Macro-F1, 0.8522 ROC-AUC, and 0.8130 PR-AUC.

These results indicate that EfficientNetB0 provided the strongest balance between threshold-dependent and ranking-based classification performance on the independent clean test set. Although MobileNetV2 was highly competitive in Accuracy and Macro-F1, EfficientNetB0 remained superior in ROC-AUC and PR-AUC, indicating better ranking quality across decision thresholds. Based on these holdout results, EfficientNetB0 was selected as the primary candidate for final deployment.

Table 2. Holdout Evaluation Results on the Clean Test Set

Model	Accuracy	Macro-F1	ROC-AUC	PR-AUC
EfficientNetB0	0.7900	0.7898	0.8829	0.8461
MobileNetV2	0.7892	0.7892	0.8654	0.8167
ResNet50	0.7383	0.7356	0.8522	0.8130

3.2. Ten-Fold Cross-Validation Results

To evaluate model robustness beyond a single holdout split, ten-fold cross-validation was conducted. As shown in Table 3, the summary results indicate that MobileNetV2 slightly outperformed EfficientNetB0 in terms of mean Accuracy and Macro-F1, achieving 0.7819 ± 0.0118 Accuracy and 0.7818 ± 0.0118 Macro-F1. In comparison, EfficientNetB0 achieved 0.7801 ± 0.0093 Accuracy and 0.7796 ± 0.0094 Macro-F1. However, EfficientNetB0 remained superior in ROC-AUC (0.8780 ± 0.0090) and PR-AUC (0.8469 ± 0.0158), indicating stronger ranking-based discrimination across all ten partitions. ResNet50 consistently remained the weakest architecture across all metrics, achieving 0.7371

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

± 0.0126 Accuracy, 0.7335 ± 0.0124 Macro-F1, 0.8568 ± 0.0110 ROC-AUC, and 0.8224 ± 0.0148 PR-AUC.

Ten-fold cross-validation was selected to balance computational efficiency and evaluation robustness [28], [29]. The per-fold accuracy results for all three models are illustrated in Figure 3, which presents the fold-by-fold performance variation to facilitate visual comparison of model stability across partitions. The cross-validation summary results are presented in Table 3 for quantitative reference.

Table 3. Ten-Fold Cross-Validation Results

Model	Accuracy ± std	Macro-F1 ± std	ROC-AUC ± std	PR-AUC ± std
EfficientNetB0	0.7801 ± 0.0093	0.7796 ± 0.0094	0.8780 ± 0.0090	0.8469 ± 0.0158
MobileNetV2	0.7819 ± 0.0118	0.7818 ± 0.0118	0.8636 ± 0.0131	0.8199 ± 0.0211
ResNet50	0.7371 ± 0.0126	0.7335 ± 0.0124	0.8568 ± 0.0110	0.8224 ± 0.0148

The per-fold accuracy results for all three models are illustrated in Figure 3, which presents the fold-by-fold performance variation to facilitate visual comparison of model stability across partitions.

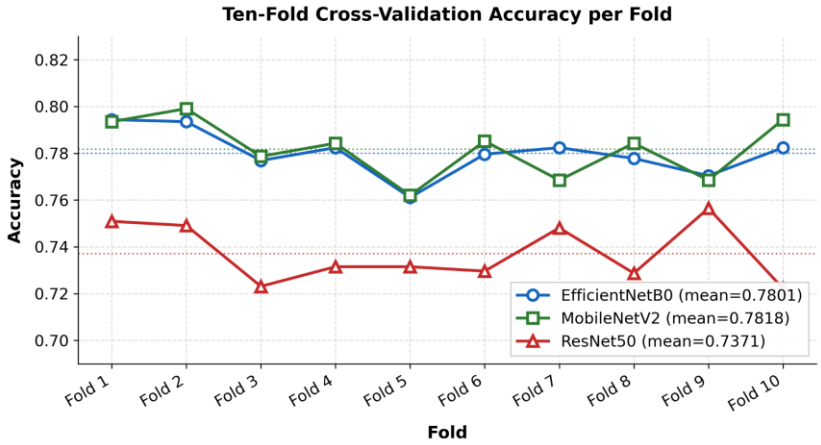

Figure 3. Ten-Fold Cross-Validation Accuracy per Fold for Each Model

These results suggest that the performance gap between EfficientNetB0 and MobileNetV2 is relatively small in threshold-based metrics, whereas EfficientNetB0 provides a stronger advantage in ranking-based performance. The standard deviation of EfficientNetB0 across folds (± 0.0093 in Accuracy) was notably lower than that of MobileNetV2 (± 0.0118), indicating that EfficientNetB0 produced more consistent performance across different data partitions. Therefore, the cross-validation results reinforce the interpretation that the two leading architectures are closely competitive, while ResNet50 is less suitable for the final pipeline.

3.3. Pairwise McNemar Test Results

To determine whether the differences between models were statistically meaningful, pairwise McNemar tests were conducted on the clean test set. As shown in Table 4, the results indicate that the comparison between EfficientNetB0 and ResNet50 was statistically significant, and the comparison between MobileNetV2 and ResNet50 was also statistically significant. In contrast, the comparison between EfficientNetB0 and MobileNetV2 was not statistically significant.

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

Table 4. Pairwise McNemar Test Results

Model Comparison	p-value	Decision
EfficientNetB0 vs ResNet50	0.00003317	Significant
EfficientNetB0 vs MobileNetV2	1.00000000	Not Significant
ResNet50 vs MobileNetV2	0.00011860	Significant

These findings indicate that both EfficientNetB0 and MobileNetV2 significantly outperformed ResNet50, while the top two architectures produced highly similar disagreement behavior. This statistical result supports the interpretation that EfficientNetB0 and MobileNetV2 are closely matched, but EfficientNetB0 remains preferable due to stronger ROC-AUC, PR-AUC, and deployment alignment.

3.4. Calibration and Confidence Interval Analysis

In addition to predictive performance, confidence reliability was assessed using calibration analysis. As shown in Table 5, the calibration comparison indicates that EfficientNetB0 achieved the best calibration performance, with the lowest Expected Calibration Error (ECE = 0.0558) and the lowest Brier Score (0.1449). MobileNetV2 also showed competitive calibration behavior, with ECE = 0.0623 and Brier Score = 0.1512. In contrast, ResNet50 exhibited much poorer calibration, with ECE = 0.1760 and Brier Score = 0.2116, despite having the highest average confidence.

