applied sciences

MDPI

Article

Person Identification and Gender Classification Based on Vision Transformers for Periocular Images

Vasu Krishna Suravarapu and Hemprasad Yashwant Patil *

School of Electronics Engineering, Vellore Institute of Technology, Vellore 632014, Tamilnadu, India
* Correspondence: hemprasad.patil@vit.ac.in; Tel.: +91-909-6019-335

Simple Summary: The idea of identifying persons using the fewest traits from the face, particularly the area surrounding the eye, was carried out in light of the present COVID-19 scenario. This may also be applied to doctors working in hospitals, the military, and even in certain faiths where the face is mostly covered, except the eyes. The most recent advancement in computer vision, called vision transformers, has been tested for the UBIPr dataset for different architectures. The proposed model is pretrained on an openly available ImageNet dataset with 1 K classes and 1.3 M pictures before using it on the real dataset of interest, and accordingly the input images are scaled to 224 × 224. The PyTorch framework, which is particularly helpful for creating complicated neural networks, has been utilized to create our models. To avoid overfitting, the stratified K-Fold technique is used to make the model less prone to overfitting. The accuracy results have proven that these techniques are highly effective for both person identification and gender classification.

Abstract: Many biometrics advancements have been widely used for security applications. This field's evolution began with fingerprints and continued with periocular imaging, which has gained popularity due to the pandemic scenario. CNN (convolutional neural networks) has revolutionized the computer vision domain by demonstrating various state-of-the-art results (performance metrics) with the help of deep-learning-based architectures. The latest transformation has happened with the invention of transformers, which are used in NLP (natural language processing) and are presently being adapted for computer vision. In this work, we have implemented five different ViT- (vision transformer) based architectures for person identification and gender classification. The experiment was performed on the ViT architectures and their modified counterparts. In general, the samples selected for train:val:test splits are random, and the trained model may get affected by overfitting. To overcome this, we have performed 5-fold cross-validation-based analysis. The experiment's performance matrix indicates that the proposed method achieved better results for gender classification as well as person identification. We also experimented with train-val-test partitions for benchmarking with existing architectures and observed significant improvements. We utilized the publicly available UBIPr dataset for performing this experimentation.

Keywords: convolutional neural networks; vision transformers; computer vision; periocular biometrics

check for updates

Citation: Suravarapu, V.K.; Patil, H.Y. Person Identification and Gender Classification Based on Vision Transformers for Periocular Images. *Appl. Sci.* **2023**, *13*, 3116. https://doi.org/10.3390/app13053116

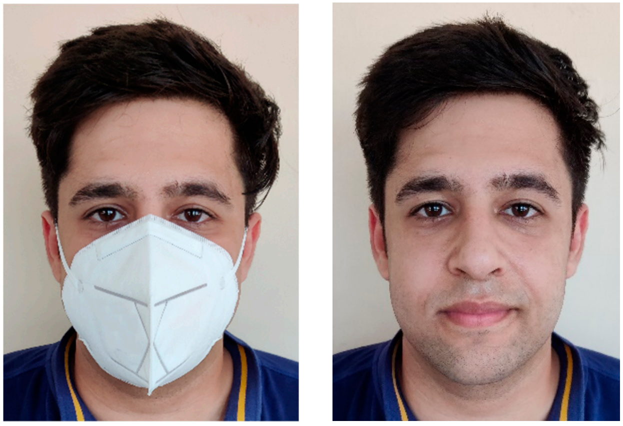
Academic Editor: Jan Egger

Received: 12 December 2022
Revised: 14 January 2023
Accepted: 27 January 2023
Published: 28 February 2023

1. Introduction

Biometrics refers to the science of uniquely identifying an individual based on their physiological (face, iris) or alternatively behavioral (signature, gait) peculiarities [1]. The requirement for biometrics in human identification creates a mutual trust for a wide range of applications that is in everyone's best interest. Building complex models for biometrics is difficult since the differentiating characteristics could be identical. When more pictures of a single individual are taken, it occasionally results in resemblance with other people. While behavioral biometrics are also impacted by socio-environmental variables, physiological biometrics progressively change over time. There are seven requirements for any feature (modality) that can be utilized for biometric identification. They can be listed as permanence, performance, circumvention, universality, acceptability, uniqueness, and measurability [1]. The biometric modalities are fingerprint, face, iris, hand geometry, signature, gait, ear, voice, palmprint, hand vein, and teeth [1]. Digital image processing techniques, along with computer vision, pave the way to develop robust systems to integrate biometrics in day-to-day life. Biometric technology is rapidly growing, and the applications are widespread from security at the airport to mobile banking and law enforcement, etc.

One of the challenges we have been facing for a few years is the COVID-19 pandemic, which is compelling humans to wear a mask. The features available during face recognition earlier [2] are not available for classification. With the way the new variants and mutations of existing ones are on the rise, alternate methods are necessary to identify subjects at the workplace. It can be either for attendance purposes or to identify persons on the go at any place based on the need. The physical contact with the hand needed for biometrics may cause infectious diseases [3], so we need to explore the possibilities of contactless biometrics. In the scenario of wearing a mask, the challenges faced are plenty since the region of interest is limited, and the central part of the face is covered. When we have signboards saying "No entry without a mask", we should have mechanisms to ascertain the person's identity. It is a cause of concern in scenarios where the movement of the public is restricted due to security issues. Figure 1 depicts the sample portion which is not available for identification and highlights indirectly the challenges faced. It is well known that identifying a person based on iris recognition needs good infrastructure, and it is costly and requires cooperation from the user [4]. At the same time, the subject cannot be brought too close to the measuring device because proximity can cause infections. In such a scenario, periocular biometrics is useful for person identification as well as related tasks. A periocular image is defined as the peripheral area of the human eye, which comprises the eyebrows, eyes, and pre-eye orbital portion [5].



(**a**) (**b**)

Figure 1. (**a**) Person's Image with a mask; (**b**) Person's image without a mask to highlight the challenges in identification.

To address the issue of person identification using periocular images, several algorithms have been proposed by the researchers. Convolutional neural networks are popular algorithms that helped to achieve competitive performance metrics for visual classification. Recently, vision transformers [6] have been utilized by computer vision researchers for image classification.

The generic steps followed by vision transformers [6] are as follows:

(A) Perform image resizing and normalization. For resizing, the size of the images will be the same as the image size used during the development of the pre-trained model.

(B) The motivation of vision transformers comes from sentence transformers in NLP (natural language processing). The sentence transformers expect sequences of words as input. However, in the image processing context, sequences of words are not available. Hence, the entire image is split into patches of equal size so that those patches could be further arranged as sequences to be fed to the encoder of the vision transformer. It can be noted that even after splitting the original image, the patches are still in 2-D form.

(C) Each patch is subsequently flattened to form a 1-D vector from a 2-D matrix.

