A Metric for Evaluating Face Image Deep Learning Classification Models

Fernando Quiroz Jr.
School of Technology and Computer Studies
Biliran Province State University
Naval, Biliran Philippines
evan.quiroz@bipsu.edu.ph

Vladimir Y. Mariano
College of Computing and Information Technologies
National University Manila
Sampaloc, Manila Philippines
vymariano@national-u.edu.

***Abstract*— This study proposes a multidimensional metric for evaluating deep learning models in face image classification, addressing long-standing limitations of accuracy-centered evaluation. While modern architectures such as convolutional neural networks and Vision Transformers demonstrate strong predictive performance, conventional metrics fail to capture deeper structural behaviors, including fairness, representational quality, interpretability, and computational viability. To address this gap, the proposed framework integrates six components: validation F1 score, linear-probe embedding separability, fairness via skin-tone gap reduction, compute efficiency, anatomical interpretability, and embedding stability, yielding a holistic assessment of model performance. Applied across six benchmark architectures and four Baumann Skin Type tasks, the metric reveals consistent superiority of transformer-based models, which exhibit stronger fairness, stability, and interpretability compared to CNNs. Findings demonstrate that responsible facial model evaluation requires multidimensional criteria that move beyond accuracy to ensure equitable, transparent, and deployment-ready systems. This work contributes toward trustworthy and fairness-aware facial AI.**

Keywords—*AI fairness, CNN, facial image classification, evaluation metrics, vision transformers*

I. Introduction

The rapid expansion of facial image classification has been driven by breakthroughs in deep learning, particularly convolutional neural networks (CNNs) and transformer-based architectures. Landmark models such as VGGFace, FaceNet, ArcFace, and Vision Transformers have demonstrated unprecedented performance in face recognition, expression analysis, attribute prediction, and identity verification tasks. As datasets such as CelebA and FairFace expanded in size and diversity, research in face image classification continued to accelerate, which deepens model sophistication and societal reach.

Despite these advancements, model evaluation in facial image classification remains overwhelmingly centered on accuracy-based metrics, particularly accuracy [1] and the F1 score [2]. While these metrics quantify predictive correctness, they fail to capture deeper, structural properties of model behavior. Studies revealed that commercial facial analysis systems, despite reporting high overall accuracy, exhibited error rates up to 34.7 times higher for darker-skinned women compared to lighter-skinned men, which highlights the inadequacy of global accuracy as a measure of equitable model performance [3]. Similarly, a study demonstrated that accuracy often masks disparate impacts, where models maintain high aggregate performance while performing disproportionately poorly on marginalized demographic groups [4]. Complementary studies reinforce these findings, which shows that facial recognition models remain sensitive to demographic variables such as age, race, and sex [5], [6], [7]; while other studies argue that fairness-aware evaluation requires metrics beyond conventional classification performance, including subgroup error analysis and representational quality measures [8], [9]. Collectively, these studies demonstrate that sole reliance on accuracy obscures essential concerns including dataset imbalance, representational bias, and disparate performance across demographic groups. Such blind spots pose risks when evaluating benchmark models, as they reward models optimized for majority groups and encourage architectures that reinforce bias rather than ensuring equitable representation. In fairness-critical domains like facial analysis, accuracy alone is an insufficient, and potentially misleading, basis for model selection.

In response to these limitations, several alternative evaluation strategies have emerged. Researchers have explored balanced accuracy, ROC-AUC, precision–recall trade-offs [10], embedding separability [11], and demographic performance disaggregation [12] . Other works evaluated computational metrics such as inference latency, model size, or energy efficiency, especially in mobile deployment contexts [13]. However, these approaches are typically assessed in isolation, or used only in pairwise comparative studies, which leaves a persistent gap: the absence of a unified, multi-dimensional evaluation framework that holistically captures predictive performance, feature representation quality, fairness, interpretability, and computational efficiency. Current literature lacks an integrated metric that systematically consolidates these factors into a single evaluative structure for comparing deep learning benchmark models in facial image classification.

To address this gap, the study introduces a comprehensive evaluation metric tailored for modern facial classification models. The framework integrates six dimensions to provide a richer, more actionable view of model behavior. This multi-metric approach promotes transparent, responsible model selection by evaluating not only accuracy but also representational fairness and deployment readiness, contributing to broader efforts in AI fairness and trustworthy facial analysis.