Table 5. Calibration Comparison Across Models

Model	Accuracy	Average Confidence	ECE	Brier Score
EfficientNetB0	0.7900	0.8412	0.0558	0.1449
MobileNetV2	0.7892	0.8440	0.0623	0.1512
ResNet50	0.7383	0.9144	0.1760	0.2116

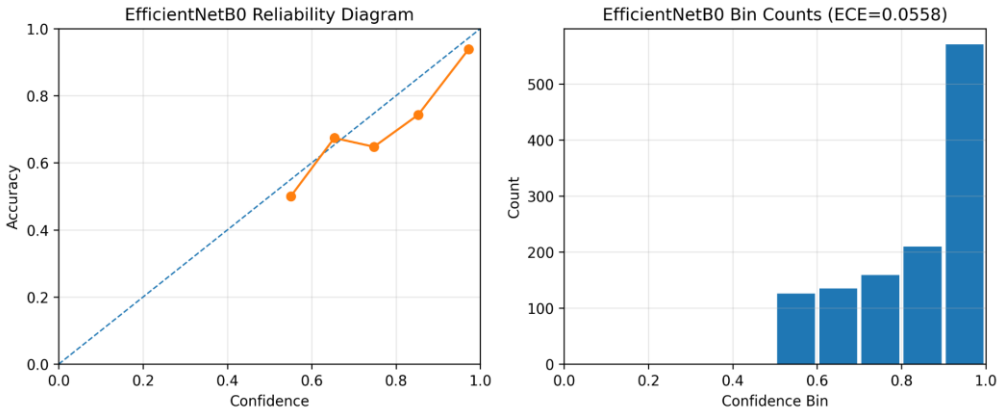

Figure 4/ illustrates the reliability diagram of EfficientNetB0.

This result indicates that ResNet50 tended to produce overconfident predictions, whereas EfficientNetB0 generated confidence estimates that were better aligned with actual correctness. Therefore, EfficientNetB0 was not only the best-performing model in the holdout evaluation, but also the most reliable in terms of confidence calibration. This additional evidence further supports its selection as the final deployed model.

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

To quantify uncertainty around the main holdout accuracy values, approximate 95% confidence intervals were also computed. As shown in Table 6, EfficientNetB0 achieved a confidence interval of 0.7660–0.8121, while MobileNetV2 achieved 0.7652–0.8113. These overlapping intervals confirm that the two strongest models are closely competitive. In contrast, ResNet50, with a confidence interval of 0.7127–0.7624, remained clearly lower than the top two architectures.

Table 6. Approximate 95% Confidence Intervals

Model	Accuracy	95% Confidence Interval
EfficientNetB0	0.7900	0.7660 – 0.8121
MobileNetV2	0.7892	0.7652 – 0.8113
ResNet50	0.7383	0.7127 – 0.7624

3.5. Confusion Matrix and Error Summary

To examine class-wise prediction behavior, the confusion matrix of EfficientNetB0 is analyzed. As shown in Figure 5, the confusion matrix illustrates the class-wise prediction distribution, while Table 7 provides a tabular error summary. EfficientNetB0 correctly classified 493 natural_makeup images and 455 bold_makeup images. However, it misclassified 107 natural_makeup images as bold_makeup and 145 bold_makeup images as natural_makeup.

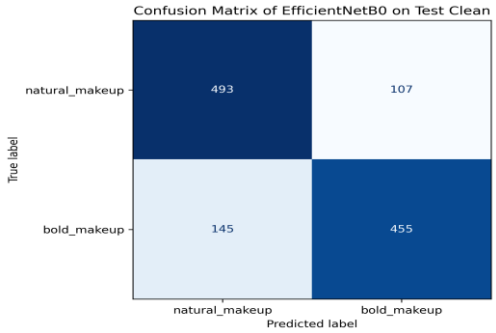

Figure 5. Confusion Matrix of EfficientNetB0

The confusion pattern reveals that the dominant error occurred when bold_makeup was predicted as natural_makeup, indicating that the bold class remained more difficult to classify consistently. This suggests that the visual distinction between the two classes was not always strong, particularly in borderline appearances where eye emphasis, eyebrow contrast, or lip color did not appear highly dominant. Such a pattern is consistent with the practical difficulty of separating subtle and prominent makeup styles under diverse visual conditions.

Table 7. Error Summary of EfficientNetB0 on the Clean Test Set

Actual Class	Predicted Class	Count	Interpretation
Natural Makeup	Natural Makeup	493	Correct classification
Natural Makeup	Bold Makeup	107	False positive for bold makeup
Bold Makeup	Bold Makeup	455	Correct classification
Bold Makeup	Natural Makeup	145	False negative for bold makeup

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

3.6. Grad-CAM Visualization

To improve interpretability, Grad-CAM was applied to the selected EfficientNetB0 model. As shown in Figure 6 and Figure 7, the resulting visualizations represent correctly classified examples from the bold_makeup and natural_makeup classes, respectively. The activation maps indicate that the model primarily focused on visually meaningful facial regions, including the eye area, eyebrow region, and lip area.

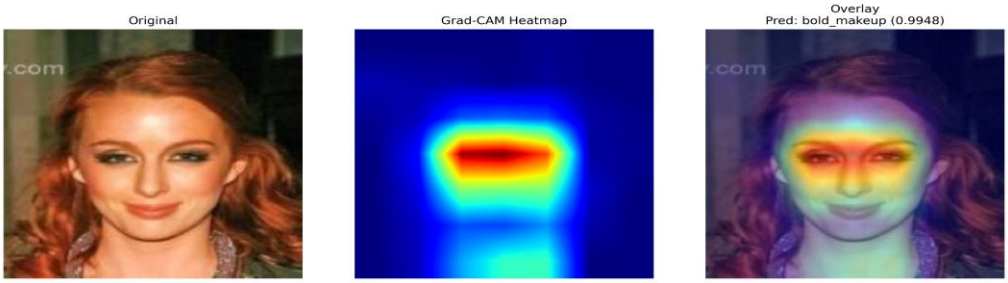

Figure 6. Grad-CAM on a Correctly Classified Bold Makeup Image

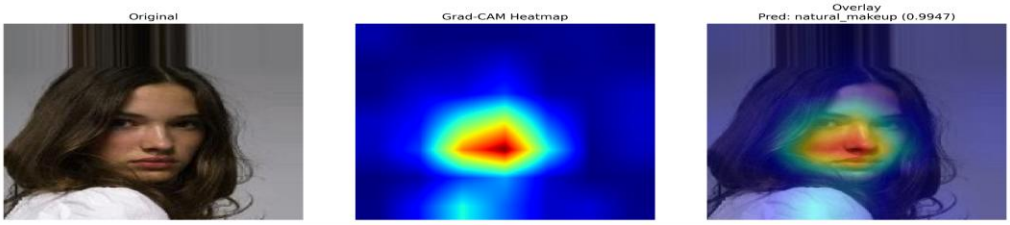

Figure 7. Grad-CAM on a Correctly Classified Natural Makeup Image

These visualizations suggest that the selected model captured semantically relevant appearance cues rather than relying on arbitrary background information. Thus, the Grad-CAM analysis provides qualitative support that the model's decisions were guided by plausible facial makeup features. This interpretability component is important because the project was designed not only as a classification benchmark, but also as an explainable applied system.