(D) The 1-D vectors are further subjected to linear mapping. This assigns an intermediate representation to the original 1-D vector. It may be a more reduced dimension than the original 1-D vector's dimensionality.

(E) A classification token is appended to every linearly mapped 1-D vector. This additional information will preserve the classification label.

(F) Using a positional encoding function, the vectors are mapped onto a set of values that typically follow frequency patterns.

(G) The vectors are further subjected to an encoding process. This encoding is a two-step procedure that involves the normalization of layers and the multi-head self-attention algorithm. During normalization, a layer-specific mean subtraction and division by standard deviation is performed. For multi-head self-attention, the following process is performed: For every processed 1-D vector pertaining to its 2-D patch, a dot product is computed using query (Q) and key (K). This will yield a cosine similarity between Q and K. The product is normalized with the dimensions of the key vector (division operation). The result is subjected to the softmax function, which will yield a matrix whose values are in the range of 0 to 1. This matrix is multiplied with a value matrix. This process will assign the similarity amongst patches, which is a final output of the multi-head self-attention layer. At the end of this process, the linear mapping is applied to reduce the dimensionality of the concatenated multi-head self-attention matrix.

(H) The MLP (multi-layer perceptron) head defines intermediate layers that finally yield the probability of predefined categories.

The main contributions of this work are (i) competitive performance metrics for person identification as well as gender classification using periocular images and (ii) introduction of the use of concatenated average pool and max pool layers in the vision transformer architecture for periocular image classification. It can be inferred that the proposed technique yields higher accuracies for person identification as well as gender classification.

The goal of this paper is to propose a technique for obtaining enhanced performance metrics for periocular person identification and gender classification.

The remainder of the article is organized in this manner. Section 2 presents a literature review of techniques related to periocular biometrics. Section 3 illustrates the resources and techniques employed, which include the dataset used and transformers in NLP and vision, followed by the proposed model, ROI (region of interest) for pre-processing, and a step-by-step understanding of the proposed approach to obtain better accuracy. In Section 4, we exhibit the results of the experimental work in a tabular form compared with previous results obtained and graphical analysis. Finally, in Section 5 discussion on the results is presented followed by the conclusion in Section 6.

2. Related Works

Among the first researchers in this area were Park et al. [7], who mentioned periocular biometrics categorized under standalone biometric peculiarities. They have emphasized extracting global and local information and used surface-texture and limited-neighborhood operators to obtain a feature matrix for the purpose of categorization. Position variations, masking of the iris, template aging, face occlusion, and eye regions were the utmost discerning features, with pose deviations, template aging, occluded iris, face occlusion, and eye areas being distinguished factors that cause overall accuracy to deteriorate [7,8]. The authors designed a user interface to show human subjects who were shown a pair of ocular images in an experiment. The features were ranked, starting with the most beneficial and ending with the least beneficial, as (i) eye shape, (ii) eyelids, (iii) eyelashes, (iv) eyebrows, (v) tear duct, (vi) skin, and (vii) outer corner. The testers were provided with multiple periocular images and were tasked with ascertaining the validity of the same [9]. When shown images taken in the visible and NIR spectra, participants were told to identify periocular pictures in pairs, whether they belong to the same or a different individual [10].

Projections-based methods like eigenvector-based PCA, derivatives, and discriminant function analysis are also investigated [11]. When the performance of iris identification algorithms deteriorates, the periocular area has been demonstrated to give exceptional recognition rates in complicated picture acquisition settings considering inadequate image blur, iris segmentation, specular reflections, and occlusions caused by eyelids and eyelashes. The findings suggest that with the ocular region's size, system accuracy does not necessarily improve [12]. A periocular image is broken down into several boxlike areas, improving the ability to differentiate the matched biometric pattern. It is achieved by an unsupervised patch selection procedure [13]. A significant space of the periocular neighborhood is viable for obstructed images. The joint representation of periocular texture and structure effectively expresses and poses invariant representation [14]. For less restrictive periocular matching outcomes, a periocular identification technique based on SCNN (semantics-assisted convolutional neural networks) can outperform alternatives in accuracy and matching time [15]. It uses explicit semantic information to extract natural periocular properties automatically. Another study found five local textures or geometrical features based on periocular area landmarks to possess discriminating racial information. Adaboost training combined these features into a prominent feature that enhanced the average accuracy rate [16]. A multi-scale technique was retrieved in which characteristics were from the face, periocular, shoulders, and head area. They were then combined in a two-stage process with a pair of classifiers, resulting in increased efficiency [17]. A merging model technique outperforms the outcome of employing a particular CNN model for the right and left pictures in terms of low-resolution deterioration and blockage. The authors developed two separate CNN models for each eye, one for the left and one for the right, and then combined them to produce a new CNN model [18]. Instead of iris texture, gender-related information was mainly found in the periocular region. Linear SVM and CNN were compared with handcrafted and deep features [19]. Another work implemented periocular recognition for different pose variations and eye comparisons. Based on the transfer learning approach, it was implemented on seven distinct standard deep-learning-based CNNs [20]. In a feature fusion technique, a multiclass SVM classifier was used to evaluate a mix of HOG and non-handcrafted features for three non-ideal conditions, including the impact of spectacles, the effect of eye occlusion, and posture variations [21]. Polygon and Rectangular shape ROIs were recommended for a person wearing a mask based on the optimal size periocular ROI. They found that a broader area was associated with better recognition accuracy [5]. Vision transformers have been utilized in diagnosing COVID-19 in a framework consisting of following major phases: (i) lung localization applying the UNet algorithm succeeded by (ii) classification [22]. The transformer was scaled down to create the CMTs (CNN meets transformers) family of models, which outperformed earlier convolution- and transformer-based models in terms of precision and efficiency [23]. A thorough investigation of multiple indices of ViT model robustness was conducted, with the results compared to ResNet baselines. The authors observed that ViT models are, for the most extent, as reliable as ResNet equivalents on a wide sphere of perturbations when pre-trained with enough data [24]. Transformers are used to solve challenges that are not just restricted to NLP but also include computer vision. ViT paves a new direction of research in this domain of periocular biometrics.

Transformers, first described in 2017 in the well-known work "Attention is All You Need" [25], have quickly become one of the most widely utilized and exemplary designs in natural language processing. In 2020, ViT was utilized for computer vision work, as observed in [6].

The inductive biases that assume spatial structure as an input to CNN-based deep neural networks are not present in vision transformers, allowing them to capture a broader and global range of correlations at the cost of more extended data training. One of the important properties of vision transformers is its ability to extract features using adaptable and changeable receptive fields. Therefore, ViT models can handle defects in the image, like patches negatively affecting the image or other imperfections [21]. To summarize, after the advent of transformers, researchers are exploring the application of ViT architectures to perform computer-vision-related tasks such as image classification, image segmentation, etc.

