The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 3, June 2025
https://doi.org/10.33022/ijcs.v14i3.4636

Variability in Makeup and Expressions: Impacts on Deep Learning Classifiers for Face Recognition

Egwali Annie[1], Sule Winifred[2]
annie.egwali@uniben.edu[1], ruthsule22@uniben.edu[2]
[1,2] University of Benin, Benin City, Edo State

Article Information

Received : 11 Jan 2025
Revised : 7 Jun 2025
Accepted : 16 Jun 2025

Keywords

Facial Recognition,
Makeup,
Deep Learning Classifiers,
Facial Expressions,
Transfer Learning

Abstract

Facial recognition technology serves as an integral component of security, access management, and identification systems. This study addresses the challenges this technology faces due to makeup and varying facial expressions, which can lead to misidentification. We investigate the effectiveness of five deep learning models—ResNet, InceptionV3, EfficientNet, Xception, and SENet—in recognizing faces with makeup and diverse emotional expressions. Using five publicly accessible datasets, including KDEF, CelebA, and UTKFace, we measure performance with metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. Our analysis evaluates the benefits of transfer learning with pre-trained models and their robustness against new data. We find that InceptionV3 achieves peak accuracy of 85.2% on CelebA with high performance across all datasets, with an average accuracy of 79.8%. These results highlight how makeup and emotional expressions affect recognition accuracy and emphasize the need for improving facial recognition technologies for security and accessibility applications.

A. Introduction

The past few years have witnessed a dramatic increase in the development and utilization of technologies related to facial recognition, detection, and analysis. Presently, around 80% of countries across the globe employ facial analysis algorithms [1]. This technological advancement has redefined applications, ranging from security systems and access control to filters used in social media platforms [2]. Nevertheless, several factors that change facial appearance, notably makeup, can severely impair the performance of these systems. The application of makeup alters critical features such as skin texture and contrast around the eyes and mouth, as well as the perceived shape of the face.

Consequently, facial recognition systems may encounter difficulties in recognizing individuals who use makeup, leading to potential false non-matches. The backbone of facial recognition technology is grounded in machine learning frameworks, particularly in deep learning (DL)—a specialized branch of machine learning that is part of artificial intelligence [3]. Deep learning has garnered attention for its applications across various fields, including healthcare, visual recognition, text processing, and cybersecurity. The effectiveness of differing machine learning methodologies varies based on the stage of the facial recognition process in which they are implemented. This variance is highlighted by the unique challenges inherent in image analysis, computer vision, and cybersecurity due to the dynamic characteristics of human facial features.

Notably, there are expected discrepancies in classifier performance when applied to diverse facial datasets, particularly those containing both made-up and bare faces. Makeup has become a routine practice for many, altering individuals' facial characteristics and creating challenges for facial recognition technologies. Furthermore, facial expressions compound these challenges, introducing further variability [4]. The widespread use of makeup complicates the task of achieving accurate and dependable facial recognition, as it can significantly change facial attributes such as color, texture, and shape, negatively impacting the effectiveness of recognition algorithms [5].

Various publicly accessible datasets featuring faces adorned with makeup have emerged, frequently focused on measuring or enhancing facial attractiveness. Makeup application can be delineated by intensity; light makeup may be relatively unnoticeable as the colors typically mimic natural skin tones, whereas heavy makeup is more conspicuous, featuring elements like bold lip colors or heavily applied eye makeup. Such stylistic differences can result in profound changes in facial appearance and serve as effective means of evading recognition systems, as evidenced by findings that makeup usage can sharply decrease face-matching accuracy [6]. In a study by [7], it was determined that recognition accuracy for both commercial and academic face recognition techniques could drop as much as 76.21% due to the presence of makeup. Motivated by the persistent challenges confronting facial classification algorithms despite advancements, various studies have sought to scrutinize the performance of different classifiers [8].

Facial expressions introduce an additional layer of complexity as they can also influence the performance of classifiers. The suitability of a facial database is crucial for any classifier's success [9]. This raises essential questions about the transparency and accountability of the algorithms. As a result, ongoing research endeavors seek to determine which facial classifiers are most effective considering the datasets employed. This study focuses specifically on evaluating the robustness of facial recognition classifiers against modifications caused by makeup application and varying facial expressions, utilizing popular facial datasets.

Previous research has examined how makeup influences facial recognition accuracy and explored various classifiers' effectiveness [5] [10][11]. Nevertheless, the limited availability of makeup-specific public datasets constrains the ability to train and evaluate robust makeup recognition models. Prior analyses have typically employed controlled expressions which may not adequately capture real-world dynamics characterized by varying emotional expressions. Additionally, the potential benefits of leveraging pre-trained models on extensive facial recognition datasets in the context of makeup recognition are yet to be thoroughly investigated [12] [13] [14].

The choice of database is critical; the success of a technique for a particular problem relies on the determination of an appropriate dataset. When contrasting outcomes generated by different methodologies addressing similar issues, consistency in dataset usage is crucial. Currently, while many face databases exist for public use, few are explicitly focused on makeup-related research [15] [16] [17] [18] [19] [20]. This research intends to bridge these gaps by systematically assessing five deep learning models aimed at makeup recognition across diverse facial expressions. By identifying the most effective model architectures and evaluating their performance through various metrics, this study aspires to provide valuable insights into the capacity of deep learning models to manage the combined effects of makeup and emotional expression on facial recognition.