3.7. ROC, PR, and Threshold Sensitivity Analysis

The ROC and Precision–Recall comparisons across the three evaluated models are presented. As illustrated in Figure 8 and Figure 9, EfficientNetB0 achieved the strongest discrimination capability, followed by MobileNetV2 and ResNet50. A similar trend is visible in the PR comparison, where EfficientNetB0 maintained the best precision–recall trade-off across the operating range. These findings are consistent with the ranking-based metrics reported earlier in the holdout evaluation.

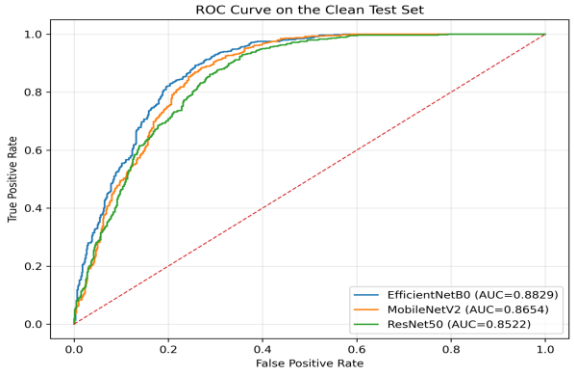

Figure 8. ROC Curve Comparison Across Models

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

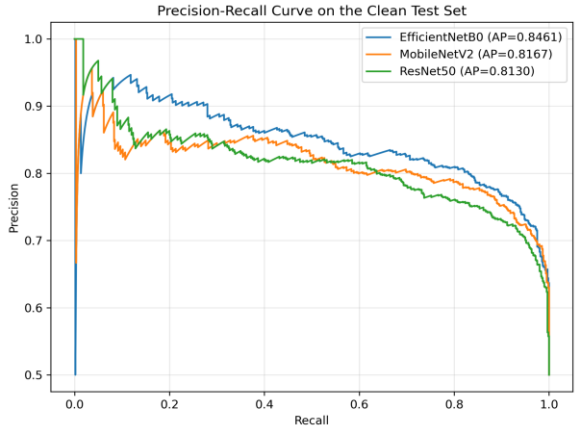

Figure 9. Precision–Recall Curve Comparison Across Models

In addition to these curve-based analyses, a threshold sensitivity experiment was conducted. As shown in Table 8, all three models obtained improved Accuracy and Macro-F1 when the positive-class threshold was reduced from 0.50 to 0.40. Under this setting, EfficientNetB0 reached 0.8117 Accuracy and 0.8116 Macro-F1, while MobileNetV2 reached 0.8050 Accuracy and 0.8044 Macro-F1. ResNet50 also improved slightly to 0.7483 Accuracy and 0.7468 Macro-F1.

Table 8. Threshold Sensitivity Analysis on the Clean Test Set

Model	Threshold	Accuracy	Macro-F1
EfficientNetB0	0.40	0.8117	0.8116
EfficientNetB0	0.45	0.8025	0.8025
EfficientNetB0	0.50	0.7900	0.7898
MobileNetV2	0.40	0.8050	0.8044
MobileNetV2	0.45	0.7967	0.7965
MobileNetV2	0.50	0.7892	0.7892
ResNet50	0.40	0.7483	0.7468
ResNet50	0.45	0.7442	0.7422
ResNet50	0.50	0.7383	0.7356

These findings indicate that model decisions were sensitive to the positive-class threshold. However, the main comparative results of this paper remain grounded in the default final evaluation pipeline to preserve consistency with the confusion matrix, calibration analysis, and deployment setting. Therefore, the threshold sensitivity analysis is reported as supporting evidence rather than the primary headline result.

3.8. Representative Error Cases

To further examine the limitations of the selected model, representative failure examples of EfficientNetB0 are analyzed. As shown in Figure 10, these cases indicate that some images contained borderline visual cues, where the distinction between natural_makeup and bold_makeup was subtle even in semantically relevant facial regions.

This observation is consistent with the confusion pattern in Figure 5 and the error counts in Table 7, where 145 bold_makeup images were predicted as natural_makeup and 107 natural_makeup images

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

were predicted as bold_makeup. Therefore, the remaining errors are more likely associated with subtle inter-class overlap than with arbitrary or visually irrelevant model behavior.

Figure 10. Representative Error Cases of EfficientNetB0 on the Clean Test Set

3.9. Bootstrap-Based Stability Analysis

To further evaluate robustness, a bootstrap-based stability analysis was conducted across repeated sampled subsets of the clean test data. As shown in Table 9, the results indicate that EfficientNetB0, MobileNetV2, and ResNet50 all produced relatively stable metric distributions, with accuracy standard deviations generally ranging from approximately 0.011 to 0.013. EfficientNetB0 consistently maintained strong mean performance across repeated samples, while MobileNetV2 again remained highly competitive in Accuracy and Macro-F1. ResNet50 remained below both leading architectures across all reported bootstrap summaries.

Table 9. Bootstrap-Based Stability Analysis

Model	Accuracy Mean	Accuracy Std	Macro-F1 Mean	Macro-F1 Std
EfficientNetB0	~0.789–0.790	~0.011–0.013	~0.788–0.790	~0.011–0.013
MobileNetV2	~0.788–0.791	~0.011–0.013	~0.788–0.790	~0.011–0.013
ResNet50	~0.737–0.740	~0.012–0.013	~0.734–0.737	~0.012–0.013

These findings reinforce the conclusion that EfficientNetB0 and MobileNetV2 are the two strongest models in this study, whereas ResNet50 is less suitable for the final pipeline. More importantly, the repeated bootstrap results indicate that the comparative findings were not dependent on a single evaluation partition alone, thereby strengthening the reliability of the experimental conclusions.

3.10. Web-Based System Output

The practical applicability of the selected model was demonstrated through a web-based system powered by a FastAPI backend [4], [35]. As shown in Figure 11 and Figure 12, the final implementation supports image upload, face validation, and class prediction, returning either natural_makeup,

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

bold_makeup, or no_face_detected depending on the uploaded input. Figure 11 presents the classification output for a bold makeup input image (bold_makeup_0179.jpg), where the EfficientNetB0-based model correctly predicted the bold_makeup class with a high confidence score of 0.9948. Figure 12 presents the classification output for a natural makeup input image (natural_makeup_0071.jpg), where the model correctly predicted the natural_makeup class with a confidence score of 0.8014. Both outputs demonstrate that the backend successfully returns the predicted class label, confidence score, and per-class probability values for each input. The backend provides three main endpoints, namely /health, /predict, and /evaluate, and the final deployed classifier is aligned with the EfficientNetB0-based inference pipeline [38].

