ARTIFICIAL INTELLIGENCE RESEARCH and APPLICATIONS
W. Z. Tareq, Vol.1, No.3, September, 2025 |Research Article|

Exploring Gender Classification Performance on AI-Generated Facial Datasets Using Transfer Learning Models

Wadhah Zeyad Tareq *‡

* Computer Engineering Department, Faculty of Engineering and Natural Sciences, İstinye University, İstanbul, Türkiye
(wadhah.tareq@istinye.edu.tr)

‡ Computer Engineering Department, Faculty of Engineering and Natural Sciences, İstinye University, İstanbul, Türkiye
wadhah.tareq@istinye.edu.tr

Received: 06.08.2025 Accepted: 02.09.2025

Abstract- The demand for gender classification has increased significantly with the growth of smart and automated security systems. Artificial Intelligence (AI), particularly Deep Learning (DL), has emerged as a promising approach for building reliable classification systems. Despite advancements in classification techniques, the availability of high-quality datasets for training and testing remains a notable challenge. In this work, we propose utilizing a synthetic facial dataset to train several well-known classification models. We evaluate different Convolutional Neural Network (CNN) architectures, including transfer learning approaches, on a set of unreal face images generated entirely using a Generative Adversarial Network (GAN). To the best of our knowledge, this is among the few works that investigate gender classification trained exclusively on fully synthetic GAN-generated datasets, highlighting its novelty. The models achieved strong performance on the synthetic dataset, with up to 99% training accuracy and 95% test accuracy. However, as a limitation, the generalization of these results to real-world datasets remains uncertain, since synthetic images may not capture all demographic and natural variations. This study demonstrates the viability of using artificially generated data when real data is scarce or difficult to obtain.

Keywords- Gender classification, synthetic datasets, transfer learning, ResNet50, MobileNetV2, InceptionV3.

1. Introduction

Human gender is one of the most significant attributes in modern automated vision tasks. Many real-world applications perform differently based on an individual's gender—targeted advertising being a prime example. In product recommendation systems, gender plays a key role in guiding users toward suitable products or services [1]. Deep learning (DL), a subfield of machine learning, currently dominates computer vision tasks such as detection and classification. These models can automatically learn image features through powerful discriminative architectures. Over the years, several deep learning models have been proposed to achieve high accuracy in image-based tasks. Among the most well-known are AlexNet [2], VGG-16 [3], GoogleNet (Inception-v1) [4], Inception-v3 [5], ResNet [6], UNet [7], and MobileNet [8].

Numerous studies have investigated gender classification using DL models. Rajee and Mythili [9] applied ResNet50 to digital dental X-ray images, achieving 98.27% accuracy on a dataset of 1,000 images. Abbas et al. [10] proposed a 64-layer architecture called 4-BSMAB, based on AlexNet, and attained 93% accuracy on the CIFAR-100 dataset using a SoftMax classifier. Sharma et al. [11] introduced a SegNet-based model combined with SVM for age and gender classification, reporting accuracies of 92.48% on FG-Net for age and 95.1% on IOG for gender. Similarly, Singh et al. [12] developed a hybrid model using self-attention and BiLSTM, achieving performance improvements of approximately 10% and 6% over existing age and gender classification methods, respectively.

For any deep learning system to be effective and reliable, the availability of large, diverse, and well-annotated datasets is essential [13]. However, collecting such datasets for facial analysis is challenging due to privacy concerns. Generative Adversarial Networks (GANs) [14] have shown remarkable potential in generating synthetic data that mimics real-world variations. However, despite this progress, synthetic datasets

