DLSU Research Congress 2025
De La Salle University, Manila, Philippines
June 26 to 28, 2025

BiasX: A Feature-Based Framework for Explaining and Quantifying Gender Bias in Face Classification

Rixdon Niño R. Mape[1,*], Jerwin Glen A. Lucero[1], and Jan Wilhelm T. Sy[1]
[1] *Computer Science Department, College of Science, Bicol University*
**rixdonninorecario.mape@bicol-u.edu.ph*

Abstract: Gender bias in face classification models poses significant ethical challenges, yet traditional fairness metrics like Equalized Odds often fail to explain why these biases occur, limiting effective mitigation. This research aimed to develop and validate BiasX, a novel framework designed to quantify and explain gender bias at the facial feature level. The methodology involved creating the BiasX framework, which integrates explainable AI (XAI) techniques like class activation mapping (CAM), specifically Grad-CAM++, and facial landmark detection to calculate bias scores indicating how specific facial features contribute to misclassifications. A Python library implementing this framework and an interactive Streamlit user interface were developed to provide practical tools for analysis. Validation employed an optimized convolutional neural network (CNN) architecture, chosen for its ability to generate interpretable activation maps, trained on controlled UTKFace dataset variations with skewed gender distribution and targeted facial feature masking. Rigorous statistical analyses and software testing (unit, integration, system) confirmed the framework's reliability and effectiveness in detecting feature-specific biases. Key results demonstrated that model architecture significantly impacts heatmap interpretability and that BiasX offers distinct insights compared to Equalized Odds by showing greater stability under data skew and revealing feature-specific impacts and compensatory attention shifts during masking experiments. BiasX provides a validated, practical approach for explainable, feature-level gender bias analysis that offers crucial insights beyond aggregate fairness metrics.

Key Words: gender bias; face classification; Explainable AI (XAI); fairness metrics

1. INTRODUCTION

1.1 Background

Face classification, particularly gender categorization, stands as a cornerstone application within computer vision, seeing widespread use across security, healthcare, and social media domains (Mosayyebi et al., 2024; Zhao et al., 2024). These systems frequently utilize convolutional neural networks (CNNs) owing to their effectiveness in learning complex features from extensive datasets (Zhao et al., 2024). Although recent advancements have significantly improved classification accuracy, positioning these models as essential tools, their increasing deployment in socially sensitive areas gives rise to critical ethical concerns, especially surrounding system fairness and transparency (Kolla & Savadamuthu, 2023).

Many existing gender classification models demonstrate performance disparities across

demographic groups, an issue commonly referred to as bias. While traditional fairness metrics, such as equalized odds, can successfully identify when these models perform differently between groups, they offer no insight into the underlying reasons for these discrepancies or which specific model components contribute to the unfair outcomes (Hardt et al., 2016). This limitation is particularly evident concerning underrepresented populations, including women or individuals with darker skin tones, where models often exhibit higher error rates stemming from imbalanced training datasets (Buolamwini & Gebru, 2018). Research confirms that models predominantly trained on datasets skewed towards lighter-skinned individuals or males yield higher error rates when classifying faces from other demographic groups (Krishnapriya et al., 2020).

Although Explainable AI (XAI) techniques like class activation mapping (CAM) have enhanced the interpretability of classification models, current methods have not been effectively integrated with traditional fairness metrics to address bias comprehensively (Zhou et al., 2020). This disconnect leaves developers without adequate tools to understand which specific facial features are driving the biases detected by statistical measures. Previous research has often concentrated on general demographic biases without sufficiently tackling the nuanced biases related to gender misclassifications, particularly how specific facial characteristics might lead to these errors (Adebayo et al., 2020; Chattopadhyay et al., 2018). Bridging this significant gap requires understanding which features cause gender-specific errors, thereby providing actionable insights for improving fairness within classification models.

1.2 Objectives

This research aims to develop and validate BiasX, a novel framework that unifies statistical fairness metrics with feature-level explainability to measure and explain gender biases within face classification models. The specific objectives are to: (1) prepare face classification models and balanced facial image datasets with varied gender biases to serve as inputs for evaluating the fairness framework; (2) design a fairness framework that quantifies gender bias through feature-based analysis by considering contribution of facial features to classification decisions; (3) develop a library and interface that implements the fairness framework, allowing users to upload their pre-trained models and receive analysis report; and (4) validate the framework against Equalized Odds and the library through software testing across different face classification models and datasets.