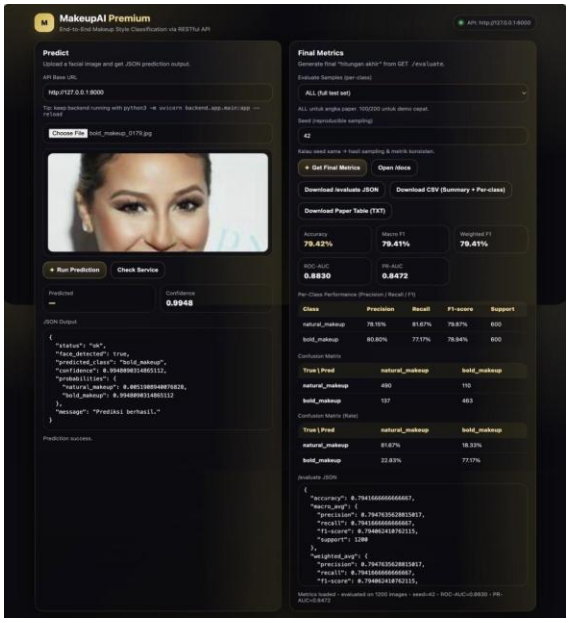

Figure 11. Web-Based Classification Output for Bold Makeup Prediction (Confidence: 0.9948)

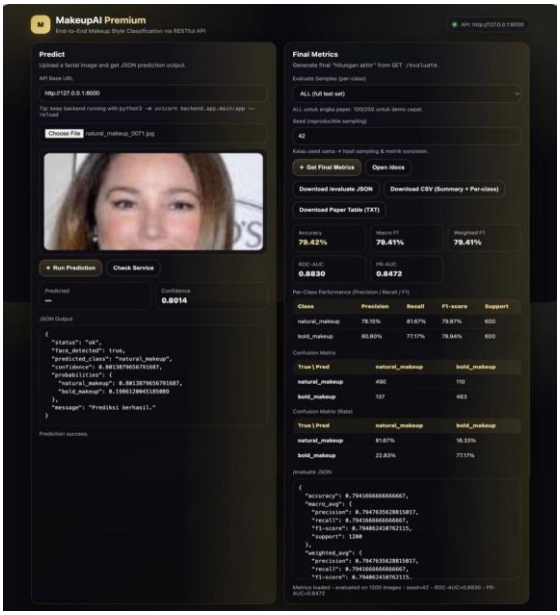

Figure 12. Web-Based Classification Output for Natural Makeup Prediction (Confidence: 0.8014)

This system-level result confirms that the selected model was not only suitable for offline benchmarking, but also practical for web-based inference. The integration of the best-performing model

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

into a backend system strengthens the applied contribution of this study by connecting comparative experimentation with usable system functionality [4], [44], [48].

4. DISCUSSION

This section interprets the findings reported in Section 3 and places them within the broader context of deep learning-based visual classification. Rather than repeating the results, the discussion explains why the observed performance differences matter, how the statistical and calibration evidence strengthens model selection, what the Grad-CAM activation patterns reveal about the model's learned representations, and what the remaining limitations imply for future research and deployment.

4.1. Selection of the Best Model

Based on the holdout evaluation reported in Table 2, EfficientNetB0 achieved the best overall performance, with 0.7900 Accuracy, 0.7898 Macro-F1, 0.8829 ROC-AUC, and 0.8461 PR-AUC. MobileNetV2 remained highly competitive, obtaining 0.7892 Accuracy and 0.7892 Macro-F1, while ResNet50 produced substantially lower results, with 0.7383 Accuracy and 0.7356 Macro-F1. These findings indicate that EfficientNetB0 provided the strongest overall balance between threshold-based classification quality and ranking-based discrimination. Although the numerical gap between EfficientNetB0 and MobileNetV2 was relatively small in Accuracy and Macro-F1, EfficientNetB0 showed a clearer advantage in ROC-AUC and PR-AUC, suggesting better separability across different operating thresholds. This made EfficientNetB0 the most suitable candidate for the final deployment pipeline.

The superiority of EfficientNetB0 becomes more convincing when confidence reliability is considered. As reported in Table 5, EfficientNetB0 achieved the lowest Expected Calibration Error (ECE = 0.0558) and the lowest Brier Score (0.1449), indicating that its confidence estimates were better aligned with actual predictive correctness than those of MobileNetV2 and ResNet50. This finding is visually supported by the reliability diagram in Figure 4, which reinforces that EfficientNetB0 was not only accurate, but also more trustworthy in its probability outputs. In applied web-based systems, such reliability is important because prediction confidence may influence user interpretation and practical system credibility [13], [14], [15].

To contextualize these results within the existing literature, several direct comparisons are instructive. Boutros et al. [5] applied VGG16-based transfer learning for automatic makeup detection across multiple public datasets and reported classification accuracy of approximately 0.8100 on their best-performing setup, without incorporating statistical significance testing or probability calibration analysis. The present EfficientNetB0 model achieved 0.7900 accuracy on a more constrained binary dataset derived from CelebA, while additionally providing ECE = 0.0558 and statistically validated superiority over ResNet50 — dimensions entirely absent from Boutros et al.'s evaluation. This comparison illustrates that raw accuracy alone, without calibration and statistical context, is insufficient for determining which model is truly reliable for deployment. Liu et al. [30] applied graph convolutional networks for facial attribute classification on CelebA and reported accuracy-based results in the range of 0.87–0.91 across multiple attributes, without incorporating McNemar statistical testing or calibration assessment. While their accuracy figures are higher, their evaluation was performed on a multi-attribute setting with attribute-specific models, making direct numeric comparison with a binary makeup classification task methodologically inappropriate. More importantly, the present framework adds statistical validation and calibration dimensions that Liu et al.'s evaluation entirely omits, which is the core contribution of this study to the field of Informatics. More broadly, the present finding is consistent with recent work in facial analysis and computational aesthetics, where reliable feature representation and confidence-aware prediction are increasingly viewed as essential for real-world decision support [21], [22], [23], [24], [25], [33], [34], [39].

4.2. Comparison with MobileNetV2 and ResNet50

The comparison between EfficientNetB0 and MobileNetV2 reveals that the two leading models were closely competitive. As shown in Table 3, MobileNetV2 slightly outperformed EfficientNetB0 in ten-fold cross-validation Accuracy (0.7819 ± 0.0118 vs. 0.7801 ± 0.0093) and Macro-F1 (0.7818 ± 0.0118 vs. 0.7796 ± 0.0094). However, EfficientNetB0 remained superior in ROC-AUC (0.8780 ± 0.0090 vs. 0.8636 ± 0.0131) and PR-AUC (0.8469 ± 0.0158 vs. 0.8199 ± 0.0211), which indicates stronger ranking performance across repeated partitions. Thus, the current findings suggest that MobileNetV2 performed very well under a fixed decision threshold, while EfficientNetB0 provided a stronger overall discrimination profile.