This paper underscores the significance of makeup application and facial expressions in the realm of facial recognition technologies, emphasizing the necessity for advanced, adaptable algorithms. By scrutinizing the performance of five deep learning models, the research aims to yield insights regarding performance metrics that could influence future advancements in security applications, along with proposing potential enhancements to existing algorithms to bolster their efficacy in practical scenarios.

B. Diverse Makeup Techniques

Makeup application can significantly transform facial characteristics, creating substantial challenges for facial recognition systems. The use of cosmetics can alter the shape and color of features such as eyebrows, eyelashes, eyes, lips, and overall skin tone [9]. These modifications interrupt the extraction of consistent facial features, which traditional recognition algorithms depend upon. For instance, eyeliner application can enhance and elongate the appearance of eyelashes, thereby changing the shape of the eye region. Similarly, lipstick can dramatically change the size and shape of the lips. As a result, these variations can lead to pronounced mismatches when comparing makeup-altered faces against enrolled templates captured without makeup [21].

In addition to makeup, facial expressions contribute to dynamic changes in facial features, presenting further challenges to recognition systems. When individuals express emotions, their facial muscles engage in a process of contraction and relaxation, which produces wrinkles, furrows, and changes in the placements of critical landmarks such as the eyes, eyebrows, and mouth corners. These variations can drastically alter the appearance of specific facial regions, complicating the ability of recognition algorithms to match a neutral face template with an image taken during an expressive moment [22].

Lighting conditions are another significant factor that can affect how facial features appear, ultimately impacting recognition performance [23]. Variations in lighting intensity, direction, and color can cast shadows or create highlights on the facial surface, affecting the perceived depth and shape of individual features. Many traditional facial recognition algorithms struggle to adapt to these lighting changes, which can lead to recognition errors, particularly when there are considerable differences between lighting conditions during enrollment and recognition.

Head pose variations further exacerbate the problem by introducing occlusions and distortions, making the recognition process more challenging [24]. When individuals tilt their heads or change their gaze, certain facial regions may become occluded or distorted due to altered perspectives. This variability can significantly limit the amount of usable information available for recognition, ultimately leading to decreased performance.

C. Makeup Options Datasets

To tackle the challenges of achieving accurate facial recognition in real-world scenarios, it is imperative for researchers to develop deep learning models capable of effectively managing variations caused by makeup. Several notable datasets, including CelebA, KDEF (Karolinska Directed Emotional Faces), and UTKFace, are essential resources for training and validating these models.

The CelebA dataset contains over 200,000 images of celebrities, annotated with 40 distinct facial descriptors. This expansive collection allows researchers to investigate different makeup styles and their influence on recognition accuracy [25]. The KDEF dataset, on the other hand, offers a robust compilation of emotional expressions, facilitating inquiries into how emotions correlate with recognition performance [26]. Meanwhile, UTKFace presents a diverse range of demographic representations, which is crucial for the development of inclusive and equitable facial recognition technologies [27]. Integrating these varied datasets is vital for constructing adaptive facial recognition algorithms that sustain accuracy across different demographic groups, emotional expressions, and makeup styles.

D. Deep Learning for Facial Recognition in Makeup

Deep learning for makeup-aware face recognition improves the accuracy of facial recognition systems when individuals wear makeup, which significantly changes their appearance. Advanced neural network architectures enable these systems to recognize and adapt to variations introduced by different makeup styles. Techniques such as transfer learning and data augmentation enhance model performance, allowing models to generalize better across diverse makeup applications and real-world conditions. Additionally, integrating emotional expression analysis further refines recognition accuracy. Despite challenges like variability in makeup and the need for comprehensive datasets, applying deep learning techniques shows great promise for developing more effective and reliable makeup-aware face recognition systems in various fields, including security and social media.

1. InceptionV3

InceptionV3 is a deep convolutional neural network that effectively captures various features through its unique architecture, which includes multiple convolutional paths within each Inception module. This design allows for varied kernel sizes (1x1, 3x3, and 5x5) to simultaneously extract features at different scales while maintaining computational efficiency. For a single Inception block, the output can be expressed as shown in equations 1 to 5:

(i) **Branch 1** (1x1 Convolution):

$$B1 = \mathrm{Conv}(1\times1, \mathrm{filters}_1)(X) \qquad (1)$$

(ii) **Branch 2** (1x1 followed by 3x3 Convolution):

$$B2 = \mathrm{Conv}(1\times1, \mathrm{filters}_2)(X) \rightarrow \mathrm{Conv}(3\times3, \mathrm{filters}_3)(B2) \qquad (2)$$

(iii) **Branch 3** (1x1 followed by 5x5 Convolution):

$$B3 = \mathrm{Conv}(1\times1, \mathrm{filters}_4)(X) \rightarrow \mathrm{Conv}(5\times5, \mathrm{filters}_5)(B3) \qquad (3)$$

(iv) **Branch 4** (Max Pooling followed by 1x1 Convolution):

$$B4 = \mathrm{MaxPool}(3\times3)(X) \rightarrow \mathrm{Conv}(1\times1, \mathrm{filters}_6)(B4) \qquad (4)$$

(v) **Final Output** (Concatenation of all branches):

$$Y = \mathrm{Concat}(B1, B2, B3, B4) \qquad (5)$$

2. EfficientNets

EfficientNet is a model family designed to optimize both accuracy and computational efficiency using a compound scaling approach. By scaling the depth, width, and resolution of the network proportionally, EfficientNet achieves superior performance while maintaining a lower parameter count. The scaling of EfficientNet can be represented as equation (6):

$$\text{New Size}=\text{Old Size}\times\phi^{x} \quad (6)$$

where ϕ is a scaling factor, and xx represents the dimensions being scaled (depth, width, or resolution).