II. METHODOLOGY

This study adopts a quantitative, experimental research design to develop and validate a multidimensional evaluation metric for deep learning-based face image classification models. The methodology consists of three stages: model selection, dataset preparation and preprocessing, and evaluation metric development.

A. Model Selection

This study evaluates six deep learning architectures that represent two dominant paradigms in modern face image classification: (1) Convolutional Neural Network (CNN) backbones, which encode strong spatial inductive biases [14], and (2) Vision Transformer (ViT) models, which rely on patch tokenization and global self-attention [15].

CNNs are widely used in facial image classification because their architecture leverages spatial structure through localized receptive fields, shared filters, and hierarchical feature extraction. Early layers learn edges and textures, while deeper layers capture higher-level facial geometry and semantic patterns. With non-linear activations, pooling, and expanding receptive fields, CNNs build strong multi-scale facial representations. Weight sharing further provides translation invariance, helping the network recognize features despite small shifts or distortions.

Three architectures, ResNet, EfficientNet, and DenseNet were selected because they represent complementary strengths in depth, efficiency, and feature reuse. ResNet introduced residual learning, which enables the training of very deep neural networks by addressing vanishing gradients [16], [17]. A residual block is expressed as:

$$y = F(x, W) + x$$

where $F(x, W)$ is the residual mapping of the identity shortcut connection *(x)* and weights *(W)*. These residual connections allow gradients to flow directly through the identity path, which stabilizes optimization in deep networks.

EfficientNet introduces compound scaling, where depth d, width and resolution are scaled uniformly to allow the model to maintain balanced capacity across layers [18], [19]. EfficientNet is built around the Mobile Inverted Bottleneck Convolution (MBConv) with Squeeze-and-Excitation (SE) attention:

$$y = W_{proj} \bullet \sigma(W_{exp} \bullet x \odot SE(x))$$

where the SE block computes:

$$s = \sigma\big(W_2 \delta(W_1 z)\big), z = GAP(x)$$

DenseNet introduces dense connectivity, where each layer receives inputs from all previous layers [20]. This connectivity pattern encourages feature reuse and strengthens gradient flow. The growth rate k controls the number of new feature maps added per layer:

$$C_l = C_0 + k \bullet l$$

A dense block applies:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}])$$

where $H_l$is BN → ReLU → Conv. DenseNet's concatenation property ensures earlier textural cues (such as freckles, eye corner details) remain accessible in deeper layers.

Meanwhile, Vision Transformers (ViTs) excel in facial image classification by modeling long-range dependencies and global context through self-attention. Instead of relying on local convolutions, ViTs divide an image into fixed-size patches and embed each as a token, treating the image as a sequence. This approach captures fine-grained patch-level detail while enabling global interactions across all facial regions via multi-head self-attention.

In this study, three Vision Transformer patch embedding strategies were implemented to evaluate how different positional encoding mechanisms influence facial representation learning. Standard patch embedding [21] divides into non-overlapping patches x_i, each flattened and projected using a learnable matrix W_E:

$$z_i = W_E \bullet Flatten(x_i) + P_i$$

with P_i denoting sinusoidal or learned positional encodings. This method is computationally efficient and allows global self-attention but lacks strong locality modeling, which may limit performance on fine-grained facial features. To address this limitation, sequential overlapping patch embedding [22] introduces overlapping windows $x_{i:i+k}$ with stride $s < k$, improving spatial continuity across patches:

$$\hat{x}_i = x_{i:i+k}, \qquad z_i = W_E \bullet Flatten(\hat{x}_i) + P_i$$

often combined with sequential positional biases $P_i = f(i) + g(i-1)$ to preserve ordering. This approach reduces patch-boundary artifacts and enhances the modeling of local facial structures such as eyes and mouth regions. Finally, convolutional patch embedding [23] incorporates a convolutional layer prior to projection to infuse CNN-like inductive biases:

$$\tilde{x}_i = Conv(x_i), \qquad z_i = E(\tilde{x}_i) + P_i$$

This method improves spatial coherence and is particularly advantageous for faces, where subtle geometric patterns and texture cues are critical.

