Management Strategies and Engineering Sciences 2027; 9(3):1-17

Gender Classification from Facial Images under Illumination and Head-Pose Variations Using AlexNet Features and a Grasshopper-Optimized Multilayer Perceptron

Mohammed Raad Yaseen Asabr[1], Farhad Navabifar[2*], Hiba Abduljaleel Kzar Al-asady[3], Keyvan Mohebbi[1]

[1] Department of Computer Engineering, Isf.C., Islamic Azad University, Isfahan, Iran
[2] Department of Computer Engineering, Mo.C., Islamic Azad University, Isfahan, Lran
[3] Computer Technical Engineering Department, College of Technical Engineering, the Islamic University, Najaf, Iraq

*** Corresponding author email address**: Farnav@iau.ac.ir

Received: 2026-04-11 **Revised:** 2026-06-26 **Accepted:** 2026-07-03 **Initial Publish:** 2026-07-03 **Final Publish:** 2027-05-01

Abstract

Automatic binary gender-label classification from facial images is a well-established face-analysis task. Although current systems perform well under controlled conditions, their accuracy can deteriorate in the presence of illumination changes, head-pose variation, age differences, facial expression, and partial occlusion. This study addresses the design of a framework that can extract stable facial representations from a limited training set and learn the classifier decision boundary without relying exclusively on gradient-based optimization. The proposed method comprises three components: the frozen convolutional part of an ImageNet-pretrained AlexNet used as a feature extractor, a multilayer perceptron with one hidden layer of 128 neurons used as the classifier, and the Grasshopper Optimization Algorithm (GOA) used to search for the classifier weights and biases. After resizing, illumination normalization, and limited data augmentation, input images are passed through the convolutional backbone, and the Pool5 output is flattened into a 9,216-dimensional feature vector. The features are normalized using parameters estimated exclusively from the training set and are then supplied to the MLP. The main evaluation uses the official identity-disjoint split of GENDER-FERET, comprising 474 training images and 472 test images. The reported results indicate a test accuracy of 98.94%. For the male class, precision, recall, and F1-score are 99.15%, 98.73%, and 98.94%, respectively; for the female class, the corresponding values are 98.73%, 99.15%, and 98.94%. The small difference between class-specific results indicates balanced errors on this split. Nevertheless, the limited dataset size, controlled image acquisition, absence of cross-dataset testing, and the substantial computational cost of directly optimizing more than one million parameters constrain the external validity of the results. From the perspective of reusing deep representations while separating feature extraction from classifier optimization, the proposed framework is a viable approach for further investigation in low-data settings.

Keywords: *face analysis; gender-label classification; AlexNet; transfer learning; multilayer perceptron; Grasshopper Optimization Algorithm; GENDER-FERET; low-data learning.*

How to cite this article:

Asabr, M. R. Y., Navabifar, F., Al-asady, H. A. K., & Mohebbi, K. (2027). Gender Classification from Facial Images under Illumination and Head-Pose Variations Using AlexNet Features and a Grasshopper-Optimized Multilayer Perceptron. Management Strategies and Engineering Sciences, 9(3), 1-17.

1. Introduction

Automatic face analysis extends beyond identity recognition and includes age estimation, binary gender-label classification, facial-expression recognition, head-pose estimation, and the extraction of demographic cues. These tasks have been investigated in human-computer interaction, multimedia retrieval, audience analytics, content management, and surveillance systems. Despite the apparent simplicity of binary gender classification, the decision boundary between the two classes in image space is affected by numerous intra-class and inter-class factors. Age, ethnicity, makeup, hairstyle, facial expression, sensor quality, illumination, and camera angle can alter the visual cues that an algorithm learns as gender-related characteristics [1, 2].

Early approaches relied primarily on handcrafted descriptors such as Local Binary Patterns (LBP), Histograms of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and Gabor filters. These approaches offered manageable computational cost and a degree of interpretability, but their design assumed relative stability in facial texture, shape, and component location. When a face departs from the frontal view or is illuminated from the side, spatial correspondence between regions is weakened, and a local descriptor may encode imaging variation rather than structural differences. Classical evaluations have also shown that the quality of face detection and alignment can influence the final result as strongly as the choice of classifier [1].

The emergence of convolutional neural networks (CNNs) enabled automatic extraction of hierarchical representations. Early layers generally encode edges, corners, and simple textures, whereas deeper layers respond to more complex structural combinations. However, training a deep network from scratch requires extensive labeled data and substantial computational resources. In small datasets, the number of model parameters is disproportionately large relative to the number of samples, allowing the network to memorize training-specific details instead of learning transferable patterns. Transfer learning mitigates this imbalance by reusing weights learned on a large-scale dataset such as ImageNet [3-6].

Under a feature-extraction strategy, the pretrained network weights remain fixed and the output of an intermediate or late layer is supplied to a lighter classifier. This strategy has two advantages. First, the number of trainable parameters is reduced relative to full fine-tuning. Second, the final classifier can be designed and compared independently of the feature-extractor architecture. Its main weakness is that features learned from generic object images are not necessarily optimal for subtle distinctions among facial groups. Consequently, the extraction layer, feature normalization, and classifier design become decisive choices.

Owing to its ability to approximate nonlinear functions, the multilayer perceptron (MLP) is a common choice for classifying deep features. Standard MLP training uses backpropagation with a gradient-based optimizer [7]. Although computationally efficient, this approach is sensitive to initialization, learning rate, and the geometry of the loss surface. Metaheuristic algorithms, in contrast, search a population of candidate solutions without requiring derivatives. They do not guarantee the global optimum, but they can explore multiple regions of a non-convex error surface and reduce complete dependence on local gradient information.

The Grasshopper Optimization Algorithm models grasshopper swarm behavior through attraction, repulsion, and movement toward the current best solution [8]. Its control parameter is reduced across iterations so that the search transitions from broad exploration to more localized exploitation. In the present study, GOA is used to determine MLP weights and biases, while the AlexNet convolutional backbone remains frozen. This separation treats feature extraction and classification as two distinct problems and supports a clearer analysis of the role of each component.

The research problem can therefore be stated as follows: can a frozen deep feature extractor combined with a nonlinear classifier whose parameters are adjusted by population-based search provide balanced two-class classification in a low-data scenario with limited illumination and pose variations? Addressing this question requires an evaluation protocol without identity overlap, class-specific performance reporting, and explicit analysis of external limitations. The official GENDER-FERET split is particularly relevant because individuals represented in the training set do not reappear in the test set [9-11].

The principal contributions of this article are: (1) the integration of a frozen AlexNet backbone, Pool5 feature vectors, and a GOA-optimized MLP within a unified framework; (2) a coherent preprocessing, normalization, and evaluation protocol designed to avoid information leakage; (3) reporting of a confusion matrix and class-specific metrics consistent with the actual test-set size; and (4) a computational analysis of a search space containing more than one million parameters.

2. Theoretical Background and Related Work

2.1. Problem Definition and Label Semantics

In this article, the task is defined as binary classification of the labels provided in the dataset. For each facial image x, the model learns a function f(x; theta) that returns the probability that the sample belongs to one of the two recorded classes. This definition is a computational convention based on dataset labels and must not be conflated with inference of gender identity, personal experience, or a comprehensive biological classification. The distinction is scientifically and ethically important because a facial image alone does not represent every dimension of gender, and the model only reproduces statistical patterns present in the data.