2. METHODOLOGY

2.1 Theoretical Foundation

A quantitative foundation is established for analyzing gender bias in face classification models through core fairness criteria and metrics. This includes established fairness concepts and novel feature-based bias metrics developed for this framework. Equalized Odds provides a fundamental fairness criterion demanding equal predictive performance across gender groups, given the true outcome (Hardt et al., 2016). It ensures model accuracy consistency regardless of gender identity.

Definition 1 (Equalized Odds). For a model predictor $\widehat{Y}$, protected attribute A, where $A = 0$ is female and $A = 1$ is male, and true outcome Y, where $Y = 1$ is the positive class, Equalized Odds requires:

$$Pr(\widehat{Y} = 1|A = 0, Y = y) = Pr(\widehat{Y} = 1|A = 1, Y = y) \quad \text{(Eq. 1)}$$

where $y \in \{0, 1\}$. This mandates equal true positive rates $(Y = 1)$ and false positive rates $(Y = 0)$across genders. To measure adherence, the equalized odds score captures the maximum performance disparity:

Theorem 1 (Equalized Odds Score). For a binary classifier, the score is:

$$EO_{score} = max(|TPR_0 - TPR_1|, |FPR_0 - FPR_1|) \quad \text{(Eq. 2)}$$

where $TPR_a = Pr(\widehat{Y} = 1|A = a, Y = 1)$ and $FPR_a = Pr(\widehat{Y} = 1|A = a, Y = 0)$ are true/false positive rates for group $a \in \{0, 1\}$. The score ranges from 0 (perfect fairness) to 1 (maximum disparity).

While EO quantifies overall fairness, it doesn't identify contributing facial features. Feature probability addresses this by measuring the frequency of specific facial feature presence in misclassifications per gender.

Definition 2 (Feature Probability). For facial feature $f \in F$, its probability measures presence $(X_f = 1)$ when $\widehat{Y} \neq Y$ for gender A

$$P_f^a = Pr(X_f = 1 \mid A = a, \widehat{Y} \neq Y),\ a \in 0, 1,\ f \in F \quad \text{(Eq. 3)}$$

This links specific features to model errors for each group, indicating potential bias sources if disparities exist. Building on this, feature-specific bias quantifies the disparity in how a feature contributes to errors across genders.

Theorem 2 (Feature-Specific Bias). For feature f , the score is

$$B_f = | P_f^0 - P_f^1 |,\ f \in F \quad \text{(Eq. 4)}$$

where P_f^0 and P_f^1 are feature probabilities for females and males. A score of 0 means equal contribution to errors; higher values mean greater disparity. This identifies problematic features for targeted mitigation. To assess overall fairness across all features, an aggregate metric is used.

Corollary 1 (BiasX score). The score is the mean of feature-specific bias scores

$$\overline{B} = \frac{1}{|F|} \sum_{f \in F} B_f \quad \text{(Eq. 5)}$$

where $|F|$ is the total number of features. Ranging from 0 to 1, it allows model comparison but should be viewed with individual B_f scores to avoid masking significant disparities.

2.2 Implementation Details

The BiasX system operates through a defined data pipeline, as shown in Figure 1. Initially, model parameters provided by the AI model developer, along with facial images from a testing dataset, are processed for image classification. The resulting classification outcomes are then channeled into two parallel streams: one for generating visual explanations and another for storage in an explanation dataset. This dataset centralizes classification results and the annotated regions derived from the visual explanation process. Finally, a bias analysis process synthesizes this collected data, computing feature probabilities based on misclassifications, calculating feature-specific and BiasX scores, and culminating in a comprehensive analysis report delivered back to the Al Model Developer.