B. Dataset Preparation and Preprocessing

The dataset was compiled from publicly available facial image repositories identified through targeted searches: *"facial skin images dermatology dataset," "skin texture face photos"*. Only datasets with explicit open-access licenses or verifiable consent documentation were included. The final collection integrates facial images from multiple reputable sources, which include CelebA, FairFace, Caltech Faces, Labeled Faces in the Wild and IMDB-WIKI were incorporated to capture variations in pose, illumination, age, and real-world conditions. To further expand diversity, images were compiled from open-source Roboflow Universe datasets (2021–2025), and images from various Kaggle repositories focusing on skin type, tone, and dermatological conditions. All datasets were screened to ensure compliance with licensing, consent statements, and ethical use guidelines.

Inclusion criteria required sufficient resolution, frontal orientation, and unobstructed cheek and nasal regions. Images were excluded for heavy makeup, major occlusions, severe blur, or non-human content. Automated filtering first checked image type, sharpness, and facial visibility, removing grayscale images without chromatic data and samples too blurred for texture analysis. Facial landmarks were detected using a 68-point model to verify boundaries, symmetry, and region visibility. Images with excessive head tilt, cut-off areas, or open-mouth expressions were discarded to ensure consistent facial geometry. Faces were then extracted, anonymized, cropped, and resized to 224×224 pixels on a uniform background. To further reduce identifiable features, eye and mouth regions were masked while preserving the cheek and nasal areas. Final labels for Baumann Skin Types (oiliness, sensitivity, pigmentation and aging) [24] were assigned independently by two aestheticians and one dermatologist.

To ensure diversity and fairness, skin-tone distribution was evaluated using the Monk Skin Tone (MST) Scale [25]. This can be done also with other representation techniques as age group, sex, and others depending on the task. The dataset contains 3,000 images, dominated by medium and darker tones across most BST labels, particularly oily, pigmented, non-pigmented, and wrinkled categories. Light tones are minimally represented, while medium tones appear at moderate levels. This imbalance, shown in Figure 1, may affect downstream analyses and model performance.

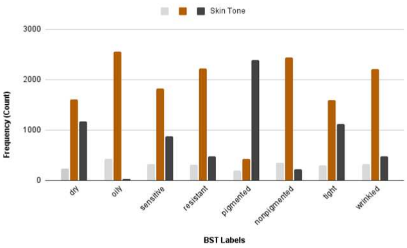

Fig. 1. Dataset distribution across BST labels and skin tone

C. *Evaluation Metric Development*

The metric is composed of six components that quantifies a distinct structural behavior of deep learning models, which allows holistic comparison beyond accuracy.

Component 1. The F1-score is a harmonic mean of precision and recall, making it more informative than accuracy when evaluating models on imbalanced facial datasets where minority classes, such as specific skin types or skin tones, are underrepresented [26]. Unlike accuracy, which can mask poor performance on small subgroups, the F1-score penalizes models that fail to detect minority instances (low recall) or frequently misclassify them (low precision). It is computed as:

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall}$$

Because the harmonic mean amplifies the effect of low values, the F1-score drops sharply when the model struggles with either minority-class sensitivity or specificity.

Component 2. Linear-probe evaluation is widely used to assess the intrinsic quality of learned representations independent of fine-tuning [27]. The method evaluates whether the backbone encodes linearly separable features. The linear-probe accuracy is computed as:

$$LP_{acc} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_i)$$

This evaluation directly measures the intrinsic quality of learned representations, free from the influence of fine-tuning or optimization tricks, and therefore offers a fair and interpretable basis for comparing backbone models.

Component 3. Bias across skin tones is a documented failure mode of facial classification systems. For this study, fairness is measured using the Monk Skin Tone (MST) scale [28]. The fairness gap is computed as:

$$Gap_{MST} = \max A_k - minA_k$$

Fairness score is computed as (higher = fairer):

$$Fairness = 1 - Gap_{MST}$$

This metric directly captures how consistently a model performs across skin tones; large gaps indicate representational or decision-boundary biases that can lead to unequal error rates and harmful downstream impacts [29].

Component 4. To support lightweight deployment, it is essential to evaluate the computational efficiency of a model, particularly its inference latency, since even highly accurate systems become impractical if they respond too slowly [30]. Inference latency per batch is measured, where t_i denotes the time required to process a batch. The mean latency is computed as:

$$\bar{t} = \frac{1}{B}\sum_{i=1}^{B} t_i$$

To make latency comparable across models and deployment settings, the score is normalized within a defined operational range $[t_{\min}, t_{\max}]$:

$$S_{lat} = 1 - \frac{\bar{t} - t_{min}}{t_{max} - t_{min}}$$

Values are then clipped to the interval [0,1], where higher scores indicate faster, more deployment-ready models. This normalization ensures that latency does not overwhelm other evaluation metrics and allows for fair comparison of systems with different computational profiles.