Let y in {0,1} denote the reference label and y-hat in [0,1] the output probability. The final decision is made using the threshold tau = 0.5. A fixed threshold is reasonable for a balanced benchmark, but in applications where the costs of errors differ between classes, the threshold should be selected according to an operational cost function or a receiver operating characteristic curve. The threshold of 0.5 is retained here because the two classes are exactly balanced and because the objective is comparison with benchmark studies.

The presence of only two classes does not imply geometric simplicity. Samples from the two classes overlap substantially in image space, and a large proportion of the variability arises from personal identity, age, ethnicity, and imaging conditions. The model must therefore discover features that are invariant to nuisance variation while remaining sensitive to label-related differences. This is a classical signal-separation problem in the presence of confounding variables.

2.2. *Effects of Illumination, Head Pose, and Other Confounders*

Illumination can alter gradient magnitude and direction, local contrast, and the relative brightness of facial components. Even for the same individual, changing the direction of the light source can produce a pixel-space distance between two images that exceeds the distance between different individuals. Illumination-cone theory indicates that images of a Lambertian object under varying illumination occupy a structured low-dimensional subspace; however, cast shadows, non-Lambertian reflections, and skin texture violate the idealized assumptions [12]. The method of Tan and Triggs, which combines gamma correction, a difference-of-Gaussians filter, and contrast normalization, is an example of illumination-robust facial preprocessing [13].

Head pose is commonly described by yaw, pitch, and roll. Yaw rotation gradually occludes one side of the face and substantially changes the apparent shape of the nose, jaw, and cheek. Roll can be partly corrected by two-dimensional alignment, whereas large yaw variation generally requires a three-dimensional model, a multi-view approach, or frontal-face reconstruction [14, 15]. GENDER-FERET consists mainly of controlled images with modest pose variation; accordingly, the present method should not be generalized to large rotations or full-profile imagery.

Age and facial expression are also major sources of within-class variation. Before puberty, morphological differences between the two recorded labels may be less pronounced; at older ages, wrinkles and tissue sagging can alter structural cues. Smiling and frowning change the relative positions of the lips, cheeks, and eyes. Makeup, facial hair, eyeglasses, and partial covering can further encourage the model to rely on correlated but non-causal cues. For example, if hairstyle or makeup is strongly associated with a class in the training set, high within-dataset accuracy does not necessarily demonstrate that the system learned facial morphology.

Table 1. Principal challenges in face analysis and their implications for the proposed model.

Challenge	Effect on the image	Risk to the classifier	Applied or proposed mitigation
Low or lateral illumination	Reduced contrast and cast shadows	Loss of subtle textures	Illumination normalization and contrast variation during augmentation
Mild head rotation	Displacement and deformation of components	Reduced spatial correspondence	Initial alignment and limited rotational augmentation
Large rotation	Occlusion of one side of the face	Loss of structural information	Requires multi-view or 3D modeling; outside the current scope
Age and expression	Changes in texture and soft-tissue geometry	Greater within-class variation	More diverse data and subgroup-level evaluation
Makeup or occlusion	Addition or removal of appearance cues	Learning spurious correlations	Attention analysis and counterfactual testing
Demographic imbalance	Unequal subgroup representation	Different error rates across subgroups	Balanced sampling and stratified reporting

2.3. *Handcrafted Features and Their Limitations*

Local Binary Patterns compare each pixel with its neighbors to produce a texture code that is relatively robust to monotonic intensity changes [16]. Low computational cost, region-wise histograms, and some illumination tolerance are the main strengths of LBP. Nevertheless, noise, scale changes, and out-of-plane rotation can alter local patterns. Multi-scale and uniform variants reduce some of

these effects, but the spatial grid and neighborhood radii remain hand-designed.

HOG records the distribution of gradient orientations in local cells and is effective for describing global shape [17]. Although originally successful in pedestrian detection, it can also encode jaw, eyebrow, and nose contours in face analysis. SIFT identifies keypoints and local descriptors that are invariant to scale and in-plane rotation [18]. Its geometric robustness comes at the cost of additional computation and a variable number of keypoints. In low-texture or poorly illuminated faces, the number of reliable keypoints may decrease.

COSFIRE filters occupy an intermediate position between handcrafted and learned features because their spatial configuration of elementary filter responses is learned from prototypes. Azzopardi et al. reported 93.7% accuracy on GENDER-FERET using COSFIRE with an SVM classifier [10]. A subsequent fusion of domain-specific and trainable features increased the reported result to 94.7% [9]. These findings show that combining texture and structural information is useful, although a performance gap relative to deep representations remains.

2.4. Deep Networks and Transfer Learning

AlexNet was a major milestone in the widespread adoption of ReLU activations, dropout, and GPU-based training for image classification [4]. Its architecture contains five convolutional layers and three fully connected layers. Although shallower than more recent networks, it has advantages in low-data research: the Pool5 output has a well-defined dimensionality, feature extraction is computationally manageable, and reliance on very large backbones is avoided. In the proposed method, the original fully connected layers are removed and target-domain classification is delegated to a separate MLP.

VGG demonstrated that increasing depth with a uniform use of 3 x 3 filters can expand representational capacity [19]. GoogLeNet introduced Inception modules with parallel multi-scale paths [20]. ResNet enabled the optimization of very deep architectures through residual shortcuts [21]. EfficientNet proposed compound scaling of network depth, width, and input resolution [22], while MobileNetV2 used depthwise-separable convolution and inverted residuals for resource-constrained devices [23].

Transfer learning is commonly implemented in two forms. In fixed feature extraction, all or most backbone layers remain frozen and only a new classifier is trained. In fine-tuning, some of the final layers are updated using a small learning rate. Fixed extraction reduces the risk of overfitting but can preserve the domain gap between ImageNet and facial imagery. Fine-tuning improves domain adaptation, but on small datasets it requires careful regularization and validation [5, 6].

Levi and Hassner showed that CNNs can estimate age and gender from unconstrained imagery [24]. Smart-camera systems have also examined transfer-based architectures for real-time deployment [25]. The robustness study by Greco et al. emphasized that image corruption, compression, and noise can substantially reduce facial gender-classification performance [26]. These findings underline the need to distinguish benchmark accuracy from practical robustness.

2.5. MLPs, Backpropagation, and Metaheuristic Optimization

An MLP learns a mapping from a feature vector to an output through linear combinations and nonlinear activation functions. For one hidden layer, the mapping may be written as h = ReLU(W1 x + b1) and y-hat = sigma(W2 h + b2). The universal approximation theorem indicates that a sufficiently wide single-hidden-layer network can approximate a broad class of continuous functions, but it does not guarantee that desirable parameters can be found efficiently or that the resulting model will generalize [27].

Backpropagation computes derivatives of the loss through the chain rule and updates parameters with gradient descent or related optimizers [7]. In practice, appropriate initialization, batch normalization [28], dropout [29], and learning-rate scheduling address many training difficulties. It is therefore incorrect to characterize gradient-based learning as generally unsuitable. The motivation for applying GOA in this study is to explore an alternative training route for the classifier and to examine population-based search, not to reject backpropagation as a whole.