have not yet been widely adopted for gender classification tasks, as their ability to generalize to real-world scenarios remains largely untested. The website This Person Does Not Exist [15] utilizes NVIDIA's StyleGAN [16] (and more recently, StyleGAN2 [17]) to generate highly realistic and diverse human faces by disentangling semantic attributes such as pose and identity. In this study, we address the data scarcity challenge by employing a synthetic face dataset sourced entirely from This Person Does Not Exist. We then train and evaluate four deep learning models to construct a gender classification system. It is also important to acknowledge the ethical challenges associated with gender classification, such as potential privacy risks, stereotyping, and the risk of misuse in real-world applications. In addition, both real and synthetic datasets may contain demographic biases that could influence model performance and fairness, as highlighted by recent work showing that diffusion-based face generation models can amplify demographic imbalances [18]. More recently, Korshunov et al. [19] demonstrated that while synthetic data still underperforms compared to real data, demographically balanced synthetic datasets, particularly those generated with Stable Diffusion v3.5, can reduce bias in face recognition benchmarks, emphasizing the importance of dataset composition in fairness outcomes. By addressing these gaps, this work contributes to the growing body of research on the feasibility of synthetic data, while also recognizing the limitations and ethical implications of applying gender classification systems in practice. The objectives of this work are summarized as follows:

- To construct a gender classification system using deep learning models trained exclusively on AI-generated facial images.
- To compare the performance of custom CNN and transfer learning models (ResNet50, InceptionV3, and MobileNetV2) on the synthetic dataset.
- To evaluate the feasibility and effectiveness of using synthetic face data as a substitute for real-world datasets in training gender classification systems.

The remainder of this paper is organized as follows: Section 2 describes the synthetic dataset used in this study. Section 3 outlines the methodology, covering the deep learning models employed, training configurations, and evaluation metrics. Section 4 presents the experimental results and discusses the comparative performance of the models. Finally, Section 5 concludes the paper and highlights potential directions for future research.

2. Dataset Description

The dataset of synthetic face images was sourced from a publicly available Kaggle dataset [20]. The images were manually collected and labeled by the dataset's author. In this study, the labels were directly inherited from the dataset creator and were not independently re-verified by us. As seen in Figure 1, it contains two classes Female and Male with a total of 6,873 images: 3,860 labeled as Female and 3,013 as Male. The dataset was divided into three subsets: training, validation, and test sets, with a split ratio of 70%, 15%, and 15%, respectively.

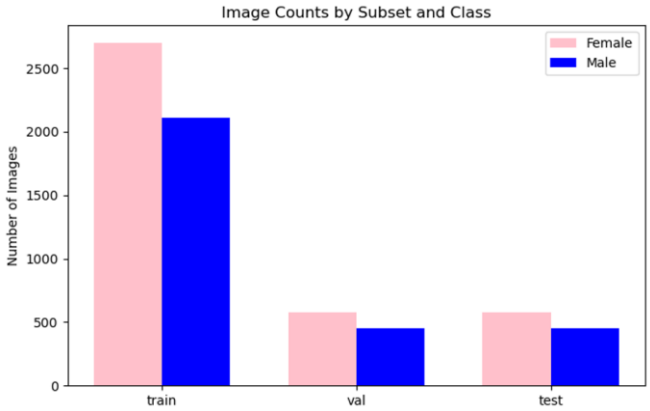

Fig. 1. Distribution of female (3,860) and male (3,013) images across the training (70%), validation (15%), and test (15%) subsets.

Prior to training, all images were normalized by rescaling pixel values to the [0, 1] range using a Rescaling layer. Additionally, data augmentation was applied only to the training set using random horizontal flipping, slight rotation, and zoom transformations. This aimed to improve generalization and reduce overfitting. All datasets were preprocessed and prefetched to optimize GPU utilization during training.

3. Methodology

This section outlines the deep learning models used for gender classification on synthetic face images. We implemented one custom CNN architecture and three pre-trained transfer learning models: ResNet50, InceptionV3, and MobileNetV2.