Pixels are the basic unit of analysis for images. ViT, on the other hand, calculates associations between numerous accompanying pixels in minuscule areas of an image (for instance, 32 × 32 pixels) rather than calculating relationships between every pair of pixels in a typical image. Units (with positional embeddings) are created in succession. The embeddings can be computed in vector form. Every segment is grouped as a sequence in linear form, sent to the transformer, and multiplied by the embedding matrix. In classification problems, the class token is essential. The single input to the final MLP head is a unique token [26]. The transformer encoder converts input tokens in the most typical image classification architectures. The decoder part of the classic transformer layout is also used in a few applications [27].

3. Materials and Methods

3.1. Dataset

For our experiment, we have chosen the UBIPr dataset, which has been publicly available since 2012 and was created by Padole and Proenca [28]. It facilitates benchmarking our results against earlier published work in this domain of periocular biometrics. A few examples of images have been illustrated in Figure 2. The total count of images accessible is 10,252, which have been stored in .bmp format and come from 342 individuals. The acquired images were taken with a Canon E05 5D camera and belong to the visible range of the electromagnetic spectrum. These images vary in the subject's distance from the camera, ranging from 4 m to 8 m, in increments of 1 m. The corresponding image resolution varies from 1001 × 801 pixels to 501 × 401 pixels. They also have variations of 0, 30, and −30 degrees in terms of pose and gaze. The subjects chosen based on gender are in the ratio of 54.4% male and 45.6% female.

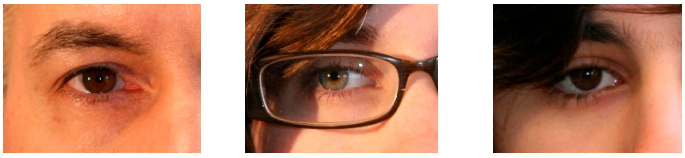


Figure 2. Sample Images (UBIPr Dataset) [28].

The metadata in the dataset highlights the gender, hair occlusion, spectacles, camera distance, canthus points, points on the inner and outer brows, the iris's central point, pose angle, gaze angle, eye closure, and size of the eyes.

3.2. Transformers in NLP

Transformer models are the go-to architectures in NLP. This topic is also gaining interest among researchers in computer vision, namely, vision transformers (ViT) [6].

Let us look into the architecture of the transformers, which has also provided a breakthrough in understanding images based on deriving query, key, and value vectors. The mathematical equation governing scaled dot product attention [25] is indicated as (1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\mathrm{Q}\cdot\mathrm{K}^{T}}{\sqrt{d_k}}\right)V \tag{1}$$

where Q indicates a query vector, K^{T} stands for a transposed key vector and V denotes a value vector. The term d_k indicates the dimension of the key vector, which is used for scaling.

The concept of Q (Query), K (Key), and V (Value) could be better understood by looking at the following. Let us take an example of the retrieval of a video on Youtube [29]. In general, the search systems find the closest match between a query and several internal key terms using similarity measures. Once a key that has smallest similarity measure has been found, a mapping from an internal dictionary that contains the key's corresponding value is returned. Hence, that corresponding value is made available to the user.

We will utilize positional embeddings during this process since the transformer does not process the input in sequential order but rather in parallel. For each element, it combines information from the other aspects through self-attention. We must explicitly encode the order of the inputs for the transformer to help further processing. At this stage, with the help of positional embeddings, the inputs are arranged as a sequence. An additional trainable (class) component is added to the series based on the location. The aforementioned class embedding has been employed to determine the categorization of the source data after it has been changed by self-attention.

Self-attention is a sequence-to-sequence procedure in which a sequence of vectors is fed into the system, and a sequence of vectors emerges. The self-attention process enables inputs to engage ('self') and decides who gets more significant attention. The context of appearance is emphasized in this process. The outputs are aggregates of these attention scores. Transformers are centered on attention mechanisms, referred to as multi-head attention. The scaled dot product attention as well as multi-head attention mechanisms are represented in Figure 3 [25].

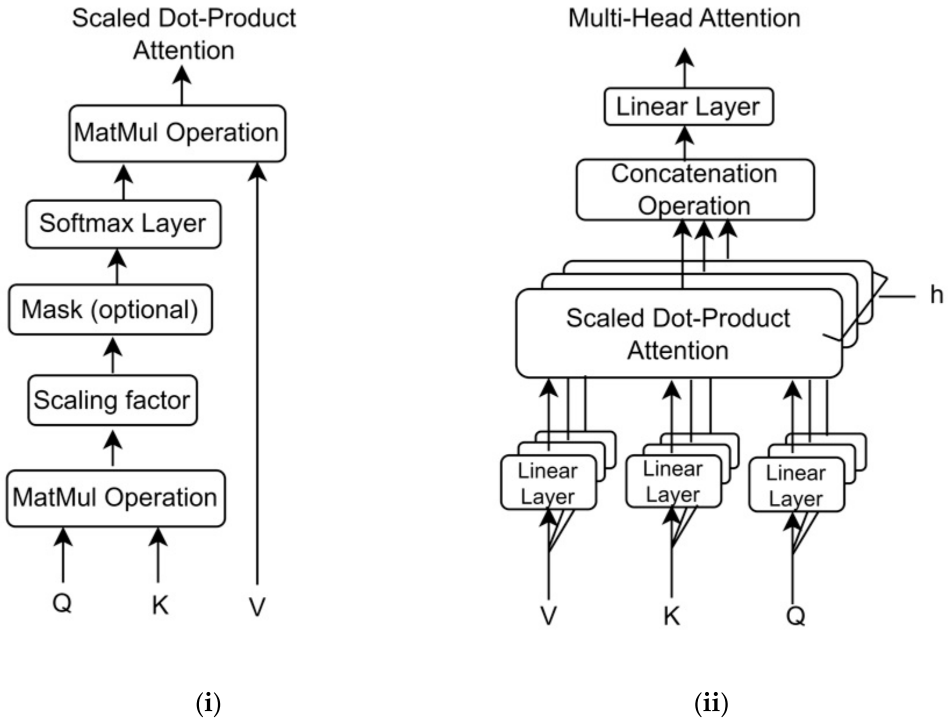


Figure 3. (**i**) Scaled dot product attention and (**ii**) Multi-head attention [25].

3.3. Transformers in Vision

ViT model was introduced in a research work designated "An image is worth 16 × 16 words: Transformers for Image Recognition at Scale" at ICLR 2021 [6]. Transformers can be considered a comprehensive training method that appropriates various types of data for better performance measures. The ViT model treats the whole image (input) as an inherent image-patches collection, similar to how word embeddings are represented when text processing is done with transformers. Figure 4 illustrates the process followed for image classification [6].