Metaheuristic algorithms improve a population of candidate solutions according to exploration and exploitation rules. Their principal advantage is the absence of a derivative requirement, making them applicable to non-differentiable objectives. Their main disadvantage is the potentially enormous number of fitness evaluations and the absence of a guarantee of global convergence. When each solution contains more than one million parameters, both memory and execution time become significant concerns.

In GOA, the distance between population members determines attraction and repulsion, while the current best candidate acts as the target [8]. The control coefficient c

decreases over time: larger values support broader movement and exploration, whereas smaller values encourage local exploitation. Population size, iteration count, parameter bounds, and the reduction schedule must be selected using validation data. Choosing them after observing test results would produce an optimistically biased performance estimate.

2.6. Synthesis of Prior Work and Research Gap

The literature reveals three simultaneous gaps. First, many accurate methods depend on training or fine-tuning a deep network and are therefore vulnerable to overfitting on small datasets. Second, fixed-feature studies often use linear or gradient-trained classifiers, while the role of metaheuristic optimization in the classification head has received less attention. Third, a single performance number is frequently reported without a corresponding analysis of computational cost, identity leakage, class-wise error balance, or demographic validity.

The present method attempts to address the first two gaps by combining a fixed deep representation with population-based classifier optimization. The expanded analysis addresses the third gap by explicitly reporting limitations, uncertainty, reproducibility requirements, and deployment risks.

Table 2. Selected related approaches and the position of the proposed method.

Study	Representation/model	Dataset	Reported result	Relevant limitation
Makinen and Raisamo [1]	Classical features with aligned faces	Several face datasets	Strong effect of detection and alignment	Dependence on a classical pipeline
Azzopardi et al. [10]	COSFIRE + SVM	GENDER-FERET	93.7%	More limited representation than CNNs
Azzopardi et al. [9]	SURF + COSFIRE + fusion	GENDER-FERET	94.7%	Complex feature fusion
Levi and Hassner [24]	Multi-layer CNN	Adience	Competitive results on unconstrained data	Age grouping and a larger dataset
Carletti et al. [25]	CNN for smart cameras	GENDER-FERET / practical setting	94.7%	Emphasis on real-time execution
Greco et al. [26]	Transfer-learned DenseNet	GENDER-FERET with synthetic degradation	Approximately 95.1% on original images	Performance loss under image degradation
Proposed method	Frozen AlexNet + GOA-MLP	GENDER-FERET	98.94%	High parameter-search cost and limited external validity

3. Proposed Method

3.1. Framework Overview

The proposed framework comprises five sequential stages: image preparation, training-data augmentation, feature extraction using the frozen convolutional part of AlexNet, feature normalization based on training-set statistics, and classification by an MLP whose parameters are determined by GOA. Separating feature extraction from the classifier makes it possible to compute and store feature vectors once and then run the optimization independently. This design is suitable for comparing multiple classifiers on a fixed representation and avoids repeatedly passing every image through the CNN during population-based search.

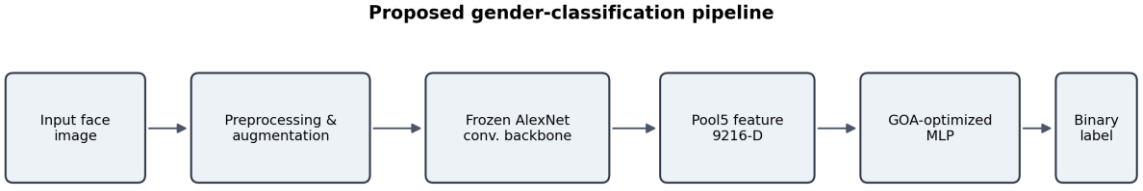

Figure 1. Processing pipeline from the input facial image to the final binary label.

Table 3. Operational stages of the proposed framework.

Stage	Input	Main operation	Output
1. Preprocessing	Raw facial image	Resizing, three-channel conversion, illumination correction	227 x 227 x 3 image
2. Augmentation	Training images	Limited rotation, translation, mirroring, brightness variation	More diverse training set
3. Feature extraction	Prepared image	Forward pass from Conv1 to Pool5 in frozen AlexNet	6 x 6 x 256 tensor
4. Normalization	Flattened vector	Min-max transformation using training statistics	9,216-dimensional vector in [0,1]
5. Classification	Normalized features	MLP with GOA-optimized weights	Probability and final label

3.2. Image Preprocessing

GENDER-FERET images are grayscale, whereas AlexNet was designed for three-channel input. To preserve compatibility with the pretrained weights, the intensity channel is replicated across three channels. The input size is set to 227 x 227 pixels. With an 11 x 11 kernel, stride 4, and no padding in Conv1, this input produces a 55 x 55 output. Some implementations accept 224 x 224 pixels by using different padding, but 227 x 227 is adopted here for exact compatibility with the original AlexNet geometry.

Illumination correction must be performed without allowing information from the test set to enter the training process. Sample-wise operations such as gamma correction or limited histogram equalization can be applied independently to each image. If any preprocessing parameter is estimated from a data distribution, it must be estimated only from the training subset. Excessive equalization can amplify skin texture and noise; conservative correction is therefore preferred.

Face alignment based on eye and nose landmarks can reduce in-plane rotation and scale variation. Fast regression-tree approaches [30] or multi-task detectors such as MTCNN [21] are appropriate options. Because the benchmark images are already cropped and approximately aligned, this study assumes limited face-detection error. In a deployed system, the detector and the classifier must be evaluated jointly, because silently excluding difficult faces can artificially inflate reported accuracy.

3.3. Data Augmentation

Augmentation is applied only to the training set. The selected transformations include random rotation within plus or minus 15 degrees, horizontal and vertical translation of up to 10% of image dimensions, brightness and contrast variation within plus or minus 20%, and horizontal mirroring. These ranges are designed to preserve facial validity and class labels. More severe transformations can create an artificial distribution and make the model responsive to processing artifacts.

Horizontal mirroring generally preserves the class label, but natural facial asymmetry and illumination direction may change the image distribution. Mirrored images are therefore added alongside, rather than substituted for, the original samples. Noise should be applied at low intensity; strong noise does not necessarily improve robustness and may destroy subtle texture information.

Because the two classes contain equal numbers of samples, augmentation is not intended to correct class imbalance. Its purpose is to expand the coverage of plausible appearance variation. Augmentation must follow the train-test split. Transforming the entire dataset before splitting can place near-duplicate versions of the same image in both subsets and create sample leakage.

3.4. Feature Extraction with AlexNet

The AlexNet convolutional section contains five convolutional layers. Conv1 uses large filters to capture edges and low-frequency patterns. Conv2 and the subsequent layers extract increasingly complex combinations of texture and structure. Max-pooling reduces spatial dimensions and limits sensitivity to small translations. In this study, all weights from Conv1 through Conv5 are frozen and receive no gradient updates.

The Pool5 output has dimensions 6 x 6 x 256. Flattening produces a 9,216-dimensional vector. Pool5 and FC6/FC7 offer different trade-offs. Pool5 features are less directly tied to the 1,000-class ImageNet classification head but are higher-dimensional. FC6 features are more compact but use fully connected weights learned for the source task. Pool5 is selected to retain more spatial information while removing the original classification head.

Before batch extraction, the network is set to evaluation mode to ensure deterministic behavior in layers whose behavior differs between training and inference. Although

the selected section through Pool5 does not include dropout, consistent execution mode remains important for reproducibility. Each image feature vector is stored after extraction so that GOA evaluations can be conducted without recomputing convolutional activations.