3.1. CNN

As a baseline, we constructed a custom Convolutional Neural Network (CNN) tailored for binary gender classification. The architecture begins with a data augmentation layer to improve generalization, followed by three convolutional blocks. Each block includes a Conv2D layer with ReLU activation and a MaxPooling2D layer to progressively reduce spatial dimensions while capturing key features. The convolutional layers use 32, 64, and 128 filters, respectively. After feature extraction, the output is flattened and passed through a dense layer with 128 units and ReLU activation. A dropout layer with a 0.5 rate is applied to mitigate overfitting. Finally, a single-neuron output layer with a sigmoid activation is used for binary classification. The model was compiled with the Adam optimizer, binary cross-entropy loss and trained for 20 epochs. Figure 2 illustrates the complete architecture of the CNN model, showing the flow from data augmentation and convolutional layers to the fully connected layers and final output neuron.

3.2. Transfer Learning Models

Transfer learning [21] is a machine learning technique where a model developed for one task is reused as the starting point for a new task.

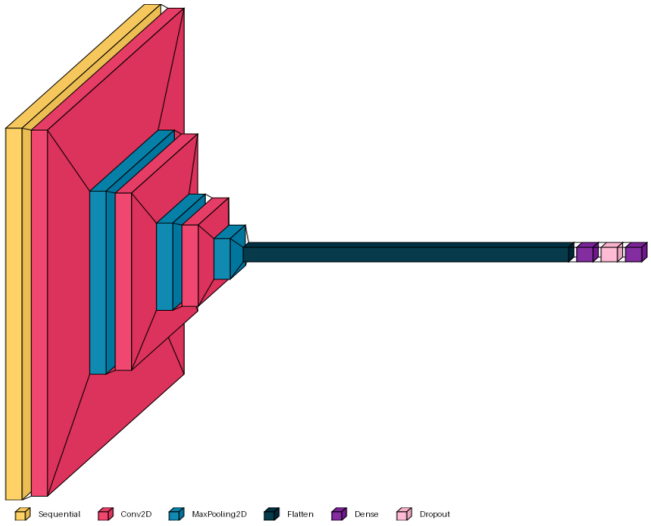

Fig. 2. CNN architecture showing the sequence of data augmentation, convolutional, pooling, flatten, dense, and dropout layers used for binary gender classification.

It significantly accelerates training time and reduces the amount of required training data, making it particularly effective in domains with limited labeled samples. In this study, we employed three widely used transfer learning architectures: ResNet50, InceptionV3, and MobileNetV2.

InceptionV3 [5] is the third generation of Google's Inception series and incorporates several innovations, such as factorized 7×7 convolutions to reduce parameters, RMSProp optimizer, batch normalization in auxiliary classifiers, and label smoothing to mitigate overfitting. It has demonstrated strong performance in diverse image classification tasks and has been used as a base architecture by researchers. ResNet [6] introduced the revolutionary concept of residual connections (skip connections), enabling the construction of ultra-deep networks by addressing the vanishing gradient problem. It also incorporates batch normalization to stabilize training. ResNet models have been designed with depths of up to 152 layers without sacrificing generalization. MobileNetV2 [8] is lightweight, yet powerful architecture designed for mobile and resource-constrained environments. It uses depth wise separable convolutions followed by pointwise convolutions to drastically reduce model complexity while maintaining accuracy.

For the transfer learning models (ResNet50, InceptionV3, and MobileNetV2), we adopted an unfreezing strategy in which the convolutional base was initially kept frozen during the early training phases and later unfrozen to allow fine-tuning on the synthetic dataset. The original classifier head of each pre-trained model was replaced with a custom dense block consisting of a global average pooling layer, a fully connected dense layer with ReLU activation, a dropout layer to prevent overfitting, and a final sigmoid layer for binary classification. During training, all models were compiled with the Adam optimizer and binary cross-entropy loss and trained for 10 epochs using the same dataset split for fair comparison. Additionally, data augmentation including random horizontal flips, slight rotations, and zoom transformations was applied to the training set to improve generalization and reduce overfitting.

4. Experiments and Results

We evaluated the performance of four deep learning models on the synthetic face dataset: a custom CNN, ResNet50, InceptionV3, and MobileNetV2. Each model was trained using the same training, validation, and test splits to ensure consistency across experiments. The CNN model was trained for 20 epochs, while the transfer learning models (ResNet50, MobileNetV2, and InceptionV3) were each trained for 10 epochs.