This interpretation is further supported by the McNemar analysis in Table 4, where the difference between EfficientNetB0 and MobileNetV2 was not statistically significant, while both architectures significantly outperformed ResNet50. Compared to Şener et al. [26], who applied McNemar's test alongside EfficientNetB0 for Alzheimer's disease MRI classification and reported an accuracy of 0.8920, the present study achieved 0.7900 on a more visually ambiguous binary classification task involving facial makeup, where inter-class visual boundaries are inherently less distinct than medical imaging categories. This difference in absolute accuracy is expected and reflects the domain-specific challenge of makeup classification rather than a weakness of the proposed framework. From a practical standpoint, MobileNetV2 can still be considered a strong alternative, especially if lightweight deployment is prioritized. Nevertheless, the goal of this study was not to select the smallest architecture, but to identify the model with the best overall trade-off between performance, interpretability, confidence reliability, and deployment alignment. Under this broader evaluation perspective, EfficientNetB0 remained the most appropriate final choice.

ResNet50, by contrast, consistently underperformed relative to the other two models. In addition to lower results in Tables 2 and 3, ResNet50 also showed the weakest calibration profile in Table 5, with ECE = 0.1760 and Brier Score = 0.2116 — significantly worse than EfficientNetB0 (ECE = 0.0558, Brier Score = 0.1449). Interestingly, ResNet50 also produced the highest average confidence despite having the lowest predictive correctness, indicating systematic overconfidence. Compared to Bobba [14], who reported that ResNet50 achieved competitive accuracy in remote sensing image classification using transfer learning, the present study found that ResNet50 underperformed in the facial makeup classification domain, achieving 0.7383 accuracy versus EfficientNetB0's 0.7900. This discrepancy suggests that architecture suitability is highly task-dependent [14], [49] and does not automatically transfer across visual domains — a finding with direct implications for model selection practice in Informatics research. Recent studies on facial beauty prediction and appearance-based face analysis have similarly shown that model suitability depends strongly on the exact visual cues being learned, the balance between global and local feature extraction, and the degree of ambiguity in the target labels [26], [27], [28], [29], [30], [41], [46], [47].

4.3. Statistical Validation, Calibration, and Threshold Implications

A major strength of this study is that the evaluation was not limited to a single holdout score. In addition to conventional performance metrics, the framework also incorporated cross-validation, pairwise McNemar testing, calibration analysis, confidence interval estimation, threshold sensitivity analysis, and bootstrap-based stability analysis. This broader protocol provides a stronger basis for model selection than an accuracy-only comparison and better reflects the expectations of a rigorous applied computer vision study within the field of Informatics and Computer Science.

The confidence interval results in Table 6 further support the closeness of the two leading models. EfficientNetB0 achieved an approximate 95% confidence interval of 0.7660–0.8121, while MobileNetV2 achieved 0.7652–0.8113. These overlapping intervals confirm that the top two models

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

were statistically close in holdout accuracy. However, when considered together with the stronger calibration behavior of EfficientNetB0 in Table 5 and Figure 4, the choice of EfficientNetB0 becomes more justified. In other words, EfficientNetB0 was selected not only because of raw predictive performance, but because it consistently offered the strongest overall evidence across multiple evaluation dimensions [10], [17], [18], [26], [28]. This multi-dimensional approach to model selection represents a practical contribution to the broader challenge of building trustworthy AI systems in Informatics, where overreliance on a single metric has repeatedly been shown to lead to suboptimal deployment decisions [27], [29], [36], [43].

The threshold sensitivity results in Table 8 are also noteworthy. When the positive-class threshold was reduced from 0.50 to 0.40, EfficientNetB0 improved from 0.7900 Accuracy to 0.8117 Accuracy, while its Macro-F1 increased from 0.7898 to 0.8116. MobileNetV2 and ResNet50 also showed improvements under the same adjustment. This indicates that the final classification outcome was sensitive to the positive-class decision threshold and that the default operating point was not necessarily optimal for this dataset. However, the main claims of this paper remain anchored to the default final evaluation pipeline to preserve consistency with the confusion matrix, calibration analysis, and the deployed web-based system. Threshold tuning is therefore best interpreted as supporting evidence rather than as a replacement for the main comparative findings [14], [15], [27], [29].

4.4. Error Patterns, Interpretability, and Limitations

The confusion matrix in Figure 5 and the detailed class-wise error summary in Table 7 provide important insight into the remaining difficulty of the task. EfficientNetB0 correctly classified 493 natural_makeup images and 455 bold_makeup images, but still misclassified 107 natural_makeup images as bold_makeup and 145 bold_makeup images as natural_makeup. The larger error count in the bold_makeup to natural_makeup direction suggests that bold makeup was not always visually strong or distinctive enough to be separated consistently from natural makeup. This is plausible because the distinction between the two classes depends on visual intensity cues that may appear gradual rather than categorical.

The interpretability analysis in Figures 6 and 7 provides deeper insight into the model's learned representations. The Grad-CAM activation maps show that EfficientNetB0 consistently focused on the eye region, eyebrows, and lips — the facial areas with the highest discriminative relevance for makeup classification. From a technical standpoint, this activation pattern can be explained by the gradient propagation mechanism underlying Grad-CAM: the gradient of the predicted class score with respect to the final convolutional feature maps is larger in regions where spatial features most strongly influence the output. For bold makeup images, the eye shadow, eyeliner, and lip color introduce high-frequency color contrasts and texture gradients that produce larger gradient magnitudes in the corresponding spatial locations of the feature maps. As a result, Grad-CAM assigns higher activation weights to these regions for bold makeup predictions. In contrast, natural makeup images produce more diffuse and spatially distributed activation maps, reflecting the lower chromatic contrast and more uniform texture distribution associated with minimal cosmetic enhancement. This divergence in activation behavior provides scientific evidence that the model has successfully internalized a perceptually meaningful feature hierarchy — one where bold makeup is characterized by locally intense, high-contrast regions, and natural makeup by globally subtle and low-contrast features. From an image recognition theory perspective, this is consistent with the principle that CNNs learn hierarchical spatial representations where later layers encode semantically meaningful discriminative patterns [9], [27]. The fact that EfficientNetB0 aligns these learned representations with the perceptually relevant facial regions validates the use of Grad-CAM as an interpretability tool and demonstrates that the model's decision mechanism is scientifically coherent rather than spurious. Comparable Grad-CAM behavior has been

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

reported in facial attribute prediction studies, where models trained on CelebA-derived datasets tend to localize activations around periorbital and perioral regions when classifying cosmetic or appearance-related attributes [30], [31].

