F. Boutros, N. Damer, M. Fang, M. Gomez-Barrero, P. Neto,
K. Raja, C. Rathgeb, A. Sequeira, and M. Todisco (Eds.): BIOSIG 2025,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2025

# Gender Classification in NIR Periocular Images: A Vision Transformer Approach

Tomaz Silva[1]

**Abstract:** Gender classification using periocular images remains a challenging task, especially when using near-infrared imaging. Existing studies based on such data are limited and often achieve results that fall short of optimal performance. In this work, a pre-trained Vision Transformer architecture was employed, using transfer learning to specialize on a small near-infrared periocular dataset. The best-performing configuration achieved an accuracy of 95.3%, outperforming prior approaches using similar data. The resulting findings suggest that the periocular region contains more gender-discriminative features than the iris alone, and that using both left and right eye regions from each subject yields the highest performance.

**Keywords:** Gender classification, Periocular, Vision Transformer.

## 1 Introduction

Periocular biometric systems have recently attracted significant attention due to their potential to complement and improve existing iris-based biometric systems.

Current iris recognition systems are mostly based on Daugman's foundation work [Da93], which involves encoding iris features and computing the distance between a biometric enrollee and a claimant's iris image. However, gender classification using the periocular region remains relatively underexplored, largely due to the scarcity of publicly available near-infrared (NIR) periocular datasets annotated with gender labels. Most existing datasets are primarily designed for biometric identification or verification tasks.

Nonetheless, gender classification is still quite relevant in topics like security, forensic science, demographic analytics, and targeted marketing.

For instance, in large-scale biometric identification systems, recognition time can increase substantially as the number of enrolled users grows. By first classifying users into gender categories, such as Male or Female, the search space can be reduced by approximately 50%, which significantly improves efficiency.

Even in systems where biometric identification is not performed, gender classification remains valuable. For example, it can support access control in gender-restricted areas or facilitate anonymity in surveys and data collection by enabling gender inference without revealing any personally identifiable information.

[1] Technical University of Denmark, DTU Compute, Fysikvej 315, Kongens Lyngby, tomaz.goncalves.silva@gmail.com

Previous studies have investigated gender classification using both iris texture and periocular regions. Bansal et al. [BAS12] trained an SVM using statistical and texture features, while Tapia et al. [TFPB16] concluded that gender-related information is distributed across the iris rather than being concentrated in concentric bands.

Other works, from Bobeldyk and Ross [BR16] and Kuehlkamp and Bowyer [KB19], suggest that the periocular region contains more discriminative information for gender classification than the iris itself. Deep learning approaches have also been explored: in Tapia and Aravena [TA18], separate Convolutional Neural Networks (CNNs) were trained for the left and right eyes and then combined, achieving the best performance, 87.26%. In Viedma et al.[Vi19] different regions of NIR periocular images were used to train the XgBoost classifier achieving 89.22%. More recently, Suravarapu and Patil [SP23] used visible-light periocular images to train a Vision Transformer (ViT) for biometric recognition and gender classification, achieving 99.04%.

Among the previously mentioned studies, only Tapia and Aravena [TA18] and Viedma et al [Vi19] utilize NIR images, which is a notable advantage, given that most operational iris biometric systems also rely on NIR imaging. This makes integration with existing systems more feasible and practical.

This paper extends previous research by investigating how a ViT model, trained on 4 different versions of a small NIR periocular dataset, can achieve competitive performance in gender classification. In particular, the impact of data preprocessing, augmentation, and model tuning on classification accuracy is explored.

The remainder of this paper is structured as follows: In Section 2, there first is an overview of ViT architectures and the reason behind selecting them for this task is explained. Then the dataset used is described as well as the four variations created from it. And finally, the hyperparameter tuning process and evaluation metrics are outlined. Section 3 reports the experimental results, accompanied by tables and analysis. At last, Section 4 provides a discussion of the findings and some reflections on the limitations and potential future directions of this work.

## 2 Methodology

It is proposed to use a fine-tuned ViT model pre-trained on ImageNet-21k (14 million images, 21,843 classes) to achieve gender classification from Periocular NIR images.

ViTs were introduced in 2020 with the paper “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale” by Dosovitskiy et al. [Do21], which showed that transformer-based architectures, originally designed for natural language processing, can outperform CNNs on image classification tasks, provided they are trained on sufficiently large datasets.