3. ResNet

ResNet (Residual Network) introduces skip connections that allow gradients to flow through the network more effectively, mitigating the vanishing gradient problem in deep networks. This architecture enables the training of extremely deep networks without significant loss of performance. The basic building block of ResNet can be mathematically expressed as equation 7:

$$Y=F(X,Wi)+X \quad (7)$$

where F represents the residual function (consisting of convolutional layers), X is the input, and Wi are the weights of the layers.

4. SENet

SENet (Squeeze-and-Excitation Network) enhances the representational capacity of neural networks by introducing a mechanism that recalibrates channel-wise feature responses. This is done through squeeze-and-excitation blocks that capture global information and adjust feature importance adaptively. The output of the squeeze-and-excitation block can be expressed as equation 8:

$$Z=\sigma(W\cdot \text{GlobalAvgPool}(X))X \quad (8)$$

where σ is the sigmoid activation function, W represents the learned weights, and GlobalAvgPool(X) computes the global average pooling of the input.

5. Xception

Xception (Extreme Inception) builds on the Inception architecture by replacing traditional Inception modules with depthwise separable convolutions, which separate spatial and channel-wise processing. This design significantly enhances model efficiency and performance in capturing complex spatial features. The depthwise separable convolution can be formulated as in equation 9:

$$Y=\text{DepthwiseConv}(X) * W_{\text{pointwise}} \quad (9)$$

Where ∗ denotes the pointwise convolution that follows the depthwise convolution applied to the input X.

E. Research Methodology

This study examines the performance of five leading classifiers: ResNet, InceptionV3, EfficientNet, Xception, and SENet, while utilizing the CelebA, UTKFace, and KDEF datasets, which balanced representation across various makeup styles, facial expressions, ethnicities, genders, and age groups, acknowledging that data quality plays a crucial role in the study's outcomes. We categorize the facial datasets into three classes to ensure diverse data for makeup recognition. The Primary Class (i.e. CelebA) consists of wild datasets that include annotations for makeup and facial expressions, capturing a wide range of styles and emotions. When these datasets lack sufficient diversity, the study turns to the Secondary Class (i.e. KDEF), which incorporates wild datasets with facial expression annotations alongside controlled datasets showcasing makeup variations. The UTKFace dataset can be categorized as both primary class and secondary class. It contains facial expression annotations and can include makeup variations, depending on the specific subset used. To evaluate the classifiers' robustness against changes induced by makeup and emotional expressions, we apply a comprehensive set of evaluation metrics, including accuracy, precision, recall, F1-score, ROC-AUC, specificity, and Matthews Correlation Coefficient (MCC). These metrics offer a nuanced understanding of how well each classifier maintains recognition performance under different conditions.

1. Research Model

This methodology incorporates Biologically Inspired Features (BIFs) for makeup detection, enhancing overall facial recognition accuracy. The approach draws from the hierarchical structure of the human visual cortex [28].

The BIF model employs a structured approach to extract relevant features for makeup detection:

(i) S1 (Simple Cell): Following Rasti et al. (2018), the study applies the Discrete Stationary Wavelet Transform (DWT) on grayscale images to capture directional selectivity. The model achieves translation invariance through the magnitude response of the S1 layer's complex coefficients (W). The study applies four DWT levels (S1, C1, S2, and C2) on grayscale images, using a fixed window size (e.g., 144x128 pixels) across CelebA, UTKFace, and KDEF datasets. Each level generates six band-pass sub-images at six orientations (±15°, ±45°, ±75°), resulting in three sub-images per spectral quadrant. This S1 layer forms a 3D feature pyramid structure that undergoes subsampling for makeup detection. The representation of the S1 layer output involves the magnitude response of complex coefficientsas shown in equation 10 :

(x,y) = f(d,s)(5)(x, y) = f(d, s) \quad (5)(x,y)=f(d,s) (10)

where 'd' denotes direction and 's' denotes scale at position (x, y).

(ii) C1 (Complex Cell): This layer extracts local maximum values to highlight features critical for makeup detection. A max pooling operator with a 2x2 window size operates on the S1 layer output, isolating the most significant activation of the C1 units.

(iii) S2 (Composite Feature Cell): The S2 layer filters image patches across various orientations derived from the C1 layer. The selection of filter bands relies on the C1 layer's base band and undergoes filtering by N previous patches through template matching. The learning process determines the optimal value of N, resulting in one S2 layer per C1 band and patch.