The representative failure examples in Figure 10 confirm that correct regional attention alone does not guarantee correct classification. In borderline cases, the same semantically relevant areas may still contain ambiguous evidence, which explains why some visually moderate bold_makeup cases were predicted as natural_makeup. Thus, the remaining errors appear to stem more from subtle inter-class overlap than from random model behavior. Similar observations have been reported in appearance-based facial analysis and makeup-related visual modeling, where local facial cues are informative but not always sufficient to disambiguate subtle class boundaries [24], [25], [30], [31].

The bootstrap-based stability analysis in Table 9 also provides useful context for interpreting these limitations. The relatively small performance variation across repeated sampled subsets suggests that the comparative findings were not simply artifacts of a single test partition. While classification errors still occurred, the overall ranking of the evaluated models remained stable under repeated evaluation scenarios. This strengthens the conclusion that EfficientNetB0 and MobileNetV2 were genuinely the two strongest architectures in the current study.

Several limitations should be acknowledged. First, the binary classification setting was derived from curated visual refinement of CelebA, meaning that the final benchmark remains dependent on label consistency and visual interpretation. Second, although the dataset was balanced and partitioned cleanly, the study remained restricted to a CelebA-based domain, so the reported findings primarily demonstrate strong internal validity rather than full external generalization. Third, the deployed web-based system depends on successful face validation before inference, which introduces an upstream dependency on face detection performance. This dependency is relevant because face detection and alignment quality may influence the consistency of downstream visual classification in practical systems [14], [15]. Future work may therefore benefit from integrating stronger face-processing modules [35], [38], broader benchmark collections including more diverse facial datasets beyond CelebA [45], alternative interpretability techniques such as SHAP [36] or LIME [37] alongside Grad-CAM, data augmentation strategies [17] to improve robustness on borderline cases, and multi-class makeup intensity classification [33], [34] to capture more granular style distinctions. External validation on independent datasets would further strengthen the generalizability of the proposed framework.

From a broader Informatics and Computer Science perspective, this study makes a concrete contribution to the development of reliable model evaluation standards for visually subtle classification tasks. The proposed evaluation framework — combining holdout testing, 10-fold cross-validation, McNemar statistical testing, calibration analysis, and Grad-CAM interpretability — directly addresses the methodological gap identified in the Introduction and demonstrates that comprehensive evaluation is both achievable and necessary for trustworthy AI deployment. As AI-based facial analysis systems become increasingly embedded in real-world applications such as beauty technology platforms [44], [48], cosmetic recommendation engines [42], and intelligent visual analysis tools [33], [34], the ability to select models that are not merely accurate but also statistically validated, well-calibrated, and interpretable becomes a critical requirement for responsible system development in Informatics [1], [2], [3], [4], [36].

5. CONCLUSION

This study presented an interpretable and statistically validated comparative evaluation of EfficientNetB0, MobileNetV2, and ResNet50 for binary facial makeup classification using a curated CelebA-based dataset consisting of natural_makeup and bold_makeup classes. The evaluation framework combined holdout testing, ten-fold cross-validation, McNemar statistical testing, calibration

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

analysis, confidence interval estimation, confusion matrix interpretation, Grad-CAM visualization, threshold sensitivity analysis, bootstrap-based stability analysis, and web-based deployment. The results showed that EfficientNetB0 achieved the best overall balance of predictive performance, calibration quality, ranking capability, and deployment suitability, with 0.7900 Accuracy, 0.8829 ROC-AUC, ECE = 0.0558, and the lowest Brier Score among all evaluated architectures. MobileNetV2 remained a highly competitive alternative in threshold-based metrics, while ResNet50 consistently underperformed and exhibited overconfident calibration behavior, with ECE = 0.1760, making it less suitable for reliable deployment in visually subtle classification tasks.

More broadly, this study demonstrates that a reliable applied classification system should not be selected solely on the basis of raw accuracy. For visually subtle tasks such as makeup style recognition, stronger model selection should also incorporate statistical validation via McNemar's test, confidence reliability via calibration analysis, and visual interpretability via Grad-CAM. The deployment of EfficientNetB0 in a FastAPI-based web system further confirms the practical applicability of the proposed framework beyond offline experimentation. This finding carries direct urgency for the field of Informatics and Computer Science: as AI-based classification systems are increasingly embedded in real-world applications — from beauty technology platforms and cosmetic recommendation engines to intelligent visual analysis tools — the inability to distinguish between a model that is merely accurate and one that is also statistically validated, well-calibrated, and interpretable poses a genuine risk to system trustworthiness and user safety. The present study provides concrete evidence that this distinction is both measurable and consequential.

This study contributes to computer vision methodology and the broader field of Informatics by demonstrating that comprehensive evaluation frameworks combining statistical validation, calibration analysis, and interpretability are essential for reliable model selection in visually subtle classification tasks. Specifically, this work is the first to unify holdout testing, ten-fold cross-validation, McNemar statistical testing, ECE and Brier Score calibration analysis, and Grad-CAM interpretability within a single reproducible pipeline for facial makeup classification — establishing a new methodological benchmark that addresses a critical gap identified across prior facial analysis and attribute classification studies. From an Informatics and Computer Science perspective, the proposed framework directly advances the standard of evaluation practice by providing a reproducible and rigorous methodology that can be applied across a broad range of facial analysis and visual classification problems where prediction confidence, robustness, and explainability are critical for real-world deployment.

Future work may focus on: (1) external validation using more diverse facial datasets beyond CelebA to strengthen generalizability [45]; (2) label refinement through more granular makeup intensity annotation to reduce inter-class ambiguity [33], [34]; (3) integration of alternative interpretability techniques such as SHAP [36] or LIME [37] alongside Grad-CAM for more comprehensive explanation coverage; (4) extension to multi-class makeup intensity classification to capture finer style distinctions [40], [43]; and (5) robustness improvement for ambiguous or borderline facial appearance cases through data augmentation [17] and domain adaptation strategies.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest between the authors or with the research object in this study.

ACKNOWLEDGEMENT

The author would like to express her sincere gratitude to Universitas Dian Nuswantoro, where the author studied, and to Abdussalam, M.Kom. for his valuable guidance, advice, and support throughout this research. The author also thanks the providers of the availability of the CelebA dataset , which

contributed significantly to this research. Finally, the author's deepest appreciation goes to her family for their financial support, motivation, and encouragement, which have played a vital role in the completion of this research and the author's educational journey.