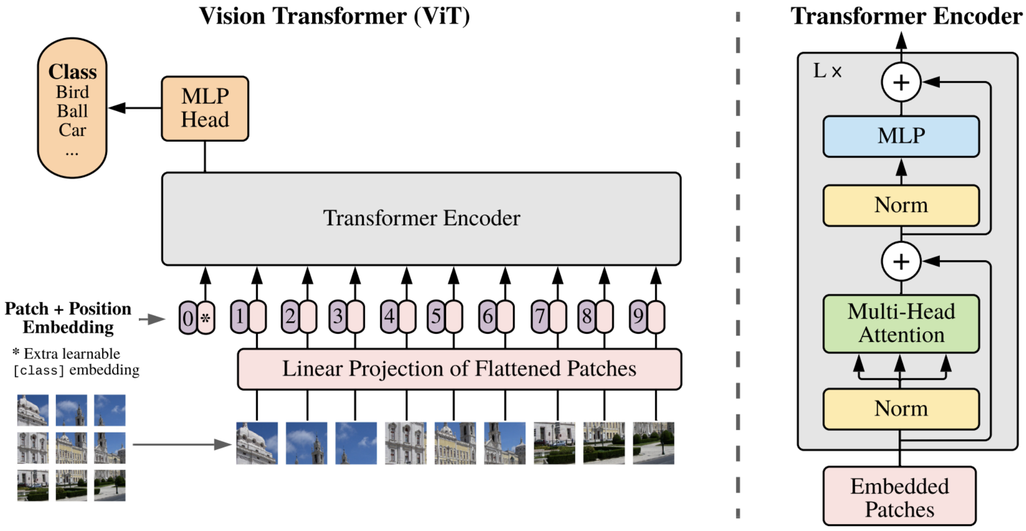


Fig. 1: Visualisation of ViT architecture, taken from [Do21]

### 2.1 Network Architecture

The Vision Transformer architecture, which can be visualised in Figure 1, segments an input image into a sequence of fixed-size, non-overlapping patches (e.g., 16×16 pixels). Each patch is then flattened and linearly projected into an embedding vector. These patch embeddings are combined with positional embeddings to retain the position of the patch in the image and then are passed through multiple transformer encoder layers.

Each encoder consists of a multi-head self-attention mechanism followed by feed-forward neural networks. These attention mechanisms are key to the ViTs' perfomance as they are responsible for ensuring different parts of the image can be correlated.

The output of the encoders is then normalized and propagated through a neural network, with 2 output neurons, corresponding to the Male and Female genders.

To convert the resulting logits from each neuron into class probabilities, the softmax function is applied, producing for each, values between 0 and 1 that sum to 1.

Unlike CNNs, ViTs typically require extensive pre-training on large-scale datasets due to their lack of inductive biases of locality [Ra24]. For this work, the `google/vit-base-patch16-224` [3] model was selected because it is pre-trained on a very large dataset (ImageNet-21k), making it well-suited for transfer learning. This is particularly advantageous when working with smaller datasets, such as the one used in this work.

[3] `https://huggingface.co/google/vit-base-patch16-224`

### 2.2 Dataset

The dataset used is composed of 3000 NIR (640×480 pixels) periocular images of 1500 individuals. Each individual has a right eye and left eye images, which results in a total of 1500 left eye and 1500 right eye images. The distribution of men and women for each eye is the same. For model training, 80% of the right eye and 80% of the left eye images were randomly selected. The remaining 20% were used for both validation and testing. The author recognizes that by reusing the same split for validation and testing that it may lead to an overestimation of the model's performance. Nonetheless, this approach was chosen due to the dataset's size and to maintain consistency with the first attempts done at training.

Each periocular image contains not only the iris and sclera, but also surrounding features such as eyelashes, eyelids, and periocular skin, as illustrated in Figure 2. These external anatomical structures often exhibit more sex specific characteristics than the iris texture itself [BR16] [KB19].