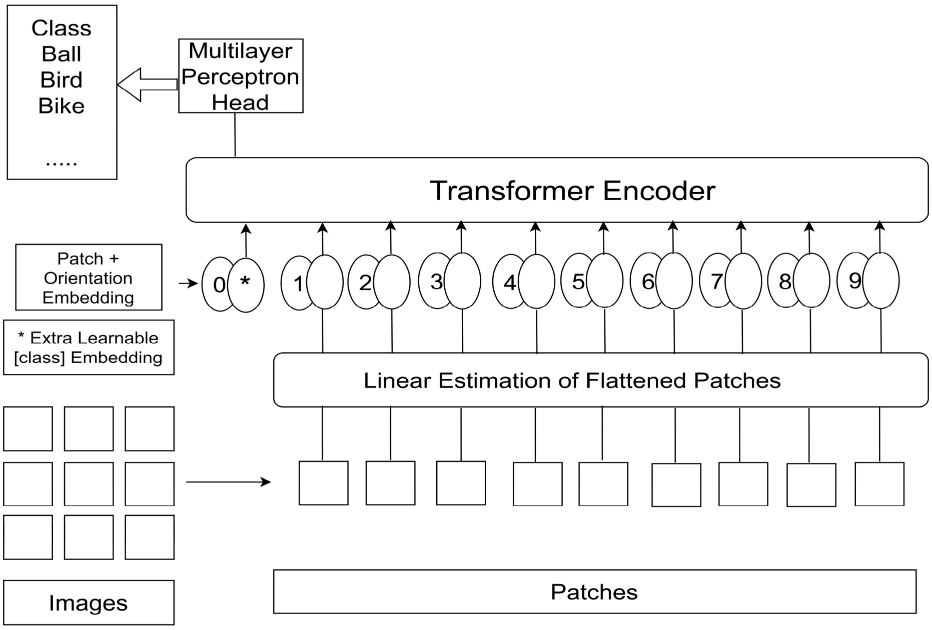

Class
Ball
Bird
Bike
.....
Multilayer Perceptron Head
Transformer Encoder
Patch + Orientation Embedding
* Extra Learnable [class] Embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Estimation of Flattened Patches
Images
Patches

Figure 4. ViT model for the processing of images [6].

For the processing of images based on the ViT Model, they have been initially broken down into fixed-size patches. If an input image is given as H × W × C, wherein H indicates height, W indicates width, C indicates channels, and P indicates patch size, then the new patch image size is P × P × C. The total count of patches (N) is the entire image size divided by the individual patches, i.e., N = (H/P) × (W/P). For example, if the input imagesize is 224 × 224 along with a patch size of 16, then [(224/16) × (224/16)] = [14 × 14] = 196 patches or sequence tokens are constituted.

The next step is flattening the image patches into one dimension, which can be given as $H \times W \times C = N \times P \times P \times C = N \times P^2C$. In general, the value of C is considered as 1 if the original vector is 1-D. The patches are projected via a linear layer that outputs the latent vector of size D. For classification, we include a trainable 'categorization token' in the process. The categorization token refers to a class token. The class token is appended to every linearly mapped 1-D vector. This additional information will preserve the classification label. Such a token is placed at the beginning of the patch sequence. After being sent through the transformer, a learnable embedding of size D will be utilized for classification. This approach is followed in BERT for NLP [30].

To maintain positional coordinates, position clusters are attached to patch embeddings. At the initial time, the position embeddings carry no information about the 2D locations of the patches; thus, all spatial interactions amidst them must be learned from scratch. The encoder receives the resultant sequence of embedding vectors. After passing through the transformer encoder, the class token is permitted to pay attention to meaningful representations of the patches to learn an embedding for classification. The class token is to be extracted from the encoder output. The dot product of the class token and the MLP output gives a vector of 1 × NumClasses, followed by layer normalization and SoftMax to help us with image classification tasks. The model versions involving tiny, small, and base architectures add a new depth to understanding ViT. Several characteristics control the shape of ViT models. The patch size, the dimensionality of patch embeddings and self-attention (width), the size of encoder units (depth), the size of attention heads, and also the MLP block's latent dimensions are all factors to consider (MLP-width) [6]. The specifics are illustrated in Table 1.

Table 1. ViT Architecture Details [30].

ViT Model/Patch	Width	Depth (Layers)	MLP	Heads	Parameters (Million)	GFLOPS
T/16	192	12	768	3	5.5	2.5
B/16	768	12	3072	12	86	35.1
S/16	256	6	1024	8	5.0	2.2
B/32	768	12	3072	12	87	8.7
S/32	384	12	1536	6	22	2.3

To express the model and input patch sizes, we employed simple notation: ViT-S/16, for example, denotes the “small” variation having a 16 × 16 input patch size. Similarly, B denotes “base” and T denotes “tiny”. The models are pretrained based on the timm PyTorch library [31].

3.4. Proposed Approach

This section discusses the pre-processing steps, mathematical framework for ROI extraction from the UBIPr dataset, and our proposed model. The overall flow of the proposed approach is indicated in Figure 5. The diagram illustrates the training and testing phases. The train:val:test based approach follows the split of the dataset in a random manner. During this random split, there may be a scenario where more samples from one class are segregated to a training subset. Subsequently, when such training data is provided to the deep learning architecture, the trained model is biased towards that class that has more samples in the training set. In order to avoid this, it is recommended to use cross validation techniques. The entire dataset is split into 5 folds for k-fold cross-validation with k = 5. At one time, 4 out of 5 (4/5) folds are considered for training the model. Subsequently, the testing data (1 out of 5) (1/5) fold is subjected to the trained model to predict the labels. In the proposed experimentation, the multilayer perceptron can be configured as per the probabilities of the number of classes required at the output layer. For person identification, the MLP is configured as an n-class classifier with $n = 342$ (as per UBIPr dataset specifications). In contrast, for gender classification, the MLP is configured as a 2-class (binary) classifier. We have configured the vision transformer separately for person identification and gender classification.