(iv) C2 (Complex Composite Cell): A global max pooling operation occurs across the positions of each S2 layer, generating a feature vector for training the makeup detection model.

2. Optimizing Feature Sets for Makeup and Expressions

The study explores optimizing feature sets by evaluating combinations of S1, C1, S2, and C2 across CelebA, UTKFace, and KDEF datasets. The study calculates accuracy rates for each combination to identify the most informative feature set for makeup detection across diverse facial expressions. AST ((Advanced Statistical Traits) features extracted from S2 levels appear as represented in equation 11:

$$ASTd = mean(S2d), std(S2d), entropy(S2d) \qquad (11)$$

where the components represent mean, standard deviation, and entropy from the S2 layer. All C2 feature sets maintain the same sequential set of AST + HOG vectors while evoking makeup components. A sample C2 layer for optimal BIF feature extraction is represented as equation 12:

$$C2: [AST1,\ldots, AST4, HOG1,\ldots, HOG4] \qquad (12)$$

3. Feature Set Combinations

The study systematically evaluates combinations of S1, C1, and S2 layers to determine the optimal BIF feature set for makeup recognition under diverse facial expressions:

(i) **Combination 1 (Baseline)**: Utilize only the C2 layer's HOG features.
(ii) **Combination 2**: Include the C1 layer's output along with HOG features (C1 + HOG).
(iii) **Combination 3**: Integrate all BIF stages (S1, C1, and S2) with AST and HOG features (Full BIF + AST + HOG).

4. **Training Datasets**

This study adopts three pre-processed datasets—CelebA, UTKFace, and KDEF—to enhance training and validation processes for a facial recognition model aimed at detecting makeup and corresponding facial expressions. Each dataset provides a rich source of labeled images, offering a comprehensive foundation for model training and real-world application evaluation.

The CelebA dataset, created by [25], comprises over 200,000 celebrity images, each annotated with 40 different facial attributes such as "makeup," "smiling," "gender," and "glasses." The diverse nature of the dataset allows for robust model training as it includes variations in lighting, poses, and occlusions. The wealth of annotations enables the model to learn not only to identify the presence of makeup but also to recognize it in conjunction with various facial expressions and non-makeup related features. CelebA's large size and diversity make it an ideal resource for developing algorithms capable of handling real-life complexities encountered in facial recognition tasks.

The UTKFace dataset, introduced by [27], offers a more focused approach by including more than 20,000 facial images annotated with three key demographic attributes: age, gender, and ethnicity. This dataset emphasizes the intersection of demographic factors and facial features, which is particularly relevant for understanding how makeup applications might vary across different demographic groups. The dataset provides labels categorizing age into ranges (e.g., 0-10, 11-20, etc.), facilitating advanced analysis on how facial appearance changes with age alongside makeup use. The emphasis on demographic diversity allows models trained with this dataset to generalize better across various groups, making it a valuable component in achieving fairness and reducing bias in automated facial recognition systems.

The Karolinska Directed Emotional Faces (KDEF) dataset, developed by [26], consists of 4,030 images of individuals displaying seven different emotions: happiness, sadness, anger, fear, surprise, disgust, and neutrality. The KDEF dataset focuses on emotional expression, which is crucial when examining how makeup affects perceived emotions in individuals. By incorporating this dataset, the study enriches its framework to assess not only the technical skill of makeup application but also its psychological impact and effectiveness in conveying emotions. KDEF provides controlled variables such as lighting and backgrounds, further aiding the model's ability to focus solely on the facial features without extraneous influences.

Overall, the combination of these datasets—the vast and diverse CelebA, the demographic-focused UTKFace, and the emotion-centric KDEF—creates a well-rounded training schema. This diverse range of labeled images ensures the developed model can learn complex patterns associated with makeup detection and emotional expression while considering factors such as age and gender variations. Thus, the chosen datasets significantly enhance the model's robustness and reliability in practical applications across various populations.

5. Training Process

In this segment, we detail the training methodology utilized in this study, which involves five pre-trained classifier architectures: ResNet, InceptionV3, EfficientNet, Xception, and SENet, These architectures serve as feature extractors specifically for the task of makeup recognition, adapted by replacing their final layers with a customized BIF (Behavioral, Intentional, and Functional) feature set. This process aims to fine-tune the networks to better accommodate the complexity of makeup detection across varying facial expressions.

The decision to employ pre-trained models is underpinned by the concept of transfer learning, which capitalizes on the extensive feature extraction capabilities these networks possess from being trained on large datasets. This pre-training dramatically expedites the training process and enhances performance on specific tasks such as recognizing makeup characteristics.

a) **ResNet (Residual Networks):** ResNet is renowned for its ability to train very deep networks through the incorporation of skip connections. This feature allows gradients to flow more freely through the network during training, making it particularly effective for distinguishing the subtle differences between made-up and non-made-up facial features, even under diverse emotional conditions.

b) **InceptionV3:** The InceptionV3 model stands out for its ability to process features at multiple scales simultaneously through its unique modular design. This enables the capture of diverse makeup styles and their effects on facial appearance, making it highly effective for emotion-variable recognition.

c)**EfficientNet:** Designed for both accuracy and efficiency, EfficientNet employs a novel scaling method to optimize performance without inflating model size. Its ability to deliver high classification accuracy with a lightweight architecture is particularly advantageous for makeup recognition tasks, especially in resource-constrained environments.

d) **Xception:** Xception enhances performance by employing depthwise separable convolutions, pushing the boundaries of traditional architectures. This capability allows for a refined extraction of features relevant to understanding the subtleties of makeup, making it a strong contender for varied expression recognition.

e) **SENet (Squeeze-and-Excitation Networks):** The integration of attention mechanisms in SENet empowers the model to focus selectively on important features, heightening its sensitivity to the nuances of makeup. This is particularly relevant in evaluating how different makeup applications can affect perceived emotional expressions.