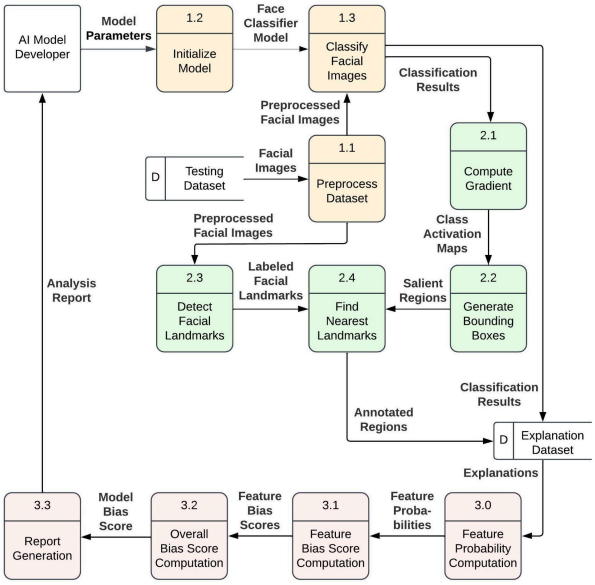

Fig. 1. Level 2 data flow diagram of BiasX system

The feature-level bias analysis uses an explainability pipeline that combines Grad-CAM++ and MediaPipe Face Mesh. Grad-CAM++ generates heatmaps to identify influential facial regions by analyzing derivatives from the final convolutional layer. MediaPipe accurately maps 468 3D facial landmarks, allowing these heatmaps to be overlaid onto anatomical features.

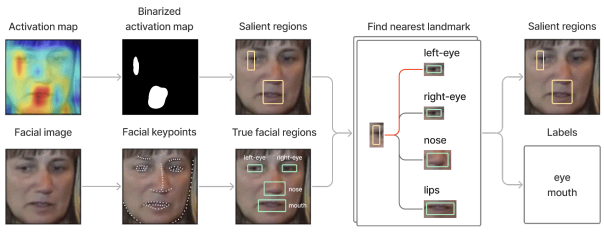

Fig. 2. Explainability pipeline for generating visual explanations of model predictions

This pipeline, depicted in Figure 2, generates a heatmap, detects facial keypoints, processes the heatmap to find significant activation areas via thresholding, and creates bounding boxes around these areas. Finally, it associates each activation box with the nearest facial landmark box if their overlap meets a set threshold, thereby labeling the activation region with a specific facial feature.

The experimental design utilized an optimized 3-block CNN architecture, indicated in Figure 3, specifically chosen for its superior ability to generate interpretable heatmaps crucial for feature-level analysis. Using this consistent architecture with the UTKFace dataset, researchers systematically generated models exhibiting predictable biases. Outcome-based bias was induced by training 45 models on datasets with deliberately skewed gender ratios (ranging from 10:90 to 90:10, with 5 replicates per condition). Attention-based bias was induced by training 40 models on datasets where specific facial features (nose, eyes, eyebrows, or lips) were programmatically masked for only one gender group (5 replicates per feature/gender condition). Baseline models were also trained on balanced data for comparison.

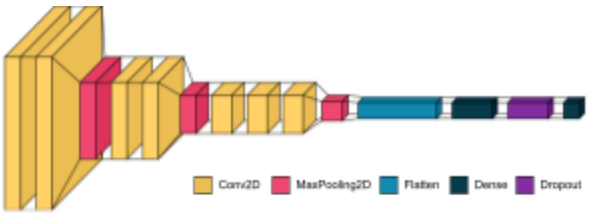

Fig. 3. Explainability pipeline for generating visual explanations of model prediction

Validation of the BiasX framework and its software implementation involved both statistical analysis and rigorous testing. Spearman's rank correlation (ρ) was used to quantitatively assess the monotonic relationship between the feature-based BiasX score and the outcome-based Equalized Odds score across the different experimental conditions. kernel density estimation was employed to examine the distribution of feature-specific bias scores for models grouped by their overall Equalized Odds performance, assessing the potential information gain from the feature-level analysis. Finally, the software library and user interface underwent comprehensive unit, integration, and system-level testing using Pytest, achieving high code coverage to confirm the system's reliability, robustness, and correctness.

3. RESULTS AND DISCUSSION

3.1 Bias Analysis

Baseline analysis across ten replicates revealed inherent bias, with statistically significant non-zero values for both Equalized Odds and BiasX ($p<0.001$ for both). As shown in Figure 4, Equalized Odds ranged from 0.027 to 0.147 (mean 0.087), while BiasX ranged from 0.085 to 0.162 (mean 0.117). These results confirm biases unlikely due to chance in the baseline model.