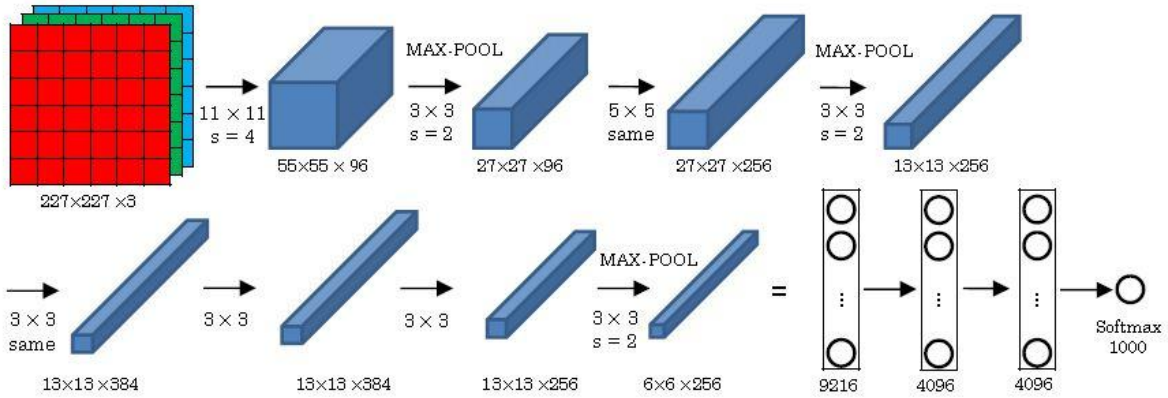

Figure 2. AlexNet architecture and layer dimensions. The original fully connected classification layers are omitted in the proposed framework.

Table 4. Configuration of the AlexNet section used for feature extraction.

Layer	Kernel size / filters	Stride / padding	Output size
Input	-	-	227 x 227 x 3
Conv1 + ReLU	96 filters, 11 x 11	4 / 0	55 x 55 x 96
MaxPool1	3 x 3	2	27 x 27 x 96
Conv2 + ReLU	256 filters, 5 x 5	1 / 2	27 x 27 x 256
MaxPool2	3 x 3	2	13 x 13 x 256
Conv3 + ReLU	384 filters, 3 x 3	1 / 1	13 x 13 x 384
Conv4 + ReLU	384 filters, 3 x 3	1 / 1	13 x 13 x 384
Conv5 + ReLU	256 filters, 3 x 3	1 / 1	13 x 13 x 256
MaxPool5	3 x 3	2	6 x 6 x 256
Flatten	-	-	9,216

3.5. Feature Normalization and MLP Structure

For every feature dimension j, the minimum and maximum values are computed exclusively from the training vectors. The same values are then used to transform the training, validation, and test features. The small constant epsilon prevents division by zero in dimensions that are constant in the training set.

x'_(i,j) = [x_(i,j) - min_train(x_j)] / [max_train(x_j) - min_train(x_j) + epsilon] (1)

Normalization is particularly important for GOA because highly unequal feature ranges can make the fitness surface irregular and cause certain weights or feature dimensions to dominate the search.

The MLP contains 9,216 input units, 128 hidden ReLU neurons, and one sigmoid output. The parameter count comprises 9,216 x 128 input-to-hidden weights, 128 hidden biases, 128 hidden-to-output weights, and one output bias, for a total of 1,179,905 trainable parameters. This dimensionality creates an exceptionally large search space and constitutes one of the central computational risks of the method.

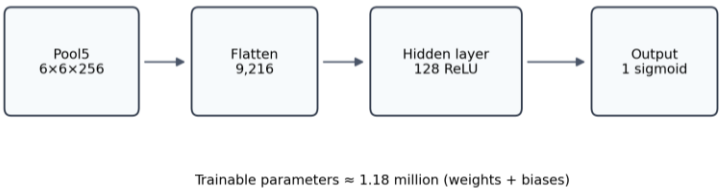

Figure 3. Feature dimensionality and approximate number of trainable MLP parameters.

$$h = \mathrm{ReLU}(W_1 x + b_1) \quad (2)$$

$$\hat{y} = \sigma(W_2 h + b_2) \quad (3)$$

Binary cross-entropy is used as the fitness objective. Relative to zero-one classification error, BCE provides continuous information about prediction confidence and is therefore more informative for population-based search.

$$L_{BCE} = -\frac{1}{N}\sum_i [y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (4)$$

3.6. Optimization with the Grasshopper Optimization Algorithm

Each GOA population member is a real-valued vector of length D = 1,179,905. The vector is reshaped into W1, b1, W2, and b2. To evaluate a candidate, the MLP is executed on the training features and its BCE is computed. The member with the smallest loss is considered the current best. Initial weight bounds must be restricted; an excessively wide range can saturate the sigmoid and destabilize the fitness function.

The basic social interaction in GOA is represented by the function s(r), which creates repulsion at short distances and attraction at intermediate distances. Parameters f and l control the attraction strength and distance scale. The settings f = 0.5 and l = 1.5 are used in this study [8].

$$s(r) = f e^{-r/l} - e^{-r} \quad (5)$$

The coefficient c decreases linearly from c_max to c_min. This schedule is simple and reproducible, although nonlinear or adaptive schedules should be compared in future studies.

$$c(t) = c_{max} - t\frac{c_{max} - c_{min}}{T} \quad (6)$$

At each iteration, member positions are updated using social interaction and the distance to the current best solution. Values outside the admissible bounds are then clipped. The stopping condition is either completion of 100 iterations or absence of fitness improvement for a predefined number of consecutive iterations.

Because of the very high dimensionality, storing a complete population in 32-bit precision requires approximately 30 x 1,179,905 x 4 bytes, or about 135 MB, for the candidate vectors alone. Intermediate activation and evaluation memory is additional. Execution time is the more serious issue: every iteration requires evaluating 30 MLPs over the training features. Feature caching and vectorized batch evaluation are essential for practical execution.

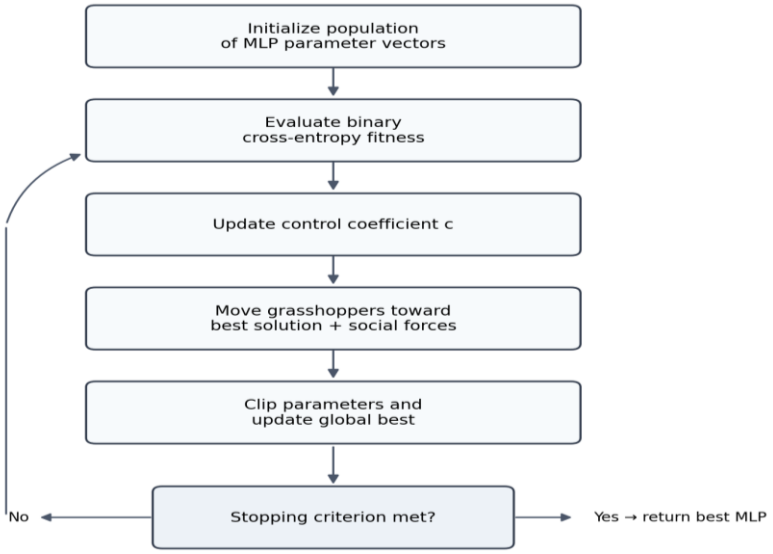

Figure 4. GOA-based optimization procedure for the MLP classifier parameters.