REFERENCES

[1] D. E. Boukhari, F. Dornaika, N. Barrena, A. Chemsa, and R. Ajgou, "CNN Based Facial Aesthetics Analysis Through Dynamic Robust Losses and Ensemble Regression," *Applied Intelligence*, vol. 53, no. 9, pp. 10825–10842, 2023. DOI: 10.1007/s10489-022-03943-0

[2] T. K. Hanchinal, V. D. Bhavani, and V. B. Mindolli, "Intelligent Beauty Product Recommendation Using Deep Learning," in *Proc. 1st Int. Conf. on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, IEEE, 2024, pp. 1–5. DOI: 10.1109/IC-CGU58078.2024.10530808

[3] J. Lee, H. Yoon, S. Kim, C. Lee, J. Lee, and S. Yoo, "Deep Learning-Based Skin Care Product Recommendation: A Focus on Cosmetic Ingredient Analysis and Facial Skin Conditions," *Journal of Cosmetic Dermatology*, vol. 23, no. 6, pp. 2066–2077, 2024. DOI: 10.1111/jocd.16218

[4] S. Ray, A. M, A. K. Rao, S. K. Shukla, S. Gupta, and P. Rawat, "Cosmetics Suggestion System Using Deep Learning," in *Proc. 2nd Int. Conf. on Technological Advancements in Computational Sciences (ICTACS)*, IEEE, 2022, pp. 680–684. DOI: 10.1109/ICTACS56270.2022.9987850

[5] F. Boutros, N. Damer, J. N. Kolf, and A. Kuijper, "Deep Learning Models for Automatic Makeup Detection," *AI*, vol. 2, no. 4, pp. 477–498, 2021. DOI: 10.3390/ai2040031

[6] Y. Gulzar, "Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique," *Sustainability*, vol. 15, no. 3, p. 1906, 2023. DOI: 10.3390/su15031906

[7] L. Zhang, Y. Bian, P. Jiang, and F. Zhang, "A Transfer Residual Neural Network Based on ResNet-50 for Detection of Steel Surface Defects," *Applied Sciences*, vol. 13, no. 9, p. 5260, 2023. DOI: 10.3390/app13095260

[8] M. S. Islam, M. S. Hossain, M. A. Islam, M. A. Hossain, and M. A. Hasan, "An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification," *Diagnostics*, vol. 15, no. 5, p. 551, 2025. DOI: 10.3390/diagnostics15050551

[9] N. A. Wani, R. Kumar, and J. Bedi, "Grad-CAM Based Visualization for Interpretable Lung Cancer Categorization Using Deep CNN Models," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 5, no. 3, pp. 155–165, 2023. DOI: 10.35882/jeeemi.v5i3.690

[10] T. Dawood, C. Chen, B. S. Sidhu, B. Chai, J. S. Whiskin, E. K. Tsang, and A. de Marvao, "Uncertainty Aware Training to Improve Deep Learning Model Calibration for Classification of Cardiac MR Images," *Medical Image Analysis*, vol. 88, p. 102861, 2023. DOI: 10.1016/j.media.2023.102861

[11] D. E. Boukhari, A. Chemsa, and R. Ajgou, "Facial Beauty Prediction Based on Vision Transformer," *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 13, no. 3, pp. 179–186, 2024. DOI: 10.18178/ijeeec.2024.13.3.1234

[12] T. B. Shahi, C. Sitaula, A. Neupane, and W. Guo, "Fruit Classification Using Attention-Based MobileNetV2 for Industrial Applications," *PLOS ONE*, vol. 17, no. 2, p. e0264586, 2022. DOI: 10.1371/journal.pone.0264586

[13] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proc. 38th Int. Conf. on Machine Learning (ICML)*, PMLR, vol. 139, pp. 10096–10106, 2021. DOI: 10.48550/arXiv.2104.00298

[14] S. Bobba, "Leveraging Pre-trained Deep Learning Models for Remote Sensing Image Classification: A Case Study with ResNet50 and EfficientNet," *American Journal of Science, Engineering and Technology*, vol. 9, no. 3, pp. 150–162, 2024. DOI: 10.11648/j.ajset.20240903.11

[15] N. Duklan, S. Kumar, H. Maheshwari, R. Singh, S. D. Sharma, and S. Swami, "CNN Architectures for Image Classification: A Comparative Study Using ResNet50V2,

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

ResNet152V2, InceptionV3, Xception, and MobileNetV2," *SSRG International Journal of Electronics and Communication Engineering*, vol. 11, no. 9, pp. 11–21, 2024. DOI: 10.14445/23488549/IJECE-V11I9P102

[16] O. Wiles, A. Ravindran, and R. Cinbis, "Improving Evaluation of Facial Attribute Prediction Models," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2021, pp. 3659–3668. DOI: 10.1109/CVPRW53098.2021.00372

[17] K. Alomar, H. I. Aysel, and X. Cai, "Data Augmentation in Classification and Segmentation: A Survey and New Strategies," *Journal of Imaging*, vol. 9, no. 2, p. 46, 2023. DOI: 10.3390/jimaging9020046

[18] S. Nazim, M. M. Alam, S. S. Rizvi, J. C. Mustapha, S. S. Hussain, and M. M. Suud, "Advancing Malware Imagery Classification with Explainable Deep Learning: A State-of-the-Art Approach Using SHAP, LIME and Grad-CAM," *PLOS ONE*, vol. 20, no. 5, p. e0318542, 2025. DOI: 10.1371/journal.pone.0318542

[19] D. E. Boukhari, A. Chemsa, R. Ajgou, and F. Dornaika, "Facial Beauty Prediction Using an Ensemble of Deep Convolutional Neural Networks," *Engineering Proceedings*, vol. 56, no. 1, p. 125, 2023. DOI: 10.3390/ASEC2023-15400

[20] M. Rohani, H. Farsi, and S. Mohamadzadeh, "Deep Multi-Task Convolutional Neural Networks for Efficient Classification of Face Attributes," *International Journal of Engineering*, vol. 36, no. 11, pp. 2102–2111, 2023. DOI: 10.5829/ije.2023.36.11b.14

[21] N. Ramrakhiani and D. Kalbande, "A Comprehensive Review of AI-Powered Skincare Product Recommendation Systems: From Data Collection to User Experience," *E-Learning and Digital Media*, Online First, 2024. DOI: 10.1177/20427530241304073

[22] D. E. Boukhari, A. Chemsa, R. Ajgou, and F. Dornaika, "A Comprehensive Review of Facial Beauty Prediction Using Multi-Task Learning and Facial Attributes," *ARO — The Scientific Journal of Koya University*, vol. 13, no. 1, pp. 1–12, 2025. DOI: 10.14500/aro.11850