During training, we tracked accuracy and loss on both training and validation sets and visualized these metrics to analyze model convergence and generalization. The CNN achieved a test accuracy of 92%, with a macro-averaged F1-score of 0.91, precision of 0.91, and recall of 0.92. The weighted averages were also consistent at 0.92, reflecting balanced performance despite the dataset's slight class imbalance. The ResNet50 model performed best, achieving 95% test accuracy, with precision, recall, and F1-score all reaching 0.95. The superior performance of ResNet50 can be attributed to its architectural depth and the use of residual (skip) connections, which enable more efficient gradient flow during training and allow the network to learn richer feature representations compared to shallower or lightweight models such as MobileNetV2 and InceptionV3. The MobileNetV2 and InceptionV3 achieved 91% accuracy, showing solid performance and efficient training with significantly fewer parameters.

We report both macro-averaged metrics, which give equal importance to each class, and weighted-average metrics, which account for class imbalance. In this paper, we use weighted averages to summarize model performance, as they more accurately reflect overall classification effectiveness given the dataset distribution.

Figures 3, 5, 7, and 9 present the training and validation curves for each model, while Figures 4, 6, 8, and 10 show the corresponding confusion matrices. These visualizations confirm that all models converged effectively, with no significant overfitting. A complete summary of test set results is presented in Table 1 where ResNet50 achieved the highest accuracy and F1-score, while all models demonstrated strong classification performance on the synthetic face dataset.

Table 1. Test set performance of the four evaluated models using weighted average metrics.

Model	Train Accuracy	Test Accuracy	Weighted Average		
			Precision	Recall	F1-score
CNN	0.92	0.92	0.92	0.92	0.92
MobileNet V2	0.98	0.91	0.91	0.91	0.91
Inception V3	0.97	0.91	0.92	0.91	0.91
ResNet50	**0.99**	**0.95**	**0.95**	**0.95**	**0.95**

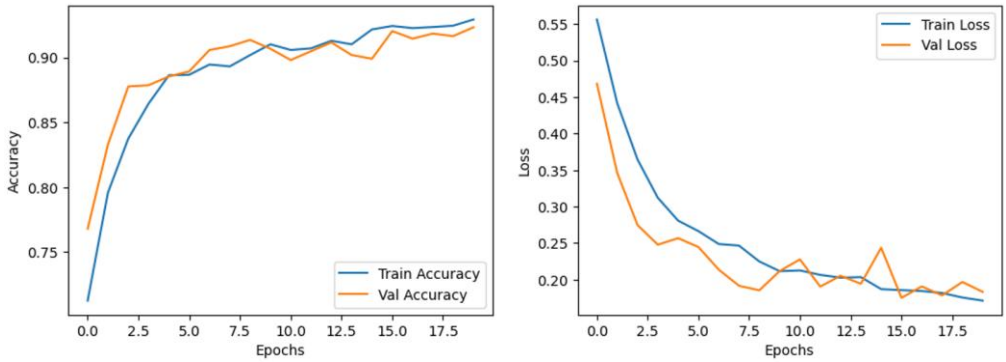

Fig. 3. Training and validation accuracy and loss curves for the CNN model over 20 epochs.

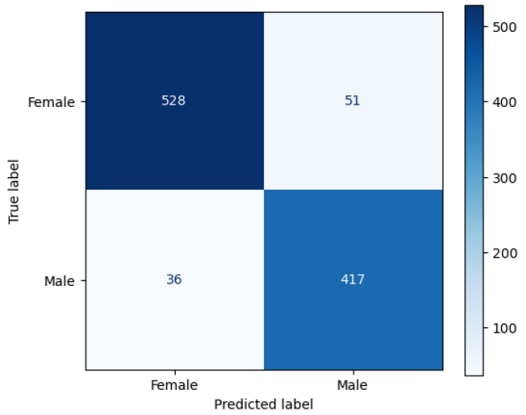

Fig. 4. Confusion matrix of the CNN model on the test set, illustrating the classification distribution between female and male classes.