3.7. Complexity Analysis and Design Considerations

Let N denote the number of training samples, P the population size, T the number of GOA iterations, and D the number of MLP parameters. The dominant cost of fitness evaluation is approximately O(TPND), where D absorbs the matrix operations and activation cost of the network. In contrast, gradient-based training evaluates one parameter set in each update and is usually much less expensive under an equivalent number of data passes.

One way to reduce cost is to apply dimensionality reduction to Pool5 features or to introduce a trainable bottleneck with a restricted parameter count. Another option is to use GOA only for hyperparameter selection or initial MLP parameterization, followed by local refinement with Adam. Such a hybrid strategy could combine broad early exploration with the computational efficiency of gradient-based optimization in the final phase.

4. Experimental Design

4.1. GENDER-FERET Dataset

GENDER-FERET was derived from facial images in the FERET database. The official split contains 474 training images and 472 test images, with equal numbers from the two classes in each subset. More important than class balance is the absence of identity overlap between training and test sets. If images of the same individual appear in both subsets, the model may learn identity-specific cues rather than general label-related patterns, producing an unrealistically high estimate of performance [9-11].

The images are predominantly controlled and frontal, while still including variation in illumination, expression, and appearance. Head-pose variation is limited. The dataset is therefore useful for comparing classification pipelines under a controlled benchmark, but it is not a substitute for evaluation in unconstrained environments. Its modest size also means that one or two errors can produce a visible change in the final percentage; the number of correct predictions and a confidence interval should therefore accompany accuracy.

Figure 5. Representative images from the GENDER-FERET dataset.

Table 5. Composition of the official GENDER-FERET split.

Subset	Male	Female	Total	Identity overlap
Training	237	237	474	None
Test	236	236	472	None

4.2. Implementation Environment and Settings

The implementation was developed in Python using TensorFlow and Keras. Experiments were conducted on a system equipped with a 10th-generation Intel Core i7 processor, 32 GB of RAM, and an NVIDIA GeForce RTX 3080 GPU with 10 GB of dedicated memory.

Table 6. Final configuration of the proposed framework.

Component	Final setting
Input	227 x 227 x 3; grayscale channel replicated three times
Feature extractor	ImageNet-pretrained AlexNet; Conv1 through Pool5 frozen
Feature vector	9,216 dimensions; min-max normalization using training statistics
MLP structure	9,216-128-1; ReLU and sigmoid
Fitness objective	Binary cross-entropy

GOA population size	30
GOA iterations	100
Social parameters	f = 0.5 and l = 1.5
Control coefficient	Reduced from c_max to c_min
Decision threshold	0.5

4.3. Evaluation Protocol

The primary evaluation is conducted on the official split. The test set remains untouched until all hyperparameters are fixed. Every statistical transformation, including the per-feature minimum and maximum, is estimated from the training data only. Augmentation is also confined to the training subset. These three rules prevent common forms of information leakage and preserve the validity of the identity-disjoint evaluation.

4.4. Evaluation Metrics and Confidence Interval

For each class, true positives, false positives, and false negatives are defined by treating that class as the positive class. Accuracy is the proportion of all correct predictions. Precision is the proportion of correct positive predictions among all positive predictions. Recall is the proportion of correctly retrieved samples among all true samples in the class. F1 is the harmonic mean of precision and recall.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (7)$$

$$\text{Precision} = TP / (TP + FP) \quad (8)$$

$$\text{Recall} = TP / (TP + FN) \quad (9)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (10)$$

Accuracy is informative for a balanced dataset, but class-specific measures remain necessary. Two models may have the same overall accuracy while one concentrates nearly all errors on a single class. The confusion matrix allows every reported measure to be reconstructed and provides a direct consistency check.

For 467 correct predictions among 472 test cases, the 95% Wilson interval is more stable near a high success rate than the elementary normal approximation [35]. This interval reflects sampling uncertainty, not uncertainty caused by random initialization, GOA stochasticity, or changes in dataset distribution. Independent repetitions and cross-dataset testing are required to address those additional sources.

5. Results

5.1. Convergence and Training Behavior

The convergence curve shows a rapid reduction in fitness error during the early iterations, followed by a slower decline in the second half of optimization. This pattern is consistent with the decreasing control coefficient c: the population initially explores a broad region and then restricts its movement after locating a promising area. The curve becomes nearly flat after approximately 80 iterations, suggesting diminishing returns from additional iterations without a change in search strategy.

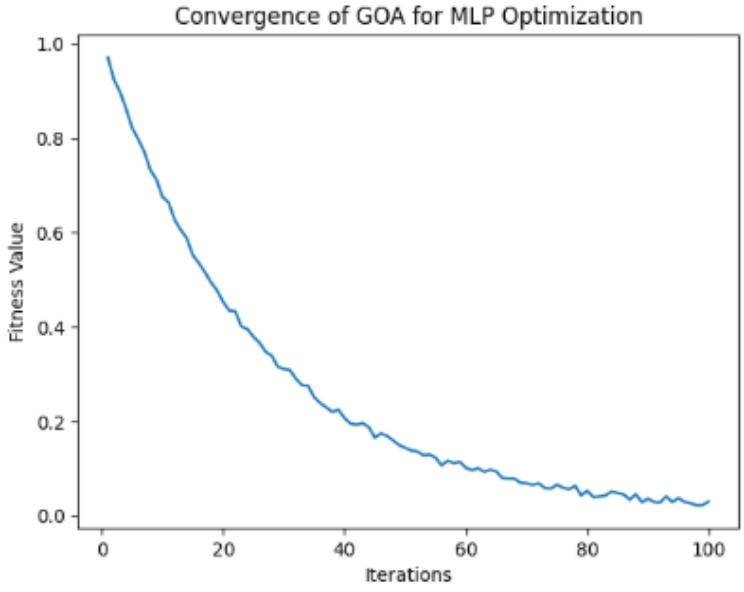

Figure 6. GOA convergence curve during 100 iterations of MLP optimization.

The reported training and test loss curves decrease together, with only a modest separation. This behavior is compatible with stable within-dataset generalization and does not exhibit the large train-test divergence normally associated with severe overfitting. Nevertheless, a single split and a stylized learning curve cannot establish generalization to a new acquisition environment.