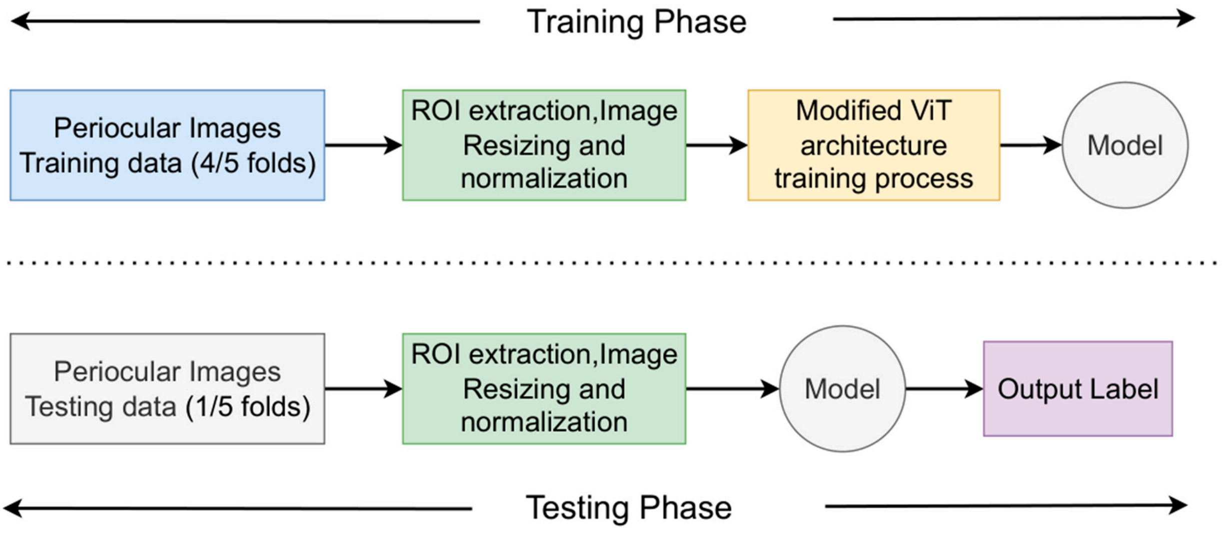

Training Phase
Periocular Images Training data (4/5 folds)
ROI extraction,Image Resizing and normalization
Modified ViT architecture training process
Model
Periocular Images Testing data (1/5 folds)
ROI extraction,Image Resizing and normalization
Model
Output Label
Testing Phase

Figure 5. Overall flow of the proposed approach.

3.4.1. ROI Extraction

The iris center has been used as a region of interest (ROI) by Park et al. [7]. This method fails when gaze angle and pose variations are involved. The individual's eyes are occasionally partially or wholly closed. Due to this, we need to identify other techniques to help recognize our image. The metadata and the UBIPr dataset enable us to retrieve the periocular region of interest, of which the canthus points are the most helpful. We utilized the approach illustrated in Liu et al. [20,32] to extract the ROI from images. The sequence of steps to extract the ROI is mentioned as follows.

Step 1—Let (a_1, b_1) be the coordinates of the medial canthus points (inner corner) and (a_2, b_2) be the coordinates of the lateral canthus points (outer corner). The Euclidean distance (D) between the medial and canthus points is determined as mentioned in (2).

$$\mathrm{D} = \sqrt{(a1 - a2)^2 + (b1 - b2)^2} \quad (2)$$

Step 2—Compute a 2-D coordinate $E_p = (E_{px}, E_{py})$ using (3). This is a point where the line connecting the medial and lateral canthus points meet.

$$E_{px} = (a1 + a2)/2,\ E_{py} = (b1 + b2)/2 \quad (3)$$

Step 3—Calculate the rectangle ROI's upper-left coordinate (a_3,b_3) and lower-right coordinate (a_4,b_4) as mentioned in (4) and (5), respectively.

$$(a_3, b_3) = \left(E_{px} - 1.2 \times \mathrm{D},\ E_{py} - 0.8 \times \mathrm{D}\right) \quad (4)$$

$$(a_4, b_4) = \left(E_{px} + 1.2 \times \mathrm{D},\ E_{py} + 0.8 \times \mathrm{D}\right) \quad (5)$$

Step 4—To get the rectangular ROI, use the calculated points (a_3,b_3) and (a_4,b_4).

3.4.2. Proposed Model

After the extraction of the ROI from the input image, the resultant region is subjected to following set of steps in order to develop the ViT model during the training phase. The trained model is stored in the system for testing purposes. During the testing phase, the model predicts labels. These predictions are utilized to compute accuracy, which determines the overall performance of the trained model.

Step 1—The initial step of development is to apply pre-processing algorithms like resizing to 224 × 224 and normalizing the images in sync with the image-net dataset, whose pre-trained weights are utilized for training, which forms the backbone of the architecture.

Step 2—After this pre-processing step, the images and corresponding labels are randomly shuffled and passed as batches to the model to obtain the raw logits. The images are first broken down into patches and flattened using the linear projection matrix, and the positional embeddings are then added to it.

Step 3—The transformer encoder block, analogous to a block contemplated by Vaswani et al. [25], consists of self-attention blocks, normalization layers, fully connected layers, and residual connections. The attention blocks are multi-headed, and hence they can focus on different patterns of the image.

Step 4—The embedding pertaining to the terminal FC layer is then passed to two pooling layers–the average pool and the max pool. These are then concatenated before finally being passed to the classification head.

Step 5—Instead of the custom train-test split method, a stratified k-fold with k = 5 is used to counter-balance the class imbalance. Instead of randomly splitting the dataset, this ensures that the classes are equally stratified into each fold and hence the model becomes less prone to overfitting.

Step 6—The fully connected layer will output the desired class prediction using the Softmax function, and the class having the foremost value becomes the predicted class.

Step 7—Setting this reliable cross-validation strategy is beneficial during inference because now we have five folds, after completing training for every fold, and we can take the mean of the model predictions as the final class.

The proposed architecture is shown in Figure 6. It may be noted that the average pool and max pool layers are added and concatenated in the ViT architecture before processing the vectors with the MLP head.

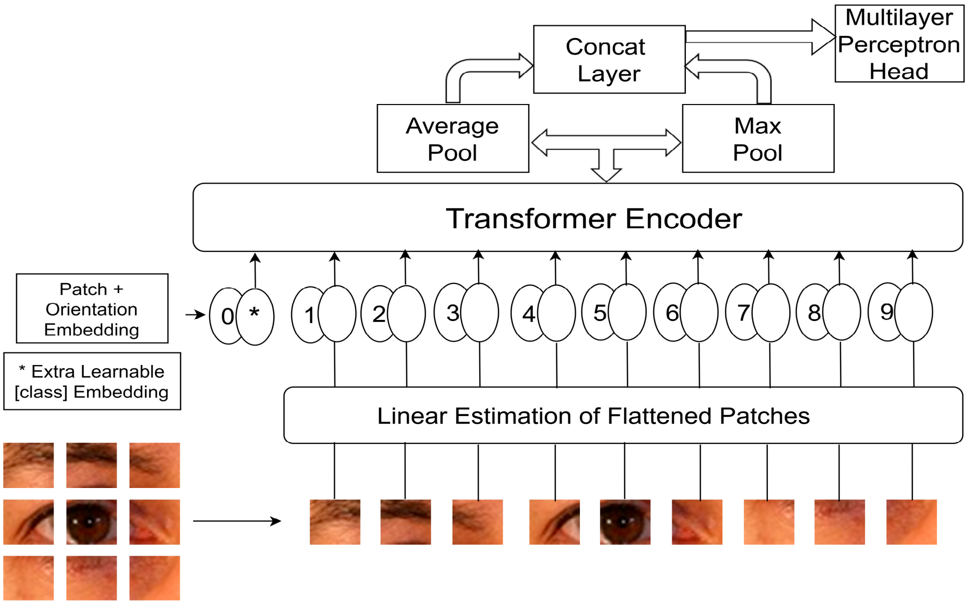

Multilayer Perceptron Head
Concat Layer
Average Pool
Max Pool
Transformer Encoder
Patch + Orientation Embedding
* Extra Learnable [class] Embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Estimation of Flattened Patches