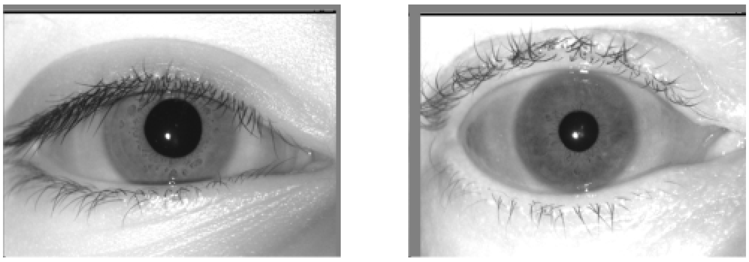


Fig. 2: Example of a woman's and man's periocular region, respectively

Prior to training, all images were processed using the `google/vit-base-patch16-224` ViT image processor. This includes standard preprocessing operations such as resizing to 224×224 pixels, normalization, and basic augmentation. The processor converts each image into pixel tensors suitable for input to the transformer architecture.

### 2.3 Dataset Variations

To evaluate how image orientation and dataset composition affect model performance, four variations of the dataset were created:

The first variation, which will be referred as *right_eyes_only*, includes 1500 images, all of them being right eyes. The second variation, *both_eyes* includes all 3000 images, comprising both right and left eyes. The third variation, *both_eyes_mirrored* augments the *right_eyes_only* variation by horizontally mirroring all left eye images to resemble right eyes, resulting in a balanced set of 1,500 original right eyes and 1,500 mirrored left eyes. Finally, the fourth variation, *both_eyes_together*, combines each subject's right and left eyes into a single image, as illustrated in Figure 3.

Variation 4, which includes images of size 280×480 pixels was preprocessed in the same way as the other variations, as described before (2.2), despite having a different width-to-height ratio.

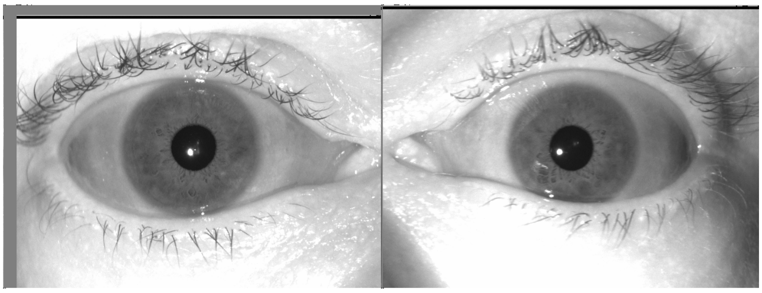


Fig. 3: Example of combined left and right eye image (*both_eyes_together* variation).

### 2.4 Hyper-parameters Selection

In order to find the best possible implementation of the model, it was necessary to explore different hyper-parameters and see how each may affect the outcome. A grid search was performed by trying different combinations for Number of Epochs, Learning Rate (LR), and Batch Size (BS). The Number of Epochs represents how many times the ViT will pass through the complete dataset during training. The Learning Rate controls how fast the model updates its parameters in response to the calculated error during each step and finally, the Batch Size determines how many samples are processed before the model's weights are updated.

In this work, the number of training epochs was evaluated across the values 2, 3, 5, 8, and 20. Batch sizes ranged from $n = 8$ to $n = 32$, increasing in powers of two. The learning rate was varied between $1.00 \times 10^{-5}$ and $4.00 \times 10^{-4}$, with each value obtained by doubling the previous one.

For each hyper-parameter combination, a unique identifier (e.g., Run_0) is assigned to facilitate direct comparison of results across different dataset variations using the same configuration.

### 2.5 Evaluation Strategy

To compare different configurations, various evaluation metrics were computed. However, for selecting the best-performing model, only Accuracy, Area Under the Curve (AUC), Equal Error Rate (EER), and F1 score were considered.

Accuracy measures the proportion of correct gender predictions made by the model:

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

AUC represents the area under the Receiver Operating Characteristic (ROC) curve; a higher AUC indicates a better trade-off between true positive rate and false positive rate, meaning there is a strong discrimination between true male and false male predictions. The Equal Error Rate (EER), typically expressed as a percentage, corresponds to the point on the Detection Error Tradeoff (DET) curve where the false acceptance rate equals the false rejection rate. In this work's context, a low EER suggests the model rarely misclassifies male images as female and vice versa.