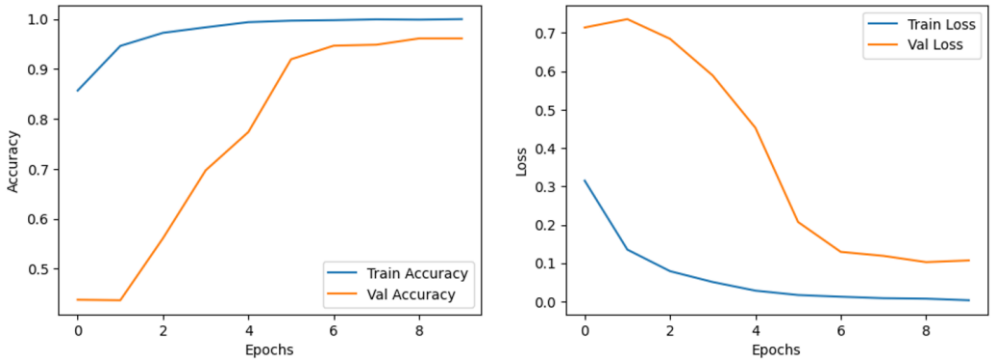

Fig. 5. Training and validation accuracy and loss curves for the ResNet50 model over 10 epochs.

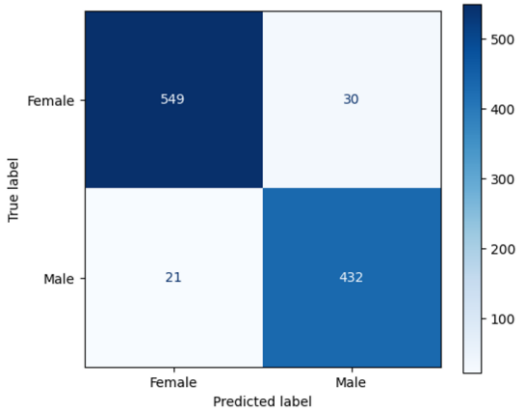

Fig. 6. Confusion matrix of the ResNet50 model on the test set, illustrating the classification distribution between female and male classes.

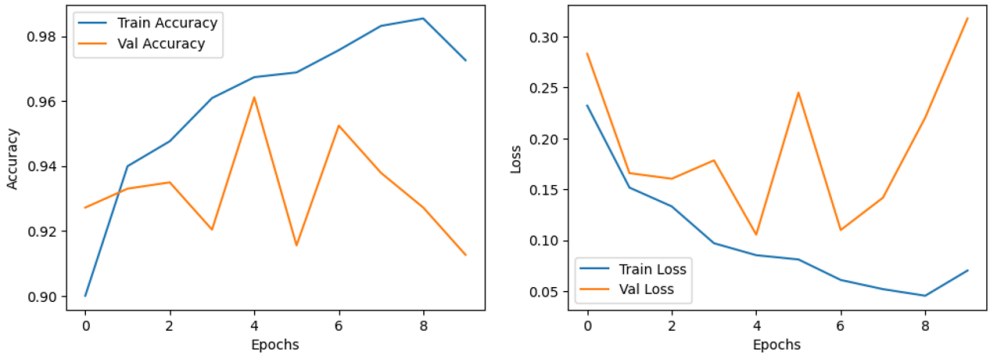

Fig. 7. Training and validation accuracy and loss curves for the InceptionV3 model over 10 epochs.