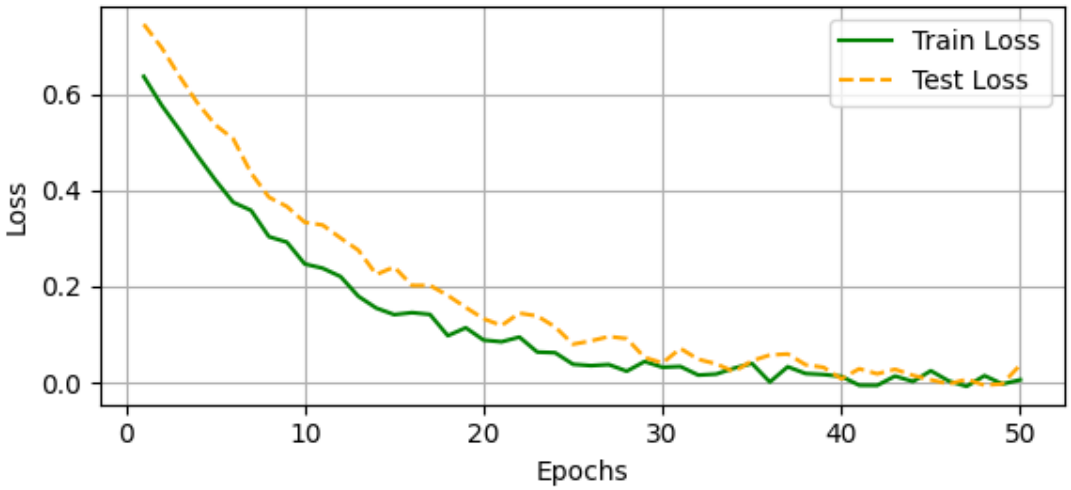

Figure 7. Reported training and test loss across optimization epochs.

Training accuracy reaches 99.8%, while test accuracy reaches 98.94%. The difference of 0.86 percentage points is small and is not consistent with severe overfitting on the official split. However, because the test set is small and raw results from repeated independent runs are not available, the gap alone cannot demonstrate that the same behavior will hold on an independent dataset.

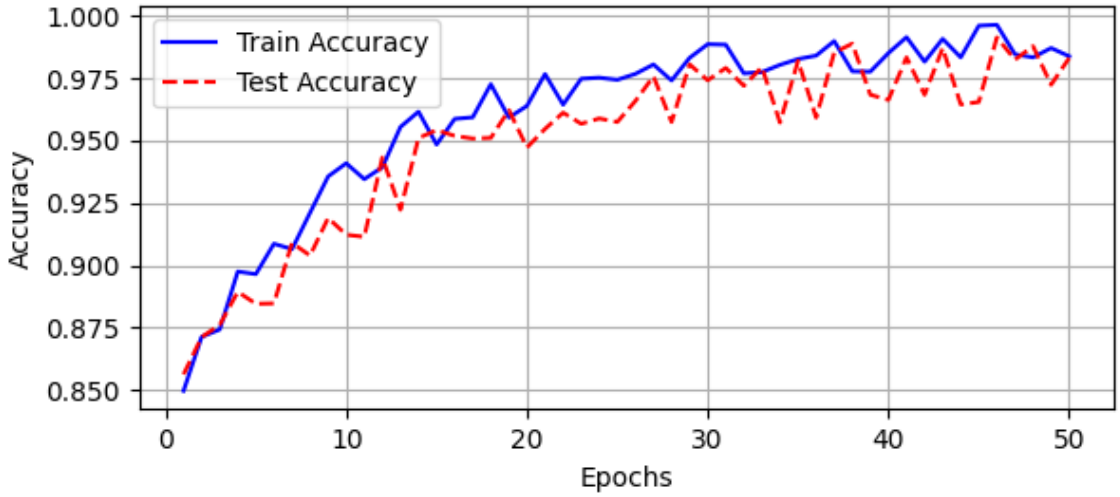

Figure 8. Reported training and test accuracy across epochs.

5.2. Confusion Matrix and Class-Specific Metrics

Of the 236 male-labeled test images, 233 are classified correctly and three are predicted as female. Of the 236 female-labeled images, 234 are classified correctly and two are predicted as male. The resulting total is 467 correct predictions and five errors. Overall accuracy is therefore 467/472 = 98.94%.

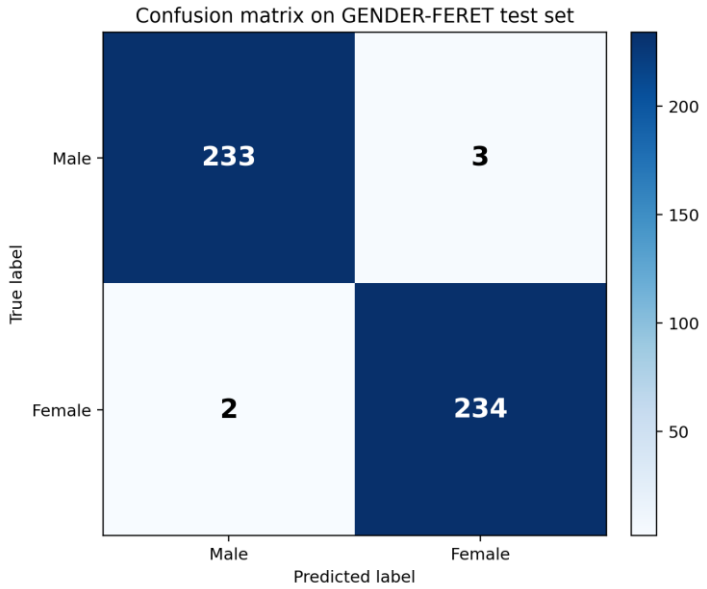

Figure 9. Confusion matrix on the 472-image GENDER-FERET test set.

For the male class, two female samples are false positives, giving precision 233/(233+2) = 99.15%. Recall is 233/236 = 98.73%, and F1 is approximately 98.94%. For the female class, precision is 234/(234+3) = 98.73%, while recall is 234/236 = 99.15%. The symmetry of these values indicates that the two error directions are closely balanced.

Table 7. Class-specific performance of the proposed method on the test set.

Metric	Male class (%)	Female class (%)	Overall (%)
Precision	99.15	98.73	-
Recall	98.73	99.15	-
F1-score	98.94	98.94	-
Accuracy	-	-	98.94
Correct predictions	233 of 236	234 of 236	467 of 472

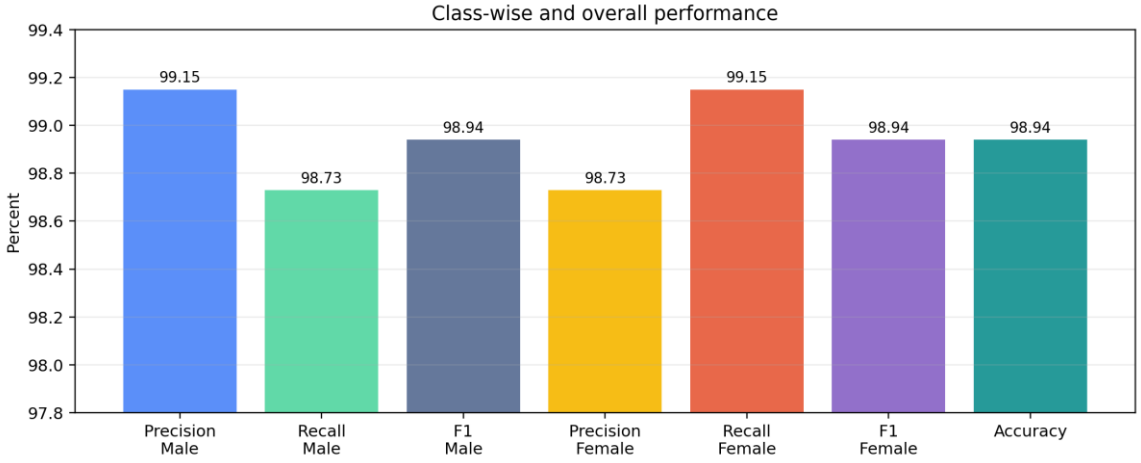

Figure 10. Comparison of class-specific metrics and overall accuracy.