Finally, the F1 score is a harmonic mean of precision and recall, calculated as:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Its main advantage lies in combining both precision and recall into a single metric, offering a balanced measure of the model's performance.

## 3 Experimental Evaluation

To import and train the ViT, PyTorch and HuggingFace's transformers library [4] were used. Matplotlib was employed for image visualization and for plotting evaluation metrics, while Scikit-learn was used to generate test splits and compute various performance metrics. For attention map visualization, publicly available code was adapted from an open-source notebook.[5]

Training 3 epochs with a Nvidia Tesla V100 took less than 15 minutes and inferring on one image around 330 ms.

### 3.1 Results

In Table 1, Table 2, Table 3, Table 4 the performance scores for each run across the different dataset variations can be seen. As shown in Table 1, the small size of the dataset leads the model to overfit after just 3 epochs, resulting in deteriorated performance. In contrast, for Table 2 and Table 3, the dataset is twice as large, and the effects of overfitting are less severe. Between these two variations (*both_eyes* and *both_eyes_mirrored*) there was a very small, yet still significant difference. This suggests that, despite the architectural design of ViTs being less dependent on specific feature orientations, spatial orientation can still influence model performance in subtle but significant ways.

In the case of Table 4, although the dataset contains only 1,500 samples, each sample includes twice as much information, which could explain the need for 5 epochs to reach optimal performance. Run_2.1 in this dataset variation achieved the highest overall accuracy of 95.3%, outperforming all other runs in every other dataset variation.

[4] `https://huggingface.co/docs/transformers/en/index`

[5] `https://github.com/jeonsworld/ViT-pytorch/blob/main/visualize_attention_map.ipynb`

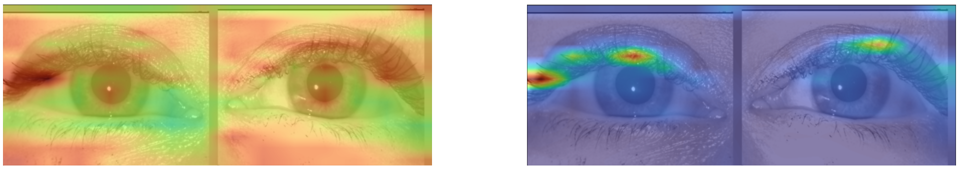
It was observed that for configurations that performed better, the attention map becomes more generalized over the image with some specific focus on key areas like the eye lashes, surrounding skin and pupil. With models that performed worse the attention was more localized and concentrated on the eye lashes mostly. A clear illustration of this phenomenon is shown in Figure 4. Notably, models trained using the *both_eyes_together* variation consistently generated broader and more informative attention distributions compared to those trained with the other dataset configurations.

All the graphs produced for Run_2.1 are shown in Figure 5. The confusion matrix is presented in Figure 5a and the ROC curve in Figure 5b. In Figure 5c it is displayed the estimated Gaussian distributions of the model's scores, obtained by running inference on the entire test set and classifying samples as male or female using a threshold of 0.5.

Finally, a DET curve was plotted for the best-performing run within each dataset variation, as presented in Figure 6. While there is a small increase in performance from *right_eyes_only* to *both_eyes*, the same cannot be said from *both_eyes* to *both_eyes_mirrored* with very similar scores being attained by both. However, the most significant improvement is observed with the introduction of the *both_eyes_together* setup.

| **Run** | **Epochs** | **LR** | **BS** | **Accuracy** | **AUC** | **EER** | **F1** |
|---|---|---|---|---|---|---|---|
| Run_0 | 20 | 2.00E-04 | 32 | 0.737 | 0.807 | 25.0% | 0.683 |
| Run_1 | 20 | 2.00E-04 | 16 | 0.773 | 0.910 | 17.5% | 0.815 |
| Run_2 | 5 | 2.00E-04 | 16 | 0.867 | 0.949 | 13.0% | 0.854 |
| Run_3 | 3 | 2.00E-04 | 16 | 0.890 | 0.959 | 11.0% | 0.883 |
| **Run_4** | **3** | **1.00E-04** | **16** | **0.917** | **0.972** | **8.0%** | **0.908** |
| Run_5 | 3 | 5.00E-05 | 16 | 0.903 | 0.966 | 9.7% | 0.898 |
| Run_6 | 3 | 8.00E-05 | 16 | 0.870 | 0.957 | 9.3% | 0.869 |
| Run_7 | 3 | 2.00E-04 | 8 | 0.813 | 0.910 | 17.7% | 0.793 |
| Run_8 | 3 | 5.00E-05 | 8 | 0.900 | 0.959 | 9.3% | 0.891 |
| Run_9 | 2 | 1.00E-04 | 16 | 0.870 | 0.951 | 12.3% | 0.865 |

Tab. 1: Performance metrics for the `right_eyes_only` variation.