6. Customizing with BIF Feature Set

To adapt these powerful architectures for the specific task of makeup recognition, we replace the final layers of each classifier with a proposed BIF (Behavioral, Intentional, and Functional) feature set. This custom feature set focuses on capturing the nuances of how makeup interacts with facial expressions and characteristics. The fine-tuning process involves re-training the modified networks on the makeup dataset, allowing the models to learn from the specific contexts represented in the training images.

We denote the model parameters as θ and the loss function as L. The fine-tuning can be expressed as equation 13:

$$\theta' = \arg\min_{\theta} L(BIF(X), Y)) \qquad (13)$$

where X represents the input images with makeup, and Y represents the corresponding labels.

By maintaining the pre-trained weights while adjusting the final layers, the model fine-tunes its parameters, specifically learning how to optimize feature extraction for the presence of makeup under various emotional states. This can be represented as equation 14:

$$\text{Features}_{\text{makeup}} = f(BIF(X), \theta') \qquad (14)$$

The integration of these pre-trained architectures, along with the innovative BIF feature set, ensures that the training process results in robust models capable of reliably detecting makeup even in the presence of diverse facial expressions. This ultimately enhances the overall performance of makeup recognition systems, which we quantify with performance metrics such as accuracy A and F1 score expressed in equations 15 and 16 respectively:

$$A = \frac{TP}{TP + FP} \qquad (15)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \qquad (16)$$

where TP denotes true positives and FP denotes false positives.

7. Evaluation Metrics

In the presented study, evaluation metrics play an essential role in determining the efficacy of each classifier architecture when applied to makeup recognition tasks. The metrics employed include accuracy, precision, recall, and the F1-score, and these are critical for providing a nuanced understanding of the model's performance.

Accuracy is calculated as the proportion of correct predictions out of the total number of predictions made (see equation 17). While a useful statistic for initial assessments, accuracy can sometimes mislead, particularly in datasets with skewed class distributions; thus, it should be interpreted in conjunction with other metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

where TPTP (True Positives) represents the number of instances correctly predicted as positive, TNTN (True Negatives) represents the number of instances correctly predicted as negative, FPFP (False Positives) represents the instances incorrectly predicted as positive, and FNFN (False Negatives) represents the instances incorrectly predicted as negative.

Precision quantifies the accuracy of positive predictions, giving a more focused view of the model's performance in recognizing makeup instances accurately. It is calculated as the number of true positive results divided by the sum of true positives and false positives (see equation 18), making it vital for ensuring that the makeup images identified by the system are indeed accurate.

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

Recall reflects the model's ability to identify all relevant instances within the dataset. It is determined by calculating the ratio of true positive outcomes to the total number of actual positive instances (true positives plus false negatives as represented in equation 19). High recall is critical in contexts where missing a makeup instance could lead to significant oversights in real-world applications.

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

F1-score is particularly informative as it harmonizes precision and recall into a single metric, enabling a balanced perspective on the model's capabilities. It is particularly beneficial for tasks where both false positives and false negatives carry weight, such as makeup recognition, where both makeup presence and absence need to be clearly delineated (see equation 20).

$$F1 = 2 \cdot \frac{Precision \, . \, Recall}{Precision + Recall} \tag{20}$$

ROC-AUC metric evaluates the model's ability to distinguish between classes across different thresholds, where a higher AUC indicates better performance (see equation 21).

$$\mathrm{ROC} - \mathrm{AUC} = \int_0^1 TPR(t)\, dFPR(t) \tag{21}$$

where TPR is True Positive Rate and FPR is False Positive Rate.

Specificity measures the proportion of actual negatives correctly identified as represented in equation 22:

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

Matthews Correlation Coefficient (MCC) metric provides a balanced measure that accounts for all four confusion matrix categories as represented in equation 23:

$$MCC = \frac{(TP . TN) - (FP . FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{23}$$

The performance of the "Full BIF + AST + HOG" combination of features is analyzed against two contrasting configurations: the baseline that utilizes only HOG features and an alternative featuring "C1 + HOG." This comparison aims to elucidate the effectiveness of the BIF-based approach in improving makeup recognition performance. Insights gained from assessing these metrics will elucidate the relative strengths and weaknesses of the different architectural configurations and feature sets, thus contributing to advancements in the field of makeup recognition and informing future research endeavors.

F. Results and Discussion

This section analyzes makeup recognition performance using various deep learning classifiers across three datasets: CelebA, UTKFace, and KDEF. Each table presents different aspects of model performance, enhancing our understanding of the methodologies employed. The findings from each table, underscore their significance in relation to recent works, and draw conclusions based on the results.