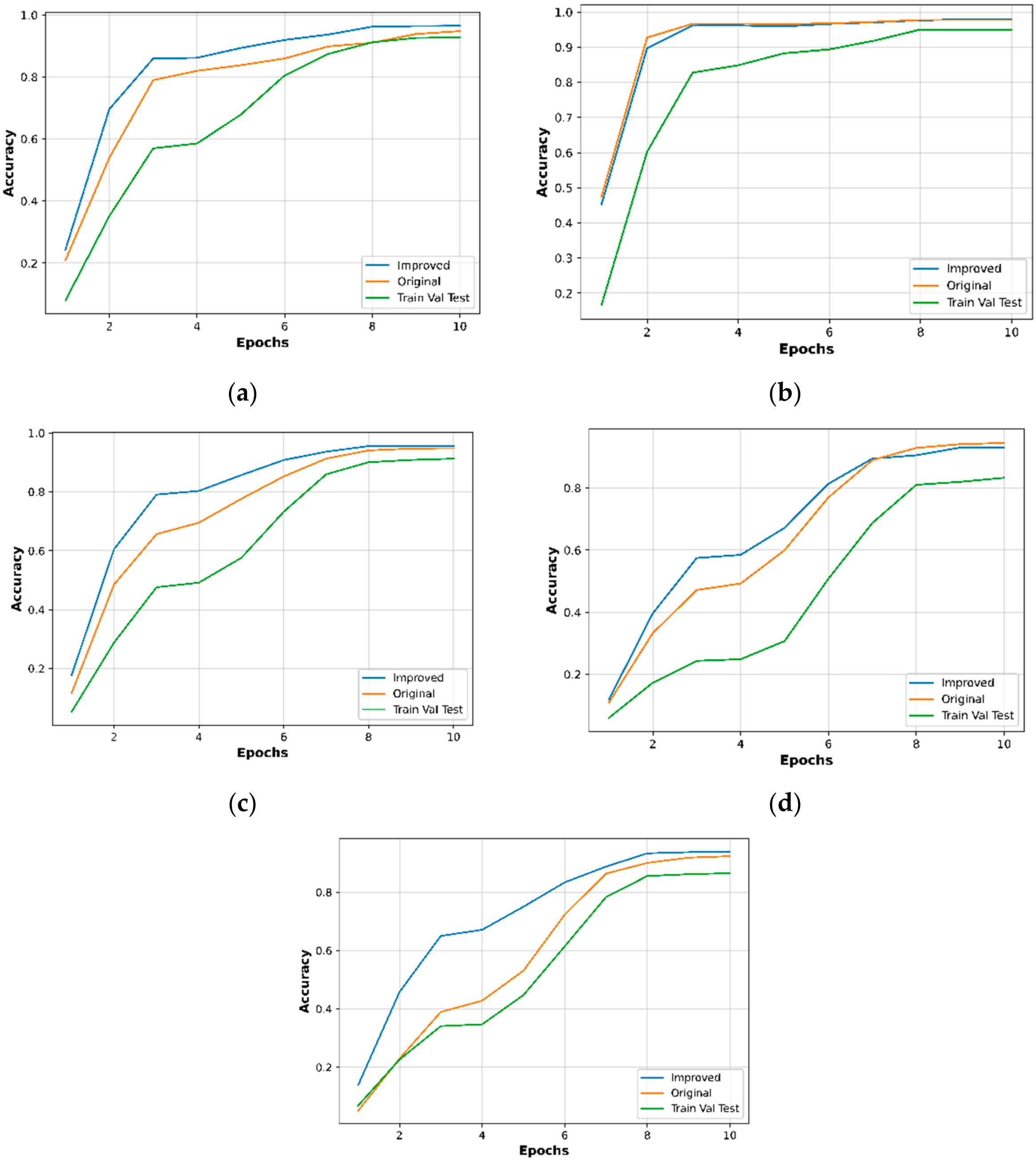
Figure 6. Proposed Architecture.

4. Results

We have implemented our models using the PyTorch framework, which is very useful for building complex neural networks with maximum flexibility and speed. We experimented using ViT for person identification as well as gender classification as illustrated in Sections 4.1 and 4.2, respectively.

4.1. Experiments for Person Identification

A classifier with ‘n’ classes (where ‘n’ stands for the number of unique persons in the dataset, which in the case of the UBIPr dataset is n = 342) is implemented through ViT for person identification. The baseline-architecture-based model was trained, and it yielded a maximum accuracy of 98.19% using 5-fold cross-validation. Subsequently, we experimented with a modified ViT architecture and reiterated the training to obtain a maximum accuracy of 98.15% as indicated in Figure 7b. All experimental results pertaining to five ViT models for person identification are depicted in Figure 7. It may be noted that for the experiments using the PI-ViT-T/16, PI-ViT-B/16, PI-ViT-S/16, and PI-ViT-S/32 architectures, the improved models (after the inclusion of concatenated average pool and max pool layers) always perform better than the corresponding baseline models. In the case of PI-ViT-B/32 based architecture, the improved model performs better than the baseline model until epoch 7. After that, the baseline model performs better than the improved model. However, in the majority of the cases, the improved model yields better accuracy than the baseline model.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 7. (**a**)—PI-ViT-T/16, (**b**)—PI-ViT-B/16, (**c**)—PI-ViT-S/16, (**d**)—PI-ViT-B/32, (**e**)—PI-ViT-S/32.

As illustrated in Figure 7a, the accuracy increases rapidly in a linear fashion up to 3 epochs for the different models with an accuracy from 60 to 80% and then stabilizes around 80 to 90% from epoch 8 to 10. An appreciable improvement is observed in the proposed ViT architecture (ViT-T/16) in the CV Score, which is 96.196 as compared to the original ViT architecture, which is 94.821. From Figure 7b, we observe that in the first two epochs, a steep increase in accuracy is observed and later it flattens for the next eight epochs. There was no significant difference noted in the original ViT architecture as compared with ViT-B/16, though there was an improvement when compared in the ViT model trained based on train-val-test. From Figure 7c, we can observe that the accuracy increases steeply up to three epochs, and then linearly through epoch 7, and flattens later. There is no significant difference found in the ViT original and the ViT-S/16 architecture, though it is significant when compared in the ViT model trained based on train-val-test. We can contemplate from Figure 7d, emphasizing ViT-B/32, that as compared with the other models, there is not a steep increase in the accuracy with the number of epochs; instead, it is more or less piecewise linear for the first eight epochs and then flattens. It can be seen from Figure 7e that the proposed architecture (ViT-S/32) performs better as compared with the other two models for all ten epochs. There is a piecewise linear increase for seven epochs and later it is constant.