(a) Run_2.1 attention map visualized.

(b) Run_2.2 attention map visualized.

Fig. 4: Comparison of attention maps from the best and the worst run in Table 4, where red is high importance.

| **Run** | **Epochs** | **LR** | **BS** | **Accuracy** | **AUC** | **EER** | **F1** |
|---|---|---|---|---|---|---|---|
| Run_0 | 20 | 2.00E-04 | 32 | 0.890 | 0.948 | 12.2% | 0.888 |
| Run_1 | 20 | 2.00E-04 | 16 | 0.877 | 0.961 | 10.0% | 0.861 |
| Run_2 | 5 | 2.00E-04 | 16 | 0.898 | 0.966 | 9.3% | 0.897 |
| Run_3 | 3 | 2.00E-04 | 16 | 0.903 | 0.960 | 10.0% | 0.901 |
| Run_4 | 3 | 2.00E-04 | 16 | 0.877 | 0.961 | 10.0% | 0.861 |
| **Run_5** | **3** | **5.00E-05** | **16** | **0.930** | **0.973** | **7.7%** | **0.929** |
| Run_5.1 | 3 | 1.00E-05 | 16 | 0.890 | 0.954 | 11.2% | 0.888 |
| Run_7 | 3 | 2.00E-04 | 8 | 0.907 | 0.959 | 9.0% | 0.902 |
| Run_7.1 | 3 | 1.00E-05 | 8 | 0.905 | 0.961 | 9.7% | 0.902 |
| Run_8 | 3 | 5.00E-05 | 8 | 0.918 | 0.969 | 9.0% | 0.918 |

Tab. 2: Performance metrics for the both_eyes variation.

| **Run** | **Epochs** | **LR** | **BS** | **Accuracy** | **AUC** | **EER** | **F1** |
|---|---|---|---|---|---|---|---|
| Run_0 | 20 | 2.00E-04 | 32 | 0.917 | 0.973 | 7.8% | 0.920 |
| Run_1 | 20 | 2.00E-04 | 16 | 0.877 | 0.944 | 12.8% | 0.881 |
| Run_2 | 5 | 2.00E-04 | 16 | 0.915 | 0.971 | 8.8% | 0.917 |
| Run_3 | 3 | 2.00E-04 | 16 | 0.918 | **0.980** | **7.5%** | 0.921 |
| **Run_4** | **3** | **1.00E-04** | **16** | **0.932** | 0.976 | 7.8% | **0.934** |
| Run_4.1 | 3 | 1.00E-04 | 16 | 0.923 | 0.966 | **7.5%** | 0.925 |
| Run_4.2 | 3 | 1.50E-04 | 24 | 0.925 | 0.975 | 7.8% | 0.927 |
| Run_5 | 3 | 5.00E-05 | 16 | 0.925 | 0.972 | 7.8% | 0.927 |
| Run_5.1 | 3 | 1.00E-05 | 16 | 0.910 | 0.963 | 9.2% | 0.910 |
| Run_7 | 3 | 2.00E-04 | 8 | 0.842 | 0.923 | 15.5% | 0.848 |
| Run_7.1 | 3 | 1.00E-05 | 8 | 0.900 | 0.960 | 10.2% | 0.910 |
| Run_8 | 3 | 5.00E-05 | 8 | 0.923 | 0.977 | **7.5%** | 0.923 |

Tab. 3: Performance metrics for the both_eyes_mirrored variation.

## 4 Conclusion

In this work, a novel approach for gender classification was explored using only Near Infrared periocular images and a Vision Transformer. Most publicly available periocular Near Infrared datasets were originally developed for identification rather than classification, making the lack of suitable data an ongoing challenge.