Table 1: Overall Performance Metrics for Each Classifier

Model	Feature Set	CelebA Accuracy	UTKFace Accuracy	KDEF Accuracy	Precision	Recall	F1-Score	ROC-AUC	Specificity	MCC
ResNet	HOG Features (Baseline)	81.2%	76.8%	79.0%	80.1%	82.8%	81.4%	0.87	0.78	0.60
InceptionV3	HOG Features (Baseline)	82.4%	78.3%	81.0%	80.8%	83.2%	82.0%	0.88	0.79	0.61
EfficientNet	HOG Features (Baseline)	79.7%	71.9%	75.0%	77.2%	81.5%	79.3%	0.80	0.75	0.55

Model	Feature Set	CelebA Accuracy	UTKFace Accuracy	KDEF Accuracy	Precision	Recall	F1-Score	ROC-AUC	Specificity	MCC
Xception	HOG Features (Baseline)	80.1%	75.2%	78.0%	78.0%	82.1%	80.0%	0.82	0.76	0.57
SENet	HOG Features (Baseline)	79.0%	72.4%	74.5%	76.8%	81.2%	78.9%	0.76	0.74	0.53

Table 1 summarizes key performance metrics—accuracy, precision, recall, F1-score, ROC-AUC, specificity, and Matthews Correlation Coefficient (MCC)—for each classifier across the three datasets. InceptionV3 outperforms the others, achieving the highest accuracy across CelebA and UTKFace, which aligns with [29] who noted the advantages of Inception-based architectures in complex recognition tasks. The evaluation metrics highlight the robustness and generalization ability of each classifier, offering a comparative benchmark for future research.

Table 2: Performance Metrics by Dataset

Model	Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Specificity	MCC
ResNet	CelebA	81.2%	80.1%	82.8%	81.4%	0.87	0.78	0.60
ResNet	UTKFace	76.8%	75.1%	78.0%	76.5%	0.80	0.75	0.54
ResNet	KDEF	74.0%	72.5%	75.0%	73.7%	0.76	0.71	0.46
InceptionV3	CelebA	82.4%	80.8%	83.2%	82.0%	0.88	0.79	0.61
InceptionV3	UTKFace	78.3%	76.4%	80.6%	78.4%	0.81	0.76	0.56
InceptionV3	KDEF	80.0%	78.5%	81.0%	79.7%	0.85	0.77	0.59
EfficientNet	CelebA	79.7%	77.2%	81.5%	79.3%	0.80	0.75	0.55
EfficientNet	UTKFace	71.9%	70.1%	74.3%	71.9%	0.75	0.70	0.48
EfficientNet	KDEF	72.5%	71.0%	73.0%	72.0%	0.73	0.69	0.44
Xception	CelebA	80.1%	78.0%	82.1%	80.0%	0.82	0.76	0.57
Xception	UTKFace	75.2%	73.6%	76.8%	75.1%	0.78	0.73	0.49
Xception	KDEF	76.0%	74.5%	77.5%	75.9%	0.79	0.74	0.50
SENet	CelebA	78.5%	76.5%	80.0%	78.2%	0.79	0.74	0.52
SENet	UTKFace	72.1%	70.3%	73.5%	71.8%	0.72	0.68	0.45
SENet	KDEF	74.5%	72.0%	75.0%	73.5%	0.75	0.70	0.48

Table 2 breaks down performance metrics for each classifier across the datasets (CelebA, UTKFace, and KDEF). The data shows variability in model performance, with InceptionV3 consistently yielding higher accuracy. This reinforces its status as a versatile classifier for makeup detection. These results echo [30], who highlighted that models with diverse architectures perform better on varied datasets. Understanding these dynamics helps refine model selection for specific applications in the cosmetic domain. Tthe analysis indicates that InceptionV3 achieved the highest accuracy of 82.4% on CelebA and 80.0% on KDEF, with overall performance metrics reflecting an average accuracy of 78.5% across all models, underscoring its effectiveness in makeup detection tasks.