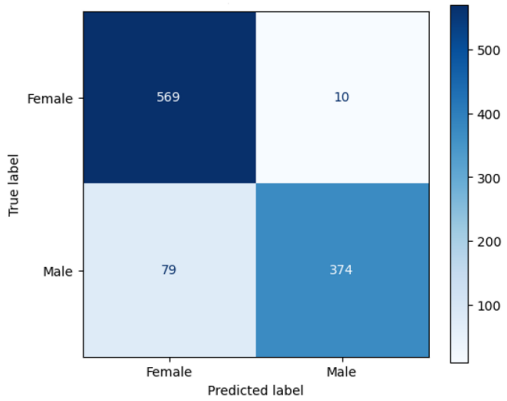

Fig. 8. Confusion matrix of the InceptionV3 model on the test set, illustrating the classification distribution between female and male classes.

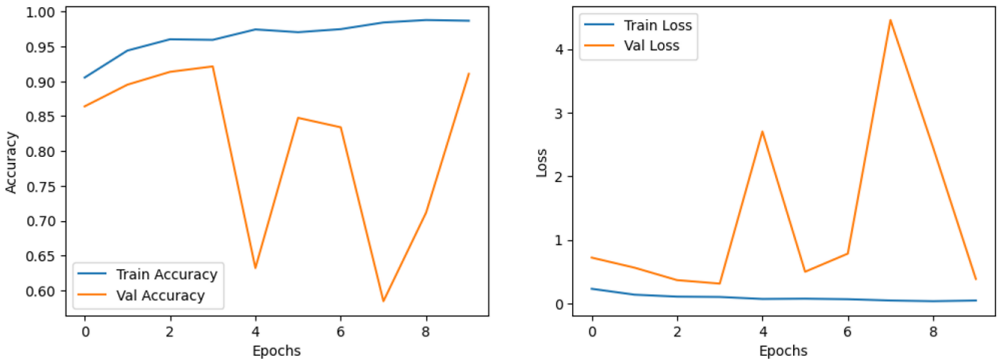

Fig. 9. Training and validation accuracy and loss curves for the MobileNetV2 model over 10 epochs.

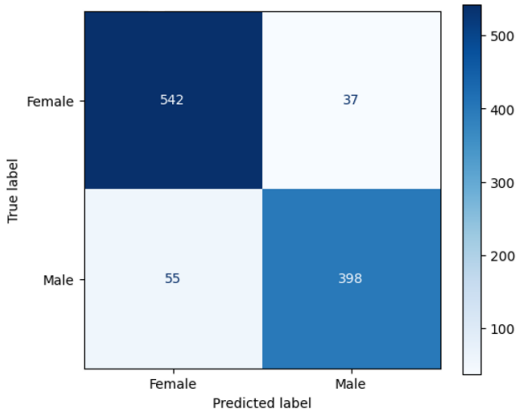

Fig. 10. Confusion matrix of the MobileNetV2 model on the test set, illustrating the classification distribution between female and male classes.

5. Discussion

The results presented in this study demonstrate that synthetic GAN-generated facial datasets can support the development of effective gender classification systems. The strong performance of ResNet50, achieving 95% test accuracy, highlights the potential of transfer learning when applied to artificially generated data. However, several aspects require careful consideration. First, although synthetic datasets offer advantages such as privacy protection and ease of data collection, their generalization to real-world applications remains uncertain. Gender classification models trained solely on synthetic faces may not transfer seamlessly to real images, as synthetic data may not fully capture the diversity and subtle variations present in natural populations.

Second, the use of GAN-generated data raises potential concerns regarding bias. Since GANs are trained on pre-existing datasets, they may inherit or even amplify demographic imbalances related to age, ethnicity, or other facial attributes. Such biases could lead to fairness and reliability issues if models trained on synthetic data are deployed in practice. Third, this study did not include a direct comparison with real-world benchmarks such as CelebA or IMDB-WIKI. While such comparisons are common in literature, the primary objective here was to investigate the feasibility of fully synthetic datasets as a training source. Real-world validation is therefore considered an important direction for future research.

Finally, the ethical implications of gender classification technologies should not be overlooked. Even when technical performance is strong, the deployment of gender recognition systems in real-world settings may pose privacy risks, enable profiling, or lead to misuse. Researchers and practitioners should remain cautious about the context in which such systems are applied.