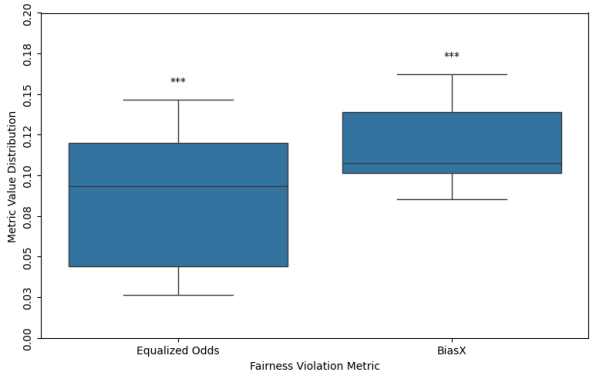

Fig. 4. Baseline values of Equalized Odds score and BiasX score

Examining model predictions requires understanding which facial features receive attention and if this reliance is consistent across genders for correct and incorrect classifications. For correct predictions, feature attention patterns differed by gender, as shown in Figure 5. Female classifications focused most on the chin (mean attention 0.595) and lips (0.595), while male classifications prioritized the nose

(0.657) and eyes (0.630). Attention to features like the nose, eyes, eyebrows, and forehead was higher for males, whereas cheek and chin attention was higher for females, indicating inconsistent use of key features for accurate predictions across genders.

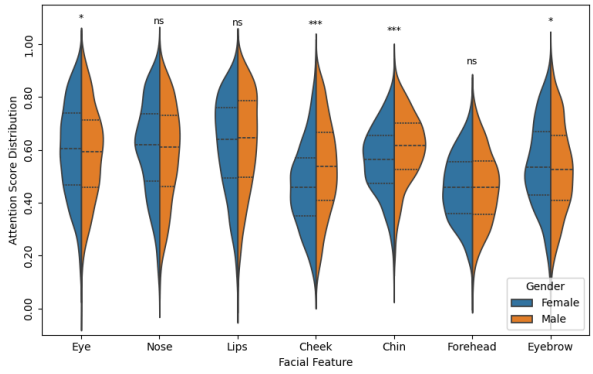

Fig. 5. Feature attention for correct predictions

Attention patterns shifted during misclassifications, as seen in Figure 6. Incorrectly classified females showed high attention to lips (0.618) and nose (0.608), while incorrectly classified males focused on lips (0.630) and chin (0.610). Although lips were prominent in errors for both genders, other key features remained inconsistent. Comparing correct versus incorrect predictions revealed dynamic changes; for example, nose attention increased significantly for females only during errors but decreased for males.

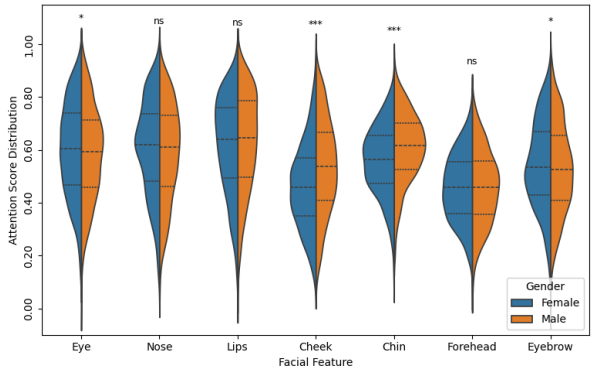

Fig. 6. Feature attention for incorrect predictions

When facial features were masked for males, the model shifted attention to remaining visible features, as shown in Figure 7. Masking the eye significantly increased attention to the chin (+0.13 difference from baseline) and cheek (+0.03). Masking the lips led to a broader redistribution, modestly increasing attention to the cheek, eye, eyebrow, and chin. Masking the nose increased cheek attention (+0.06), while masking the eyebrow resulted in minimal compensatory shifts.