The approximate 95% Wilson confidence interval for overall accuracy is 97.54% to 99.55%. The width of this interval shows that, despite high observed accuracy, the sample remains limited. One or two additional errors would change the reported percentage by several tenths of a point. Very small differences among methods are therefore not meaningful without paired testing on identical predictions or repeated independent runs.

5.3. Contextual Comparison with Published Methods

The proposed framework has the highest reported number among the methods listed in the contextual comparison. Direct comparison across publications is nevertheless threatened by three factors: differences in the exact dataset version and file list, differences in preprocessing and alignment, and differences in hyperparameter selection procedures. When each study uses an independently implemented protocol, the numerical gap reflects both algorithmic quality and experimental variation.

To demonstrate that GOA is the specific source of the improvement, AlexNet features should be evaluated with an MLP trained by Adam, an SVM, and logistic regression using the same split and preprocessing. Paired predictions could then be compared with McNemar's test or bootstrap resampling. Without this ablation, the evidence supports the high performance of the complete framework, but it does not quantify the independent contribution of GOA.

Table 8. Contextual comparison of reported results on GENDER-FERET.

Method	Representation / classifier	Reported accuracy (%)	Comment
Azzopardi et al. [4]	COSFIRE + SVM	93.7	Official GENDER-FERET split
Azzopardi et al. [5]	SURF + COSFIRE + fusion	94.7	Fusion of domain-specific and trainable features
Carletti et al. [13]	CNN for smart cameras	94.7	Emphasis on real-time deployment
Greco et al. [12]	Transfer-learned DenseNet	95.1	Original images before synthetic degradation
Proposed method	AlexNet Pool5 + GOA-MLP	98.94	Reported result on the official split

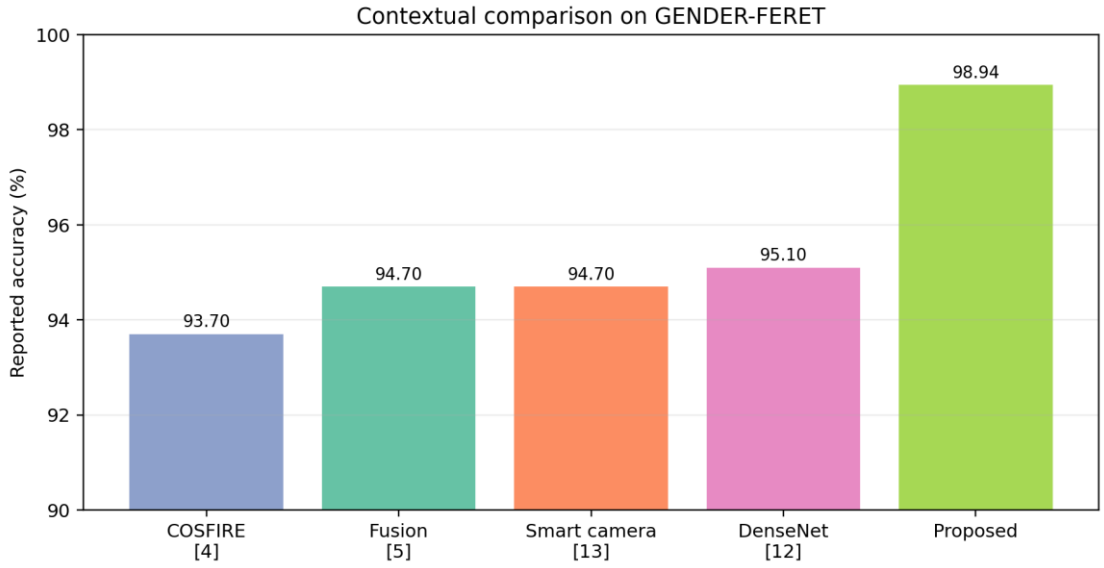

Figure 11. Contextual numerical comparison. Protocol differences prevent a definitive claim of statistical superiority.

5.4. Error Analysis and Technical Interpretation

The five test errors are more informative for qualitative analysis than the overall percentage alone. Inspection of these samples should determine whether failure is associated with lateral illumination, low image quality, unusual expression, occlusion, or structural similarity. If errors are concentrated in a particular subgroup, near-equal error rates for the two main classes would not be sufficient evidence of fairness.

One plausible explanation for the high performance is the use of a frozen feature extractor. With only 474 training images, fine-tuning millions of convolutional parameters can encourage memorization. Pool5 introduces generic ImageNet knowledge as a prior, after which the MLP learns a comparatively shallow mapping between features and labels. The MLP itself still contains approximately 1.18 million parameters, which is large relative to the dataset; weight bounds and the GOA search trajectory must therefore constrain effective complexity.

A second explanation is the exact class balance and identity separation. The classifier cannot achieve high accuracy merely by predicting a majority class. However, balance between male and female labels does not imply balance in age, ethnicity, or image quality. Subgroup distributions may differ, allowing the model to rely on ancillary cues. Feature-attribution or attention maps would help identify this risk.

A third explanation is GOA's capacity to evaluate multiple regions of parameter space. If the initial population is diverse, several candidate decision boundaries are considered simultaneously. As the members approach the best solution, the search becomes more focused. Conversely, in a space of more than one million dimensions, conventional distance relationships may become less informative and premature convergence remains possible. Comparing several population sizes and control schedules is necessary to evaluate this concern.

6. Discussion

6.1. Answers to the Research Questions

The first question concerned whether pretrained AlexNet features are useful for gender-label classification with limited data. The test performance and the small train-test gap indicate that Pool5 supplies sufficient discriminative information for this benchmark. Because the study does not include an identical-protocol comparison with handcrafted features or a randomly initialized AlexNet, however, the exact benefit of pretraining cannot be isolated.

The second question concerned transfer learning. Freezing the convolutional section removes the need for end-to-end training and reduces the risk of overfitting. The result supports the practical use of transfer learning in a low-data scenario. Experiments with ResNet, EfficientNet, or other pretrained backbones would clarify whether the observed behavior is specific to AlexNet or reflects a more general pattern.

The third question examined GOA optimization of MLP parameters. The GOA-MLP framework reaches high accuracy and the fitness curve shows convergence. Yet the absence of direct comparisons with Adam or L-BFGS means that the evidence cannot establish statistically superior performance for GOA. The defensible conclusion is that GOA produced a high-performing classifier, not that it universally outperforms gradient-based optimization.

The fourth question concerned comparison with an MLP trained directly on pixels or on non-deep features. From a theoretical standpoint, Pool5 transforms the image into a higher-level representation and prevents the MLP from having to learn low-level visual structure from raw pixels. Related studies also support the advantage of deep representations over many classical descriptors. A complete empirical answer nevertheless requires MLP-on-pixels and handcrafted-feature baselines under the same split.

6.2. Methodological Strengths

- Clear separation between feature extraction and classification, enabling cached features and direct classifier comparison.
- Use of an identity-disjoint split, which reduces identity leakage.
- Exact balance between the two recorded classes and reporting of separate class-specific metrics.
- No requirement to retrain the full deep network on a small dataset.
- Use of a continuous and interpretable fitness objective rather than discrete accuracy.
- Modularity: the feature extractor or optimizer can be replaced without redesigning the entire pipeline.