Table 2 depicts the accuracy of the five ViT models for person identification under different constraints. It can be noted that the proposed model has achieved higher accuracy than the approaches presented in [33]. For this comparison, we have performed the train-val-test split, which results in (6025-1152-3075) images.

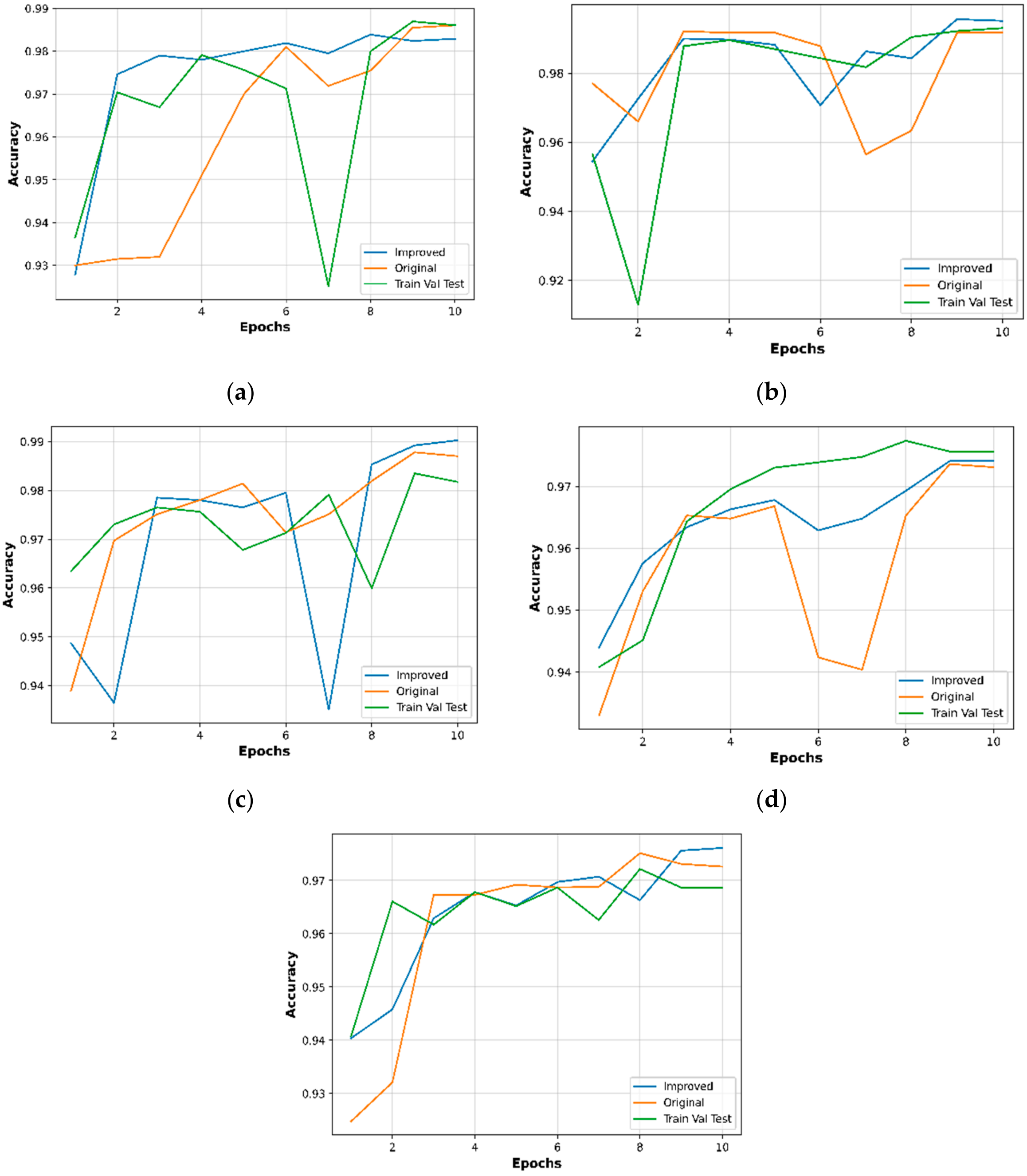
Table 2. Person identification based on different models.

Experiments on the UBIPr Dataset	Models	CV Score (Accuracy)–ViT Models			Earlier Results
		Original	Improved	Train-Val-Test	Train-Val-Test (6025-1152-3075)
		K-Fold (5)	K-Fold (5)	6025-1152-3075	
Person Identification (PI)	ViT-T/16	94.821	96.196	92.580	**93.83** [33]
	ViT-B/16	**98.186**	**98.147**	**96.000**	
	ViT-S/16	94.645	96.079	92.650	
	ViT-B/32	94.625	93.381	85.630	
	ViT-S/32	91.641	93.494	87.840	

4.2. Experiments for Gender Classification

A separate classifier with two classes has been implemented through ViT for gender classification. Hence, we divided the dataset into two classes male and female. The baseline structure was trained with a 5-fold cross-validation strategy, obtaining the maximum accuracy of 99.04%. We further experimented to improve this result by modifying the original ViT architecture. The training was iterated again, obtaining the results with a maximum accuracy of 99.13%, which is indicated in Figure 8b. All experimental results pertaining to five ViT models for gender classification are depicted in Figure 8. It may be noted that for the experiments using the GC-ViT-B/16, GC-ViT-S/16, GC-ViT-B/32, and GC-ViT-S/32 architectures, the improved models (after inclusion of the concatenated average pool and max pool layers) perform better than the corresponding baseline models at epoch 10. In the case of the GC-ViT-T/16 based architecture, the improved model performs better than the baseline model until epoch 9. After that, the baseline model performs better than the improved model. However, in the majority of the cases, the improved model yields better accuracy than the baseline model at the end of the last epoch.

From Figure 8a, we observe that there is some randomness in the values of accuracy for the train-val-test model, but there is a steep increase in the proposed architecture (ViT-T/16) in the first two epochs, and later it flattens with little variation. It is also seen that there is an improvement in the overall accuracy after 10 epochs for all the techniques. From Figure 8b, we see that after the first three epochs, the accuracy falls in the range of 96 to 98%. Though after seven epochs, there has been a dip observed for the model based on the original ViT architecture (ViT-B/16), but it also stabilizes with more than 98% accuracy later. In Figure 8c, it can be seen that the proposed architecture (ViT-S/16) showed a lot of variations in the accuracy as compared with the other techniques, though the final accuracy after 10 epochs was 98.664. The accuracy does not flatten for different epochs, which can also be observed in that fashion due to the scale on the Y axis. From Figure 8d, we can see that there is a steady increase in the accuracy up to 5 epochs, and then is close to flattening through the next five epochs. For the original ViT architecture (ViT-B/32), though, a large dip can be observed after seven epochs. From Figure 8e, we can observe that there is an increase in the accuracy in a linear fashion for the first three epochs, and later for all the remaining models. However, the values are not varying much, with a steady rise after ten epochs for the ViT-S/32-based models.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

Figure 8. (**a**)—GC-ViT-T/16, (**b**)—GC-ViT-B/16, (**c**)—GC-ViT-S/16, (**d**)—GC-ViT-B/32, (**e**)—GC-ViT-S/32.