To address this, a Vision Transformer model pre-trained on ImageNet-21k was fine-tuned using a smaller, custom periocular dataset. Four dataset variations were designed to investigate the impact of input diversity and size on performance. Despite prior work suggesting that the iris contains gender information, the models trained in this paper primarily relied on surrounding periocular features, such as eyelashes and skin texture, rather than the iris itself. Notably, using both eyes from the same subject as input resulted in the highest classification accuracy of 95.3%.

| **Run** | **Epochs** | **LR** | **BS** | **Accuracy** | **AUC** | **EER** | **F1** |
|---|---|---|---|---|---|---|---|
| Run_0 | 20 | 2.00E-04 | 32 | 0.902 | 0.982 | 7.8% | 0.888 |
| Run_1 | 20 | 2.00E-04 | 16 | 0.895 | 0.954 | 9.1% | 0.885 |
| Run_1.1 | 8 | 2.00E-04 | 16 | 0.895 | 0.965 | 8.4% | 0.874 |
| Run_2 | 5 | 2.00E-04 | 16 | 0.949 | **0.990** | **4.7%** | 0.941 |
| **Run_2.1** | **5** | **2.00E-04** | **32** | **0.953** | 0.989 | 5.4% | **0.946** |
| Run_2.2 | 5 | 4.00E-04 | 16 | 0.666 | 0.761 | 33.5% | 0.648 |
| Run_3 | 3 | 2.00E-04 | 16 | 0.899 | 0.974 | 9.3% | 0.877 |
| Run_4 | 3 | 2.00E-04 | 16 | 0.929 | 0.985 | 6.8% | 0.919 |
| Run_5 | 3 | 5.00E-05 | 16 | 0.922 | 0.978 | 7.8% | 0.908 |
| Run_5.1 | 3 | 1.00E-05 | 16 | 0.861 | 0.932 | 13.2% | 0.846 |
| Run_7 | 3 | 2.00E-04 | 8 | 0.868 | 0.943 | 15.2% | 0.859 |
| Run_8 | 3 | 5.00E-05 | 8 | 0.939 | 0.985 | 7.5% | 0.931 |

Tab. 4: Performance metrics for the `both_eyes_together` variation.

These findings suggest that the periocular region, independent of the iris, holds significant discriminative features for gender classification. Not only that, but the results also suggest that incorporating both eyes from the same subject into training and inference, even in small datasets, can substantially improve performance compared to models trained and tested on single-eye inputs.

Future work could focus on expanding the dataset and exploring the use of other Vision Transformer models for periocular gender classification across broader demographic variations.

## References

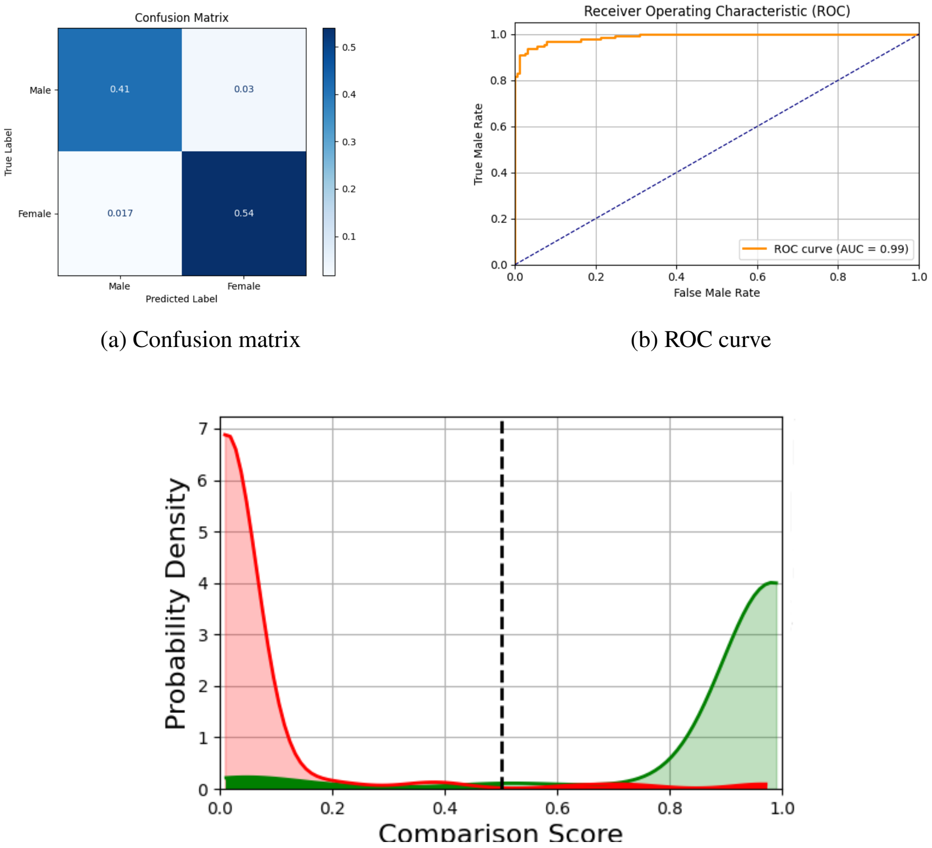
[BAS12] Bansal, Atul; Agarwal, Ravinder; Sharma, R.K.: SVM Based Gender Classification Using Iris Images. In: 2012 Fourth International Conference on Computational Intelligence and Communication Networks. S. 425–429, 2012.