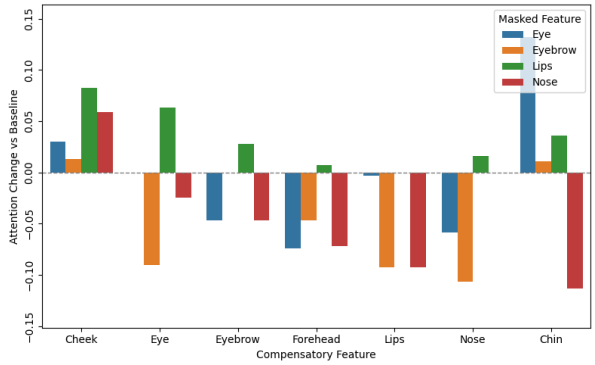

Fig. 7. Compensatory attention change for males

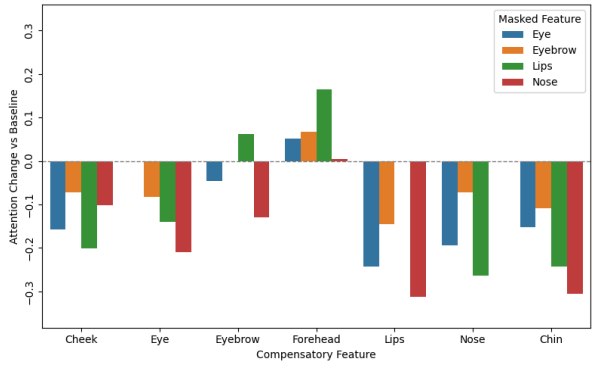

Fig. 8. Compensatory attention change for females

For females, masking induced different compensatory patterns, indicated in Figure 8. The forehead consistently emerged as a key compensatory region. Masking the eye (+0.05) or eyebrow (+0.07) increased forehead attention. Masking lips caused the largest shift, substantially increasing forehead attention (+0.16) and slightly increasing eyebrow attention (+0.06). Thus, masking prompts increased attention to

other features, but these compensatory features differ by gender: males rely more on the chin and cheek, while females primarily shift focus to the forehead.

3.2 Library and Interface Development

The BiasX Python library offers a practical toolkit for feature-level gender bias analysis in face classification models. Key functionalities include quantifying bias contributions of specific features, generating visual explanations using CAM and landmarks, and calculating BiasX and Equalized Odds scores. It supports Keras models and common datasets like UTKFace. For accessibility, the MIT-licensed library (v0.1.3) is on PyPI (pip install biasx) and GitHub (rixmape/biasx), with comprehensive MkDocs documentation. Rigorous unit, integration, and system testing confirmed its reliability and robustness, establishing BiasX as a tool for the research community.

The BiasX interface provides a structured workflow: users first upload or select a model, then adjust parameters for the model, explanation method, and dataset on a central configuration page before initiating the analysis. Results are displayed on a visualization page with three tabs. The "Feature Analysis" tab shows BiasX and Equalized Odds scores alongside feature-specific bias visualizations. The "Model Performance" tab presents metrics like confusion matrices, gender-specific precision/recall/F1 scores, and ROC/PR curves. The "Image Analysis" tab allows detailed inspection of individual samples with heatmap overlays.

3.2 Framework and Tool Validation

Correlation analysis revealed a weak, insignificant link between BiasX and Equalized Odds (Spearman's ρ=-0.138, p=0.364), as shown in Figure 8. BiasX scores remained relatively stable under varying gender distributions, while Equalized Odds scores escalated sharply with data skew. This divergence highlights their distinct measurement focuses: BiasX quantifies feature-level attention disparities, offering complementary insights to Equalized Odds' measure of error rate parity and demonstrating robustness to outcome biases from data imbalance.

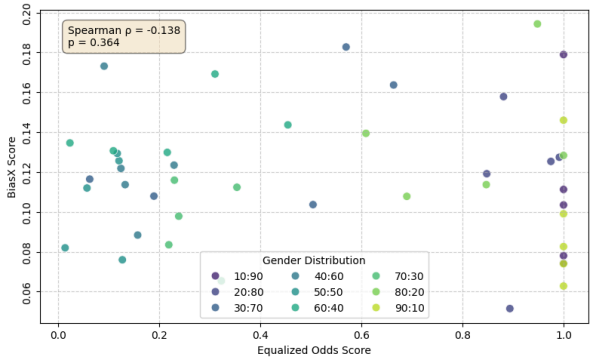

Fig. 8. Spearman's rank correlation between BiasX and Equalized Odds scores

Kernel density estimation examined feature bias score distributions across models grouped by low, medium, or high Equalized Odds scores, as shown in Figure 9. While distributions largely overlapped, the high Equalized Odds group showed a modest shift towards higher feature bias scores. This indicates models with similar aggregate fairness scores can still exhibit varied feature-specific biases, confirming that feature-level analysis provides crucial granular insights into bias nature and location beyond what Equalized Odds alone captures.