Component 5. Interpretability is essential for fairness-critical facial applications because it reveals whether a model relies on meaningful facial cues or on spurious correlations that may propagate bias [31]. In this study, the anatomical plausibility was evaluated of explanation maps generated using LIME, which produces local perturbation-based attributions highlighting image regions most influential to the model's prediction. Let $H(x, y)$ denote the normalized LIME saliency map, and let $R_j(x, y)$ represent binary masks corresponding to anatomically relevant regions of the face, specifically the forehead, nose, and cheeks. The proportion of attribution assigned to region j is computed as:

$$P_j = \frac{\Sigma_{x,y} H(x, y) R_j(x, y)}{\Sigma_{x,y} H(x, y)}$$

An overall anatomical interpretability score is then obtained by weighting these regional proportions according to their diagnostic relevance:

$$S_{XAI} = 0.3P_{forehead} + 0.4P_{nose} + 0.3P_{cheeks}$$

High-quality model explanations should concentrate attribution within core facial structures, those most relevant to human visual reasoning and clinical interpretation, rather than on hair, background areas, or occlusions. A higher S_{XAI} therefore reflects explanations that are more anatomically plausible and trustworthy for fairness-sensitive facial analysis tasks.

Component 6. Embedding stability assesses whether a model's facial embeddings maintain high intra-class similarity and low inter-class similarity, a core principle in metric learning and Siamese networks [32]. Using cosine similarity, the expected similarity for samples from the same class was computed:

$$\mu_{within} = \mathbb{E}[sim(i, j) | y_i = y_j],$$

and for samples from different classes,

$$\mu_{between} = \mathbb{E}[sim(i, j) | y_i \neq y_j],$$

A stability score is defined to quantify how well the embeddings separate identities or categories:

$$S_{stab} = \frac{1}{1 + (\mu_{between} - \mu_{within})}$$

Higher values indicate more stable and discriminative embeddings, where same-class samples cluster tightly while different-class samples remain well separated.

Each metric in the evaluation framework is first normalized to the interval [0,1]to ensure comparability across measures that naturally exist on different scales. The metrics are combined through a weighted summation to produce a single composite performance score:

$$S_{final} = 0.40S_{F1} + 0.20S_{LP} + 0.15S_{Fair} + 0.10S_{lat} \\ + 0.10S_{XAI+}0.05S_{stab}$$

The weighting scheme reflects the relative importance of each component in fairness-critical facial analysis. F1-score receives the largest weight (0.40) as it is the primary measure of predictive performance, especially under class imbalance. Linear-probe accuracy (0.20) follows, emphasizing the quality of learned representations that support reliability and fair downstream behavior. Fairness (0.15) is strongly weighted to penalize demographic disparities without overshadowing core accuracy metrics. Latency (0.10) and anatomical interpretability (0.10) are equally valued, highlighting the need for both efficient deployment and trustworthy, anatomically grounded explanations. Embedding stability (0.05) contributes modestly, capturing representational consistency while remaining secondary to accuracy and fairness.

III. Results and Discussion

The multidimensional evaluation framework was applied to six deep learning architectures: three CNN-based models and three Vision Transformer with patch embedding variants, across four Baumann Skin Type (BST) facial classification tasks: oiliness, sensitivity, pigmentation, and aging.

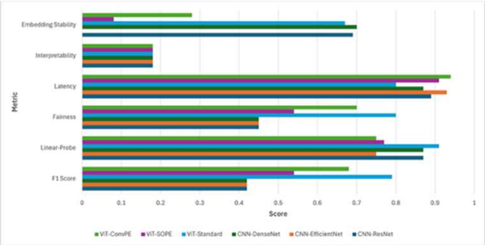

Fig. 2. Evaluation results for oiliness dimension

As shown in Figure 2, oiliness classification revealed stark contrasts between CNN and ViT architectures. All CNN models converged at an F1 score of 0.42, which mirrors traditional failures associated with class imbalance and insufficient representational generalization [33]. Although CNNs such as ResNet and DenseNet achieved relatively strong linear-probe scores (0.87), indicating potentially rich internal features, their inconsistent embedding stability and moderate fairness values (0.45) limited their overall performance [34]. The ViT family demonstrated stronger results, with ViT-Standard producing the highest composite score (0.7495). Its superior F1 score (0.79), high fairness value (0.80), and strong embedding stability (0.67) underscore the transformer's ability to integrate global facial cues, coherent, discriminative representations. These findings reinforce the notion that transformer architectures can better capture the subtle texture gradients associated with oiliness compared to CNNs' localized filters [35].