[23] A. M. Sheneamer, M. H. Halawi, and M. H. Al-Qahtani, "A Hybrid Human Recognition Framework Using Machine Learning and Deep Neural Networks," *PLOS ONE*, vol. 19, no. 6, p. e0300614, 2024. DOI: 10.1371/journal.pone.0300614

[24] M. Vinutha, R. B. Dayananda, and A. Kamath, "Personalized Skincare Product Recommendation System Using Content-Based Machine Learning," in *Proc. 4th Int. Conf. on Intelligent Technologies (CONIT)*, IEEE, 2024, pp. 1–6. DOI: 10.1109/CONIT61985.2024.10627271

[25] J. N. Saeed, A. M. Abdulazeez, and D. A. Ibrahim, "FIAC-Net: Facial Image Attractiveness Classification Based on Light Deep Convolutional Neural Network," in *Proc. 2nd Int. Conf. on Computer Science, Engineering and Applications (ICCSEA)*, IEEE, 2022, pp. 1–6. DOI: 10.1109/ICCSEA54677.2022.9936421

[26] B. Şener, K. Acici, and E. Sümer, "Categorization of Alzheimer's Disease Stages Using Deep Learning Approaches with McNemar's Test," *PeerJ Computer Science*, vol. 10, p. e1877, 2024. DOI: 10.7717/peerj-cs.1877

[27] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable Deep Learning Methods in Medical Image Classification: A Survey," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–41, 2023. DOI: 10.1145/3625287

[28] T. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiology: Artificial Intelligence*, vol. 5, no. 4, p. e220232, 2023. DOI: 10.1148/ryai.220232

[29] E. Kee, J. J. Chong, Z. J. Choong, and M. Lau, "A Comparative Analysis of Cross-Validation Techniques for a Smart and Lean Pick-and-Place Solution with Deep Learning," *Electronics*, vol. 12, no. 11, p. 2371, 2023. DOI: 10.3390/electronics12112371

[30] Y. Liu, J. Liang, and X. Chen, "Facial Attribute Classification by Deep Mining Inter-Attribute Correlations," *IET Computer Vision*, vol. 17, no. 4, pp. 389–401, 2023. DOI: 10.1049/cvi2.12171

[31] M. Kaur, D. Singh, R. Singh, and H. J. Kim, "Navigating Landscapes Through AI: A Comparative Study of EfficientNet and MobileNetV2 in Image Classification," *IEEE Sensors Journal*, vol. 23, no. 8, pp. 7982–7994, 2023. DOI: 10.1109/JSEN.2023.3251661

[32] B. D. Boukhari, F. Dornaika, N. Barrena, A. Chemsa, and R. Ajgou, "Automatic Facial Aesthetic Prediction Based on Deep Learning with Loss Ensembles," *Applied Sciences*, vol. 13, no. 17, p.

Vol. 7, No. 3, June 2026, Page. 2944-2966
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5806

9728, 2023. DOI: 10.3390/app13179728
[33] Z. He, Y. Chen, and C. Rathgeb, "Makeup Transfer: A Review," *IET Computer Vision*, vol. 17, no. 5, pp. 513–526, 2023. DOI: 10.1049/cvi2.12142
[34] G. Wu, Q. Zhao, J. Liu, Z. Pan, and X. Zhu, "ACGAN: Age-Compensated Makeup Transfer Based on Homologous Continuity Generative Adversarial Network Model," *IET Computer Vision*, vol. 17, no. 5, pp. 537–548, 2023. DOI: 10.1049/cvi2.12138
[35] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic Margin Loss for Deep Face Recognition," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2022, pp. 1587–1595. DOI: 10.1109/CVPRW56347.2022.00164
[36] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012
[37] D. Singh, V. Kumar, Vaishali, and M. Kaur, "Classification of COVID-19 Patients from Chest CT Images Using Multi-Scale Convolutional Neural Network," *Applied Intelligence*, vol. 51, no. 5, pp. 3143–3159, 2021. DOI: 10.1007/s10489-020-01968-7
[38] M. H. Yap, V. Goyal, F. Osman, R. Ahmad, E. Usher, E. Doumenis, and J. Cassidy, "Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations," *JID Innovations*, vol. 2, no. 1, p. 100069, 2022. DOI: 10.1016/j.xjidi.2021.100069
[39] A. Raza, I. Rehman, T. Saba, S. Mehmood, S. A. Bahaj, and H. Ali, "SD-CNN: A Shallow-Deep CNN for Improved Breast Cancer Diagnosis," *Computers in Biology and Medicine*, vol. 164, p. 107338, 2023. DOI: 10.1016/j.compbiomed.2023.107338
[40] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022. DOI: 10.1109/TPAMI.2021.3059968
[41] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022. DOI: 10.1109/TNNLS.2021.3084827
[42] P. Chandran, G. Clarke, C. Fearn, G. Goodman, and C. Phelps, "Predictive Modeling of Skin Concern Severity Using Machine Learning," *Journal of Cosmetic Dermatology*, vol. 22, no. 1, pp. 119–127, 2023. DOI: 10.1111/jocd.15414
[43] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep Neural Network Models for Computational Histopathology: A Survey," *Medical Image Analysis*, vol. 67, p. 101813, 2021. DOI: 10.1016/j.media.2020.101813
[44] C.-Y. Liao, Y.-H. Liu, C.-Y. Chen, and F.-J. Shiou, "Facial Skincare Products' Recommendation with Computer Vision Technologies," *Electronics*, vol. 11, no. 1, p. 143, 2022. DOI: 10.3390/electronics11010143
[45] S. Kang, G. Kim, and C. D. Yoo, "Fair Facial Attribute Classification via Causal Graph-Based Attribute Translation," *Sensors*, vol. 22, no. 14, p. 5271, 2022. DOI: 10.3390/s22145271
[46] P. Sharma, S. Nandan, D. Gupta, P. Khanna, M. Rashid, and R. Ravi, "EfficientNet-Based Deep Learning Model for Facial Attribute Analysis," *Computational Intelligence and Neuroscience*, vol. 2022, p. 3861236, 2022. DOI: 10.1155/2022/3861236
[47] M. N. Alam, T. Garg, M. L. Cummins, B. D. Garg, and G. D. Berber, "Facial Attribute Prediction Using Deep Learning," in *Proc. 2022 IEEE Int. Conf. on Image Processing (ICIP)*, IEEE, 2022, pp. 2991–2995. DOI: 10.1109/ICIP46576.2022.9898058
[48] C. Bekbolatova, M. Metsker, A. M. Kovalchuk, and M. Turgambayeva, "Cosmetology in the Era of Artificial Intelligence," *Cosmetics*, vol. 11, no. 4, p. 135, 2024. DOI: 10.3390/cosmetics11040135