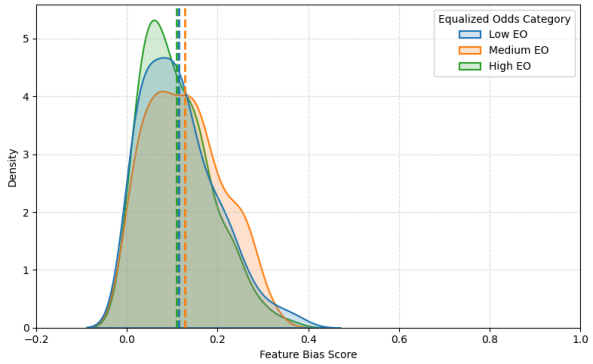

Fig. 9. Information gain analysis using kernel density estimation

The BiasX framework's reliability was validated through rigorous multi-level testing using Pytest, with all 182 test cases passing. Unit testing (122 tests) verified individual components with 100% code coverage. Integration testing (28 tests) confirmed component interactions with 84% coverage. System-level testing (32 tests) assessed end-to-end functionality and robustness across various scenarios, achieving 91% coverage. This comprehensive process confirmed the framework's dependability and readiness for application.

4. CONCLUSIONS

This research designed, implemented, and validated the BiasX framework to address the explanatory gap in traditional fairness metrics for gender face classification. By combining explainability techniques like Grad-CAM++ and landmark detection, BiasX successfully measured feature-level bias contributions. The study identified an optimal CNN architecture for interpretable visualizations and rigorously tested the Python library and Streamlit interface that enable deeper understanding of how bias manifests beyond simple detection.

Future work should explore alternative model architectures and more diverse datasets, expand analysis to intersectional biases and compensatory attention mechanisms, and enhance tool deployment and usability. Applying BiasX to real-world systems, conducting user studies, and leveraging its feature-level insights to guide targeted bias mitigation strategies are also crucial next steps for advancing fair and explainable AI.

5. ACKNOWLEDGMENTS

This work benefited significantly from the guidance and expertise of Dr. Lea D. Austero and Professor Arlene A. Satuito, as thesis advisers.

6. REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2020). Sanity Checks for Saliency Maps (No. arXiv:1810.03292). arXiv. http://arxiv.org/abs/1810.03292

Anwer, F., Aftab, S., Waheed, U., & Muhammad, S. S. (2017). Agile software development models tdd, fdd, dsdm, and crystal methods: A survey. International journal of multidisciplinary sciences and engineering, 8(2), 1-10.

Buolamwini, J., & Gebru, T. (2018.). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.

Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/WACV.2018.00097

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

Kolla, M., & Savadamuthu, A. (2023). The Impact of Racial Distribution in Training Data on Face Recognition Bias: A Closer Look. 313–322. https://doi.org/10.1109/WACVW58289.2023.00035

Krishnapriya, K. S., Albiero, V., Vangara, K., King, M. C., & Bowyer, K. W. (2020). Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone. IEEE Transactions on Technology and Society, 1(1), 8–20. https://doi.org/10.1109/TTS.2020.2974996

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A framework for

building perception pipelines. arXiv. https://doi.org/10.48550/arXiv.1906.08172

Mosayyebi, F., Seyedarabi, H., & Afrouzian, R. (2024). Gender recognition in masked facial images using EfficientNet and transfer learning approach. International Journal of Information Technology, 16(4), 2693–2703. https://doi.org/10.1007/s41870-023-01565-4

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4), 99. https://doi.org/10.1007/s10462-024-10721-6

Zhou, J., Chen, F., & Holzinger, A. (2020, July). Towards explainability for AI fairness. In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 375-386). Cham: Springer International Publishing.Abbas, Z. H., & Shaker, S. H. (2022). The Transfer Learning Models for Face Recognition: A Survey. 2022 2nd International Conference on Advances in Engineering Science and Technology (AEST), 764–768. https://doi.org/10.1109/AEST55805.2022.10412930