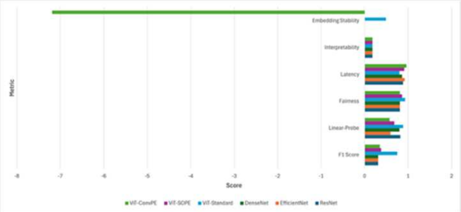

Fig. 3. Evaluation results for sensitivity dimension

As shown in Figure 3, sensitivity prediction is the most challenging BST tasks, as it depends on nuanced redness patterns, micro-irritation markers, and inflammationbased cues [35]. CNNs demonstrated performance limitations: all three models reached an F1 score of 0.31 with stability near zero, indicating weak clustering of sensitivity-related features in embedding space. Despite moderately strong linear-probe accuracies (0.80–0.82), these models did not translate representational quality into effective decision boundaries, which suggests a disconnect between learned features and the classifier's ability to separate sensitive versus resistant categories [36]. ViT-Standard achieved the best performance across all metrics (total score 0.738), driven by a strong F1 score (0.75), high fairness (0.93), and substantially improved stability (0.49). These gains likely arise from ViTs' global attention, which enables more coherent modeling of distributed facial irritation cues [37]. The poor performance of ViT-ConvPE, largely due to severe embedding instability (–7.19), which illustrates the importance of carefully selected positional encodings in transformer architectures [38].

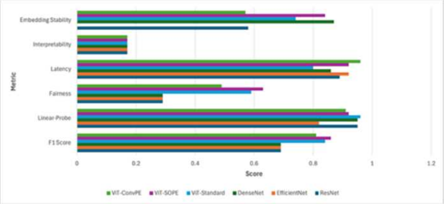

Fig. 4. Evaluation results for pigmentation dimension

As shown in Figure 6, pigmentation classification yielded the highest performance across models, likely due to distinct chromatic and melanin patterns [39]. CNNs showed strong linear-probe separability (up to 0.95) but consistently low fairness scores (0.29), which indicates uneven performance across skin tones [40]. Their embedding stability ranged from moderate (0.58) to high (0.87), suggesting good identity clustering but limited task-specific fairness. ViT-SOPE achieved the highest composite score (0.7735), with strong F1 (0.86), high linear separability (0.92), and the best embedding stability (0.84). Its overlapping patch mechanism appears well suited for modeling melanin gradients and facial discoloration [41].

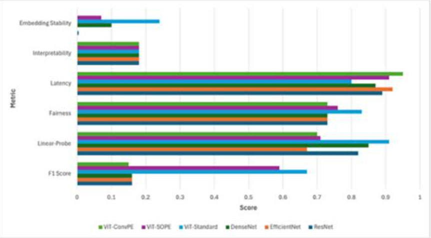

Fig. 5. Evaluation results for aging dimension

As shown in Figure 5, aging prediction demonstrated the widest disparity between CNN and ViT performance. CNNs produced uniformly low F1 scores (0.16) despite strong linear-probe accuracies (0.82–0.85), which suggests that they encoded meaningful features but failed to translate these into effective classification decisions. This aligns with literature noting that aging cues are spatially diffuse and require global contextual modeling [42]. ViT-Standard achieved the highest overall score (0.6845) with an F1 of 0.67. Interestingly, ViT-SOPE also performed well (0.6045), whereas ViT-ConvPE (0.4225) lagged, highlighting that positional encoding choices materially influence transformer performance [43].

Across all four tasks, Vision Transformers consistently outperformed CNN architectures in every major dimension of the evaluation metric: F1 score, fairness, interpretability, embedding stability, and total composite performance. CNNs occasionally achieved strong linear-probe scores, but their decision layers consistently underperformed, which indicates insufficient transfer of representational quality into classification accuracy [44]. This suggests that CNNs may encode meaningful features but struggle with tasks requiring spatially global or contextually diffuse information. Moreover, CNNs exhibited substantially lower fairness scores, demonstrating higher susceptibility to demographic performance gaps, consistent with literature documenting CNNs' bias amplification tendencies. ViTs consistent higher interpretability scores indicate that ViTs rely on meaningful facial regions, which align more closely with expert reasoning and making them well-suited for fairness-critical clinical and cosmetic applications.