[BR16] Bobeldyk, Denton; Ross, Arun: Iris or Periocular? Exploring Sex Prediction from Near Infrared Ocular Images. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). S. 1–7, 2016.

[Da93] Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(11):1148–1161, 1993.

[Do21] Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob; Houlsby, Neil: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR). 2021.

[KB19] Kuehlkamp, Andrey; Bowyer, Kevin: Predicting Gender From Iris Texture May Be Harder Than It Seems. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). S. 904–912, 2019.



(a) Confusion matrix

(b) ROC curve

(c) Score distribution for male (green) and female (red) images.

Fig. 5: Performance results of Run_2.1 from Table 4.

[Ra24] Rangwani, Harsh; Mondal, Pradipto; Mondal, Pradipto; Mishra, Mayank; Asokan, Ashish Ramayee; Babu, R. Venkatesh: DeiT-LT: Distillation Strikes Back for Vision Transformer Training on Long-Tailed Datasets. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 23396–23406, 2024.

[SP23] Suravarapu, Vasu Krishna; Patil, Hemprasad Yashwant: Person Identification and Gender Classification Based on Vision Transformers for Periocular Images. Applied Sciences, 13(5), 2023.

[TA18] Tapia, Juan; Aravena, C. Carlos: Gender classification from periocular NIR images using fusion of CNNs models. In: 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA). S. 1–6, 2018.

[TFPB16] Tapia Farias, Juan; Perez, Claudio; Bowyer, Kevin: Gender Classification From the Same Iris Code Used for Recognition. IEEE Transactions on Information Forensics and Security, 11:1–1, 08 2016.

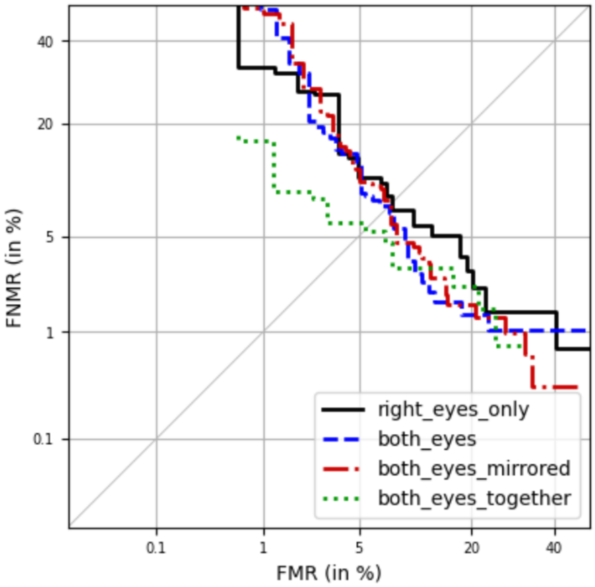


Fig. 6: DET curve of the best configurations by dataset variation.

[Vi19] Viedma, Ignacio; Tapia Farias, Juan; Iturriaga Jofré, Andrés; Busch, Christoph: Relevant features for Gender Classification in NIR Periocular Images. IET Biometrics, 8, 05 2019.