6. Conclusion

This study explored gender classification using deep learning models trained on a fully synthetic facial image dataset generated by the This Person Does Not Exist platform. Four models were evaluated—a custom CNN and three transfer learning approaches (ResNet50, MobileNetV2, and InceptionV3). Among these, ResNet50 achieved the highest performance, with a test accuracy of 95%, followed by the CNN and other models at 91–92%. These results demonstrate that GAN-generated facial datasets can be effectively used to train robust classifiers, offering a practical alternative when access to real data is limited due to privacy or availability constraints.

Despite these promising findings, this work has certain limitations. In particular, the experiments were conducted exclusively on synthetic images, and generalization to real-world datasets remains an open question. Furthermore, synthetic datasets may contain hidden biases related to age, ethnicity, or facial diversity, which could affect fairness and reliability. Future research should investigate the use of hybrid datasets that combine real and synthetic images, apply domain adaptation techniques to improve cross-dataset generalization, and scale the approach with larger and more diverse synthetic datasets. Ethical considerations also need to be addressed carefully, as gender classification systems may raise privacy, fairness, and potential misuse concerns in real-world applications.

References

[1] M. S. Devi, S. Priya, S. Singh, and R. Aruna, “Defusing dirty label backdoor attack with differentially private data through feature vectorized AlexNet for face gender classification,” IEEE Access, doi: 10.1109/ ACCESS.2024.3485804, vol. 12, pp. 159099-159120, 2024.

[2] F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Advances in Neural Information Processing Systems 25. New York: Curran Associates Inc., 2012, pp. 1097–1105.

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2015.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1–9.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception architecture for computer vision,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Cham: Springer, 2015, pp. 234–241.

[8] A. G. Howard, M. Zhu, B. Chen, et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.

[9] M. V. Rajee and C. Mythili, “Gender classification on digital dental x-ray images using deep convolutional neural network,” Biomed. Signal Process. Control, vol. 69, p. 102939, 2021.

[10] F. Abbas, M. Yasmin, M. Fayyaz, M. Abd Elaziz, S. Lu, and A. A. A. El-Latif, “Gender classification using proposed CNN-based model and ant colony optimization,” Mathematics, vol. 9, no. 19, p. 2499, 2021.

[11] S. Kumar, S. Singh, J. Kumar, and K. M. V. V. Prasad, “Age and gender classification using Seg-Net based

architecture and machine learning," Multimed. Tools Appl., pp. 1–24, 2022.

[12] A. Singh and V. K. Singh, "A hybrid transformer–sequencer approach for age and gender classification from in-wild facial images," Neural Comput. Appl., doi: 10.1007/s00521-023-09087-7, vol. 36, pp. 1149–1165, 2024.

[13] M. Culman, S. Delalieux, B. Beusen, B. Somers, "Automatic labeling to overcome the limitations of deep learning in applications with insufficient training data: A case study on fruit detection in pear orchards," Comput. Electron. Agric., vol. 213, p. 108196, 2023.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial networks," Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[15] https://thispersondoesnotexist.com. (last accessed on Aug. 1, 2025].

[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4401–4410.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 8110–8119.

[18] M. V. Perera and V. M. Patel, "Analyzing bias in diffusion-based face generation models," in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Ljubljana, Slovenia, doi: 10.1109/IJCB58384.2023.1028571, Sep. 2023, pp. 1–10,

[19] P. Korshunov, K. Kotwal, C. Ecabert, V. Vidit, A. Mohammadi, and S. Marcel, "Investigation of accuracy and bias in face recognition trained with synthetic data," arXiv preprint arXiv:2507.20782, Jul. 2025.

[20] https://www.kaggle.com/datasets/bwandowando/all-these-people-dont-exist/data (last accessed on Aug. 1, 2025).

[21] A. H. M. Linkon, A. E. Hassanien, M. S. Khan, and A. B. Khandoker, "Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study," Informatics Med. Unlocked, vol. 24, p. 100582, 2021.