IV. Conclusion and Recommendation

These findings affirm that evaluating facial classification systems must extend beyond accuracy to include structural, representational, and ethical dimensions. The proposed composite metric provides a more actionable way to identify balanced, fair, and reliable models aligned with broader AI fairness and responsible ML initiatives. Researchers and practitioners should therefore adopt metrics that does not relying on accuracy or F1 since it has risks of favoring models that perform well overall but fail on marginalized groups. Anatomical interpretability further revealed whether models used meaningful facial structures or spurious cues, which underscores the need for explanation-pattern analysis in institutional deployments. Persistent fairness disparities highlight the importance of balanced datasets across skin tones, age groups, and gender identities, motivating future work on broader demographic representation and explicit bias-mitigation techniques.

References

[1] G. Shao, L. Tang and H. Zhang, "Introducing image classification efficacies," *IEEE Access,* pp. 134809-134816, 9.

[2] K. Smelyakov, Y. Honchar, O. Bohomolov and A. Chupryna, "Machine Learning Models Efficiency Analysis for Image Classification Problem," in *In COLINS*, 2022.

[3] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *In Conference on fairness, accountability and transparency*, 2018.

[4] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[5] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on information forensics and security,* vol. 7, no. 6, pp. 1789-1801, 2012.

[6] P. Grother, "Face recognition vendor test (FRVT) part 8: Summarizing demographic differentials," *National Institute of Standards and Technology (NIST),* vol. 8, p. 8429, 2022.

[7] P. Terhörst, J. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. Moreno, J. Fierrez and A. Kuijper, "A comprehensive study on face recognition biases beyond demographics," *IEEE Transactions on Technology and Society,* vol. 3, no. 1, pp. 16-30, 2021.

[8] M. Wang, Y. Zhang and W. Deng, "Meta balanced network for fair face recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 44, no. 11, pp. 8433-8448, 2021.

[9] J. Yu, X. Hao, H. Xie and Y. Yu, "Fair face recognition using data balancing, enhancement and fusion," in *In European Conference on Computer Vision*, 2020.

[10] A. Carrington, D. Manuel, P. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh and M. McInnes, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 1, pp. 329-341, 2022.

[11] P. D. Terhörst, N. Damer, F. Kirchbuchner and A. Kuijper, "On soft-biometric information stored in biometric face embeddings. IEEE Transactions on Biometrics," *Behavior, and Identity Science,* vol. 3, no. 4, pp. 519-534, 2021.

[12] C. Herlihy, K. Truong, A. Chouldechova and M. Dudík, "A structured regression approach for evaluating model performance across intersectional subgroups," in *In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.

[13] H. I. Liu, M. Galindo, H. Xie, L. K. Wong, H. H. Shuai, Y. H. Li and W. H. Cheng, "Lightweight deep learning for resource-constrained environments: A survey," *ACM Computing Surveys,* vol. 56, no. 10, pp. 1-42, 2024.

[14] P. Purwono, A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky and Q. M. ul Haq, "Understanding of convolutional neural network (cnn): A review," *International Journal of Robotics and Control Systems,* vol. 2, no. 4, pp. 739-748, 2022.

[15] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR),* vol. 54, no. 10s, pp. 1-41, 2022.

[16] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering,* vol. 2, pp. 57-64, 2021.

[17] B. K. Durga and V. Rajesh, "A ResNet deep learning based facial recognition design for future multimedia applications," *Computers and Electrical Engineering,* vol. 104, p. 108384, 2022.

[18] I. N. Alam, I. H. Kartowisastro and P. Wicaksono, "Transfer Learning Technique with EfficientNet for Facial Expression Recognition System," *Revue d'Intelligence Artificielle,* vol. 36, no. 4, 2022.

[19] F. Mosayyebi, H. Seyedarabi and R. Afrouzian, "Gender recognition in masked facial images using EfficientNet and transfer learning approach," *International Journal of Information Technology,* vol. 16, no. 4, pp. 2693-2703, 2024.

[20] B. R. Prasad and B. S. Chandana, "Human face emotions recognition from thermal images using DenseNet," *International journal of electrical and computer engineering systems,* vol. 14, no. 2, pp. 155-167, 2023.