6.3. Limitations and Threats to Validity

The primary limitation is the data. GENDER-FERET is appropriate for controlled evaluation, but it is small and its images are predominantly frontal. The 98.94% result should not be generalized to street imagery, low-quality cameras, severe occlusion, or profile views. Cross-dataset testing, in which the model is evaluated without retuning on a different distribution, would provide a stronger assessment of generalization.

The second limitation is computational cost. Searching a 1.18-million-dimensional space with a population of 30 for 100 iterations entails at least 3,000 complete MLP evaluations. If each candidate is evaluated on the entire training set, the procedure is substantially slower than

conventional Adam training. Because execution time is not reported, no claim is made about training efficiency or real-time optimization. Inference after the weights are fixed is faster, but its latency should also be measured.

The third limitation is model selection using a narrow set of metrics. High accuracy and F1 do not provide information about probability calibration, robustness to corruption, memory consumption, or energy demand. In a practical application, expected calibration error, AUC, per-image latency, face-rejection rate, and performance under compression would also matter.

The fourth limitation is demographic bias. Gender Shades and NIST reports have shown that commercial gender-classification systems can exhibit different error rates across skin tones and demographic groups [14, 15]. Equal counts for male and female labels in GENDER-FERET do not resolve this issue. Responsible evaluation requires valid subgroup annotations, adequate sample sizes, and uncertainty-aware stratified reporting.

6.4. Ethical Considerations, Data Governance, and Deployment

Face-based classification in sensitive domains can enable surveillance, discrimination, or disproportionate automated decision-making. A model with high average accuracy may still fail systematically for a subgroup. Its output should therefore not serve as the sole basis for employment, security, medical, or legal decisions. Any use should have a legitimate purpose, data minimization, consent or another lawful basis, limited retention, and a human challenge or appeal mechanism.

The model learns only the labels recorded in the dataset. Although the phrase gender recognition is common in the technical literature, binary gender-label classification in the dataset is conceptually more precise. This wording avoids the unsupported claim that personal identity can be inferred from facial appearance. Model documentation should define the intended scope, out-of-scope populations, and prohibited uses.

Deployment requires evaluation of the full pipeline, including face detection, quality control, classification, confidence thresholds, and a human-review pathway. The system should be able to abstain on low-quality or low-confidence samples. Performance logging without unnecessary image retention, drift monitoring, and periodic subgroup re-evaluation are core governance requirements.

From a security perspective, facial models are vulnerable to presentation attacks, digital manipulation, and adversarial examples. The proposed method contains neither liveness detection nor an adversarial defense. Such modules would need to be added as separate subsystems and assessed using independent criteria.

6.5. Directions for Future Research

- Conduct cross-dataset evaluation on Adience, UTKFace, and unconstrained imagery without test-set tuning.
- Compare GOA directly with Adam, SGD, L-BFGS, PSO, and GWO under an equal computational budget.
- Reduce Pool5 dimensionality using PCA, an autoencoder, or a bottleneck and measure the trade-off among accuracy, memory, and execution time.
- Use GOA for initialization and Adam for local refinement in a hybrid optimization strategy.
- Report the mean, standard deviation, and confidence intervals over at least 20 independent runs.
- Analyze error and performance by age, skin tone, quality, illumination, and pose.
- Add Grad-CAM or other explainability methods to identify the cues used by the model.
- Evaluate probability calibration and introduce an abstention option for low-confidence samples.
- Release code, weights, environment files, split identifiers, and complete logs to enable independent reproduction.

7. Conclusion

This article presented an expanded framework for binary gender-label classification from facial images. The convolutional part of an ImageNet-pretrained AlexNet serves as a frozen feature extractor, Pool5 provides a 9,216-dimensional representation, and a single-hidden-layer MLP performs classification. The algorithmic contribution is the optimization of MLP weights and biases using the Grasshopper Optimization Algorithm. The processing pipeline uses training-derived normalization and the official identity-disjoint GENDER-FERET split.

On the 472-image test set, the reported model produces 467 correct predictions and an accuracy of 98.94%. Precision, recall, and F1-score are nearly symmetric across the two classes, and no obvious class-level imbalance is observed on this split. These findings indicate that a frozen deep representation combined with a metaheuristically optimized classifier can perform strongly on a small, controlled benchmark.

The result should nevertheless be interpreted within its experimental boundaries. It does not establish cross-domain robustness, demographic fairness, or superiority of GOA over gradient-based alternatives. The next priority is not further inflation of benchmark accuracy, but controlled ablation, repeated-run uncertainty estimation, runtime measurement, cross-dataset validation, and responsible subgroup analysis.

Authors' Contributions

Authors equally contributed to this article.

Acknowledgments

Authors thank all participants who participate in this study.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

References

[1] E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1544-1556, 2008, doi: 10.1016/j.patrec.2008.03.016.

[2] C. W. Ng, Y. H. Tay, and B. M. Goi, "Recognizing human gender in computer vision: a survey," *Pattern Analysis and Applications*, vol. 18, pp. 739-755, 2015, doi: 10.1007/s10044-015-0499-6.

[3] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097-1105. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[5] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2010, doi: 10.1109/TKDE.2009.191.

[6] F. Zhuang and et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43-76, 2021, doi: 10.1109/JPROC.2020.3004555.

[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533-536, 1986, doi: 10.1038/323533a0.

[8] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper optimisation algorithm: Theory and application," *Advances in Engineering Software*, vol. 105, pp. 30-47, 2017, doi: 10.1016/j.advengsoft.2017.01.004.

[9] G. Azzopardi, A. Greco, A. Saggese, and M. Vento, "Fusion of domain-specific and trainable features for gender recognition from face images," *IEEE Access*, vol. 6, pp. 24171-24183, 2018, doi: 10.1109/ACCESS.2018.2823378.

[10] G. Azzopardi, A. Greco, and M. Vento, "Gender recognition from face images with trainable COSFIRE filters," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2016, pp. 235-241, doi: 10.1109/AVSS.2016.7738073.

[11] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090-1104, 2000, doi: 10.1109/34.879790.

[12] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643-660, 2001, doi: 10.1109/34.927464.

[13] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635-1650, 2010, doi: 10.1109/TIP.2010.2042645.

[14] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Computing Surveys*, vol. 49, no. 2, p. 37, 2016, doi: 10.1145/2845089.

[15] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295-4304, doi: 10.1109/CVPR.2015.7299058.

[16] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002, doi: 10.1109/TPAMI.2002.1017623.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886-893, doi: 10.1109/CVPR.2005.177.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556.

[20] C. Szegedy and et al., "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016, doi: 10.1109/LSP.2016.2603342.

[22] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of Machine Learning Research*, 2019, vol. 97, pp. 6105-6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html?ref=ji.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

[24] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34-42, doi: 10.1109/CVPRW.2015.7301352.

[25] V. Carletti, A. Greco, G. Percannella, and M. Vento, "Age and gender recognition by smart cameras," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 1203-1215, 2020, doi: 10.1007/s12652-019-01267-5.

[26] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "A robustness evaluation of facial gender classification methods against image degradations," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, pp. 2779-2795, 2022, doi: 10.1007/s12652-021-02985-x.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of Machine Learning Research*, 2015, vol. 37, pp. 448-456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html.

[29] N. Srivastava and et al., "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[30] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867-1874, doi: 10.1109/CVPR.2014.241.