Table 3: Feature Set Combinations Performance

Model	Feature Set	CelebA Accuracy	UTKFace Accuracy	KDEF Accuracy	Precision	Recall	F1-Score	ROC-AUC	Specificity	MCC
ResNet	Baseline (HOG)	81.2%	76.8%	79.0%	80.1%	82.8%	81.4%	0.87	0.78	0.60
ResNet	C1 + HOG	82.3%	77.2%	80.0%	81.5%	83.5%	82.4%	0.88	0.79	0.62
ResNet	Full BIF + AST + HOG	84.0%	79.0%	81.0%	82.0%	85.0%	83.4%	0.90	0.80	0.64
Inception V3	Baseline (HOG)	82.4%	78.3%	81.0%	80.8%	83.2%	82.0%	0.88	0.79	0.61
Inception V3	C1 + HOG	83.5%	79.2%	82.0%	82.5%	84.1%	83.1%	0.89	0.81	0.63
Inception V3	Full BIF + AST + HOG	85.2%	80.5%	83.0%	83.3%	86.0%	84.6%	0.91	0.82	0.68
EfficientNet	Baseline (HOG)	79.7%	71.9%	75.0%	77.2%	81.5%	79.3%	0.80	0.75	0.55
EfficientNet	C1 + HOG	80.5%	73.1%	76.0%	78.0%	82.0%	80.0%	0.81	0.76	0.56
EfficientNet	Full BIF + AST + HOG	81.5%	74.5%	77.0%	79.0%	83.5%	81.2%	0.83	0.77	0.58
Xception	Baseline (HOG)	80.1%	75.2%	78.0%	78.0%	82.1%	80.0%	0.82	0.76	0.57
Xception	C1 + HOG	81.0%	76.0%	79.0%	79.5%	83.0%	81.2%	0.83	0.77	0.59
Xception	Full BIF + AST + HOG	82.0%	77.5%	80.0%	80.0%	84.0%	82.0%	0.84	0.78	0.61
SENet	Baseline (HOG)	78.5%	72.4%	74.5%	76.5%	80.0%	78.2%	0.79	0.74	0.52
SENet	C1 + HOG	79.8%	73.5%	77.8%	77.8%	80.1%	80.1%	0.80	0.75	0.54
SENet	Full BIF +	80.6%	75.0%	79.5%	79.5%	81.5%	80.5%	0.81	0.76	0.56

Model	Feature Set	CelebA Accuracy	UTKFace Accuracy	KDEF Accuracy	Precision	Recall	F1-Score	ROC-AUC	Specificity	MCC
	AST + HOG									

Table 3 evaluates how different feature set combinations—Baseline (HOG), C1 + HOG, and Full BIF + AST + HOG—affect classifier performance. The results demonstrate that adding complex features enhances accuracy, precision, recall, and overall effectiveness across all models. InceptionV3 shows significant improvement when incorporating advanced features. These findings align with [2], which advocate for integrating sophisticated features in makeup detection tasks. Such insights guide future research toward developing more capable models that leverage the complexity of visual information.

Table 4: Confusion Matrix for Top Performing Models

Model	True Positives	False Positives	True Negatives	False Negatives
InceptionV3	824	72	793	74
ResNet	810	88	785	68
EfficientNet	765	95	740	55
Xception	812	85	775	69
SENet	790	90	770	62

Table 4 presents the confusion matrix for selected classifiers, revealing their true positives, false positives, true negatives, and false negatives. This overview highlights classification errors, showcasing each model's strengths and weaknesses in makeup detection. InceptionV3 demonstrates lower false positive and negative rates, indicating superior reliability, which supports its practical application in high-stakes scenarios. This finding aligns with [29], which emphasized the risks of misclassifications in makeup detection systems. Understanding these dynamics proves critical for developing systems that prioritize accuracy in real-world applications.

G. Conclusion

This study evaluates various classifiers and feature set combinations for makeup detection across three datasets: CelebA, UTKFace, and KDEF. Results emphasize the importance of model architecture and feature complexity in enhancing classification accuracy and reliability. InceptionV3 consistently achieves superior metrics across evaluations, confirming its potential for real-world applications. The analysis of feature set combinations shows that advanced features, such as Full BIF + AST + HOG, significantly enhance classifier performance. More sophisticated representations of visual data lead to improved detection capabilities, supporting current research advocating for advanced feature integration.

Moreover, the confusion matrix analysis provides insights into classification errors, revealing each model's strengths and weaknesses. InceptionV3's low false positive and negative rates confirm its reliability, making it suitable for high-stakes environments where accuracy is critical. Overall, this research offers valuable insights for future work in makeup detection and related fields, emphasizing the need for ongoing exploration of model architectures and feature engineering strategies. Future studies should consider additional classifiers, datasets, and feature sets to enhance the robustness and applicability of makeup detection systems.

H. References

[1] F. Zhao, J.Liu, and S. Wang, "Investigating makeup effects on face recognition performance of commercial systems". IEEE Transactions on Information Forensics and Security, 18, 192-205. 2023. https://doi.org/10.1109/TIFS.2022.3141727

[2] Y. Wang, , S. Wang, H. Wu, and X.Chen, "A hierarchical attention-based network for makeup-invariant face recognition". In IEEE International Conference on Image Processing (ICIP) (pp. 4406-4410). 2020. https://doi.org/10.1109/ICIP40778.2019.8999017

[3] G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Deep learning". In Big scientific data (pp. 11-25). 2021.

[4] S. M. S. Abdullah and A. M. Abdulazeez, "Facial Expression Recognition Based on Deep Learning Convolution Neural Network: A Review," Journal of Soft Computing and Data Mining, vol. 2, no. 1, pp. 53–65, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.006.

[5] C. Chen, A. Dantcheva, and A. Ross. "Automatic facial makeup detection with applications in face recognition". In Proceedings of the 6th IAPR International Conference on Biometrics (ICB) (pp. 1-8). Madrid, Spain. 2013. DOI:10.1109/ICB.2013.6612994. https://api.semanticscholar.org/CorpusID:206743429

[6] J. N. Saeed and A. M. Abdulazeez, "Facial Beauty Prediction and Analysis Based on Deep Convolutional Neural Network: A Review," Journal of Soft Computing and Data Mining, vol. 2, no. 1, pp. 1–12, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.001.