[21] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu and Z. Yang, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence,* vol. 45, no. 1, pp. 87-110, 2022.

[22] W. Liu, F. Zhu, S. Ma and C. L. Liu, "MSPE: multi-scale patch embedding prompts vision transformers to any resolution," *Advances in Neural Information Processing Systems,* vol. 37, pp. 29191-29212, 2024.

[23] C. Liu, K. Hirota and Y. Dai, "Patch attention convolutional vision transformer for facial expression recognition with occlusion," *Information Sciences,* vol. 619, pp. 781-794, 2023.

[24] S. I. Cho, D. Kim, H. Lee, T. T. Um and H. Kim, "Explore highly relevant questions in the Baumann skin type questionnaire through the digital skin analyzer: A retrospective single-center study in South Korea," *Journal of Cosmetic Dermatology,* vol. 22, no. 11, pp. 3159-3167, 2023.

[25] C. M. M. E. P. Heldreth, A. T. Clark, C. Schumann, X. Eyee and S. Ricco, "Which skin tone measures are the most inclusive? An investigation of skin tone measures for artificial intelligence," *ACM Journal on Responsible Computing,* vol. 1, no. 1, pp. 1-21, 2024.

[26] J. H. Cabot and E. G. Ross, "Evaluating prediction model performance," *Surgery,* vol. 174, no. 3, pp. 723-726, 2023.

[27] H. Shi, Y. Zhang, Z. Shen, S. Tang, Y. Li, Y. Guo and Y. Zhuang, "Towards communication-efficient and privacy-preserving federated representation learning," *arXiv preprint,* vol. arXiv:2109.14611, 2021.

[28] W. Verkruijsse, A. Brancart, M. B. Jaffe and S. Groen, "Variability of printed Monk skin tone scales may cause misclassification of clinical study participants: caveats on printing," *Anesthesia & Analgesia,* vol. 138, no. 6, pp. e43-e44, 2024.

[29] J. Ali, M. Kleindessner, F. Wenzel, K. Budhathoki, V. Cevher and C. Russell, "Evaluating the fairness of discriminative foundation models in computer vision," in *In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

[30] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," *arXiv preprint ,* vol. arXiv:1907.05047, 2019.

[31] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[32] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 2005.

[33] K. Hermann, T. Chen and S. Kornblith, "The origins and prevalence of texture bias in convolutional neural networks," *Advances in neural information processing systems,* vol. 33, pp. 19000-19015, 2020.

[34] C. Zhang, X. Chen, W. Li, L. Liu, W. Wu and D. Tao, "Understanding deep neural networks via linear separability of hidden layers," *arXiv preprint ,* vol. arXiv:2307.13962, 2023.

[35] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Advances in neural information processing systems,* vol. 34, pp. 12116-12128, 2021.

[36] H. Zhao, Z. Lai, H. Leung and X. Zhang, "Feature learning and understanding," *Cham: Springer.,* 2020.

[37] Y. Gao, M. Zhou, D. Liu, Z. Yan, S. Zhang and D. N. Metaxas, "A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark," *arXiv preprint ,* vol. arXiv:2203.00131, 2022.

[38] R. Y. Y. H. D. Z. K. Z. S. Xiong, C. Xing, H. Zhang, Y. Lan, L. Wang and T. Liu, "On layer normalization in the transformer architecture," in *International conference on machine learning* , 2020.

[39] N. Tsumura, N. Ojima, K. Sato, M. Shiraishi, H. Shimizu, H. Nabeshima, S. Akazaki, K. Hori and Y. Miyake, "Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin," in *In ACM SIGGRAPH 2003 Papers*, 2003.

[40] R. Daneshjou, K. Vodrahalli, R. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. Swetter, E. Bailey, O. Gevaert and Mukherjee, "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Science advances,* vol. 8, no. 31, p. eabq6147, 2022.

[41] N. Rane, "Transformers for medical image analysis: Applications, challenges, and future scope," *Challenges, and Future Scope,* 2023.

[42] A. Lanitis, C. J. Taylor and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on pattern Analysis and machine Intelligence,* vol. 24, no. 4, pp. 442-455, 2002.

[43] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang and T. Liu, "On layer normalization in the transformer architecture," in *International conference on machine learning*, 2020.

[44] S. Kornblith, M. Norouzi, H. Lee and G. Hinton, "Similarity of neural network representations revisited," in *International conference on machine learning*, 2019.