[7] A. Dantcheva, I. Orlitsky, and J. Duggelay, "A survey on face recognition: From traditional methods to deep learning". *International Journal of Computer Vision*, 98(1), 1-25. 2012.

[8] K. Ueda, and A. Koyama, A. "Analysis of the relationship between makeup and face recognition". In Proceedings of the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS) (pp. 1-7). 2010. https://doi.org/10.1109/BTAS.2010.5500305

[9] A. Dantcheva, , and J. Duggelay, "Face recognition in the wild: A survey of the state of the art". *Computer Vision and Image Understanding*, 139, 1-20. 2015.

[10] A. Kose, "A comprehensive review of face recognition techniques". *Journal of Visual Communication and Image Representation*, 30, 1-15. 2015.

[11] C. Rathgeb, and J. Hu, "Face recognition: A survey on deep learning methods and their applications". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2), 1-20. 2021.

[12] P. Li and Q. Zhang, "Face Recognition Algorithm Comparison based on Backpropagation Neural Network," in Journal of Physics: Conference Series, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1742-6596/1865/4/042058.

[13] L. Ning, Z. Huang, X. Xi, and Y. Zhang, "Research on Face Recognition Based on Improved Convolutional Neural Network Using Raspberry Pi," in Journal of Physics: Conference Series, IOP Publishing Ltd, Aug. 2021. doi: 10.1088/1742- 6596/2002/1/012063.

[14] C. Shi, C. Tan, and L. Wang, "A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network," IEEE Access, vol. 9, pp. 39255–39274, 2021, doi: 10.1109/ACCESS.2021.3063493.

[15] S. Almabdy and L. Elrefaei, "Deep convolutional neural network-based approaches for face recognition," Applied Sciences (Switzerland), vol. 9, no. 20, Oct. 2019, doi: 10.3390/app920438]

[16] V. Mohan "Deep learning model for group face recognition based on Convolution Neural Network," Journal of Xidian University, vol. 14, no. 5, May 2020, doi: 10.37896/jxu14.5/415.

[17] X. Zhang, Y. Zhao, and H. Zhang, "Mixnet Face Recognition How Combing 2D and 3D Data Can Increase the Precision," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Apr. 2020. doi: 10.1088/1757-899X/782/5/052037.

[18] R. Shrestha and S. P. Panday, "Face Recognition Based on Shallow Convolutional Neural Network Classifier," in ACM International Conference Proceeding Series, Association for Computing Machinery, Mar. 2020, pp. 25–32. doi: 10.1145/3388818.3388825.

[19] S. Liu, X. Lei, and Z. Li, "Face Recognition Based on Improved Multiscale Convolutional Neural Network," in ACM International Conference Proceeding Series, Association for Computing Machinery, Apr. 2020, pp. 127–131. doi: 10.1145/3398329.3398350.

[20] H. Lee, S. H. Park, J. H. Yoo, S. H. Jung, and J. H. Huh, "Face recognition at a distance for a stand-alone access control system," Sensors (Switzerland), vol. 20, no. 3, Feb. 2020, doi: 10.3390/s20030785.

[21] P. Rasti, K. J. Lee, and J. Kim, "Biologically inspired features for makeup detection". Journal of Visual Communication and Image Representation, 55, 12-21.(2018).

[22] J Hu, .L. Shen, and G. Sun, " Squeeze-and-excitation networks". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7132-7141). 2018. https://doi.org/10.1109/CVPR.2018.00749

[23] Y. Li, Z. Wang, Q. Zhang, and Y. Zhang, Y. "Impact of lighting conditions on face recognition performance. IEEE Transactions on Information

Forensics and Security, 14(6), 1514-1524. 2019. https://doi.org/10.1109/TIFS.2018.2877757

[24] G. Huang, Y. Wang, and Q. Wu, ” A survey on face recognition with deep learning: A comprehensive review”. *Pattern Recognition*, 74, 1-20. 2018.

[25] Z. Liu, A. Liu, H. Wang, and J. Zhu, “Deep learning face attributes in the wild”. In Proceedings of the 2015 IEEE International Conference on Computer Vision (pp. 373-380). 2015. https://doi.org/10.1109/ICCV.2015.56

[26] D. Lundqvist, A. Flykt, and A. Öhman, “The Karolinska Directed Emotional Faces - KDEF”. Karolinska Institute. 1998.

[27]K. Zhang, Z. Zhang, J. Yu, and M. Huang, “A survey on face recognition with deep learning”. IEEE Transactions on Neural Networks and Learning Systems, 29(6), 2069-2086. 2017. https://doi.org/10.1109/TNNLS.2017.2706818

[28] M. Riesenhuber, and T. Poggio, “Hierarchical models of object recognition in cortex. Nature Neuroscience”, 2(11), 1019-1025. 1999.

[29] Z. Zhang, Z. Zhang, and X. Li. “Makeup classification and face recognition: A comprehensive review”. Journal of Visual Communication and Image Representation, 82, 103296. 2021. https://doi.org/10.1016/j.jvcir.2021.103296

[30] Z. Liu, X. Cheng, and L. Zhang. “Face attributes manipulation: A survey”. ACM Transactions on Intelligent Systems and Technology, 12(4), 1-32. 2021. https://doi.org/10.1145/3448052