Table 3 depicts the results for gender classification. Compared to person identification, we have obtained a better accuracy in gender classification with the ViT architectures. It can be noted that the proposed model has achieved higher accuracy than the approach presented in [33]. For this comparison, we have performed the train-val-test split, which results in (6025-1152-3075) images.

Table 3. Gender Classification based on different models.

Experiments on the UBIPr Dataset	Models	CV Score (Accuracy)-ViT Models			Earlier Results
		Original	Improved	Train-Val-Test	Train-Val-Test (6025-1152-3075)
		K-Fold (5)	K-Fold (5)	6025-1152-3075	
Gender Classification (GC)	ViT-T/16	98.605	98.430	97.820	**95.00** [33]
	ViT-B/16	**99.040**	**99.132**	**98.530**	
	ViT-S/16	98.440	98.664	98.170	
	ViT-B/32	97.580	97.464	96.170	
	ViT-S/32	97.776	97.386	96.580	

5. Discussion

The main goal was to showcase the performance of ViT architectures compared to the results obtained earlier with CNN-based models. Our observation from the graphical results is shown in Figure 7. A significant improvement was seen in the accuracy plot, with an increase in the number of epochs for the ViT architectures for the small, tiny, and base models. The maximum accuracy for person identification was found for the ViT base model with 16 patches at 98.186 based on the K-fold technique with K = 5. As seen in Table 3, from the accuracy of the proposed architecture, we can see that there is a significant improvement in the accuracy from the earlier work from 95.0 to 99.132. There was a significant improvement in almost all the other models. An appreciable improvement in the proposed architecture to identify gender from the dataset was observed for the ViT base model from 99.040 to 99.132 based on the K-fold technique with k = 5. There was a similar improvement that can be seen across other models, too. To benchmark our results, we tried to compare our results with earlier published work by authors in this domain. We did a split of 6025-1152-3075 for the number of images into training, validation, and testing. It was seen that the occlusion of the eyes due to hair and the closing of the eyes were a few reasons for the degradation of the system's performance.

6. Conclusions

New systems are required in order to identify people with face masks. In biometrics, the methods frequently used, like fingerprint and face recognition technologies, needed an upgrade since a significant portion of the face is covered by the mask. Our proposed solution addresses this issue by utilizing the latest innovations in deep learning, initially applied in NLP, which is also helpful in computer vision. ViT has been shown to give better performance accuracy. In our experiment, the best model was the ViT base model with 16 patches, which provided an accuracy of 98.186, based on the K-fold technique, with K = 5. It has been seen that with smaller patches, there is an improvement in accuracy. In the future, we plan to develop a hybrid technique to address the critical issues of person identification and gender classification.

Author Contributions: Conceptualization, V.K.S.; Methodology, V.K.S.; Software, V.K.S.; Validation, V.K.S.; Formal analysis, V.K.S.; Investigation, V.K.S.; Visualization, V.K.S.; Supervision, H.Y.P.; Project administration, H.Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from UBIPr Dataset owners (link below), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however, available from the authors upon reasonable request and with the permission of the UBIPr Database owners. (Link: http://iris.di.ubi.pt/ubipr.html).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jain, A.K.; Flynn, P.; Ross, A.A. *HandBook of Biometrics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; ISBN 9780387710402.
2. Agarwal, S.; Punn, N.S.; Sonbhadra, S.K.; Tanveer, M.; Nagabhushan, P.; Pandian, K.K.S.; Saxena, P. Unleashing the Power of Disruptive and Emerging Technologies amid COVID-19: A Detailed Review. *arXiv* **2020**, arXiv:2005.11507.
3. Okereafor, K.; Ekong, I.; Okon Markson, I.; Enwere, K. Fingerprint Biometric System Hygiene and the Risk of COVID-19 Transmission. *JMIR Biomed. Eng.* **2020**, *5*, e19623. [CrossRef]
4. Wei, J.; Wang, Y.; Wu, X.; He, Z.; He, R.; Sun, Z. Cross-Sensor Iris Recognition Using Adversarial Strategy and Sensor-Specific Information. In Proceedings of the 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, FL, USA, 23–26 September 2019. [CrossRef]
5. Kumari, P.; Seeja, K.R. An Optimal Feature Enriched Region of Interest (ROI) Extraction for Periocular Biometric System. *Multimed Tools Appl.* **2021**, *80*, 33573–33591. [CrossRef] [PubMed]
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
7. Park, U.; Jillela, R.R.; Ross, A.; Jain, A.K. Periocular Biometrics in the Visible Spectrum. *IEEE Trans. Inf. Forensics Secur.* **2011**, *6*, 96–106. [CrossRef]
8. Sharma, R.; Ross, A. Periocular biometrics and its relevance to partially masked faces: A survey. *Comput. Vis. Image Underst.* **2023**, *226*, 103583. [CrossRef]
9. Hollingsworth, K.; Bowyer, K.W.; Flynn, P.J. Identifying Useful Features for Human Verification in Near-Infrared Periocular Images. *Image Vis. Comput.* **2011**, *29*, 707–715. [CrossRef]
10. Hollingsworth, K.P.; Darnell, S.S.; Miller, P.E.; Woodard, D.L.; Bowyer, K.W.; Flynn, P.J. Human and Machine Performance on Periocular Biometrics under Near-Infrared Light and Visible Light. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 588–601. [CrossRef]
11. Oh, B.S.; Oh, K.; Toh, K.A. On Projection-Based Methods for Periocular Identity Verification. In Proceedings of the 2012 7th IEEE Conference on Industrial Electronics and Applications, ICIEA, Singapore, 18–20 July 2012; pp. 871–876. [CrossRef]
12. Smereka, J.M.; Kumar, B.V.K.V. What Is a "good" Periocular Region for Recognition? In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 117–124. [CrossRef]
13. Smereka, J.M.; Kumar, B.V.K.V.; Rodriguez, A. Selecting Discriminative Regions for Periocular Verification. In Proceedings of the ISBA 2016—IEEE International Conference on Identity, Security and Behavior Analysis, Sendai, Japan, 29 February–2 March 2016. [CrossRef]
14. Ambika, D.R.; Radhika, K.R.; Seshachalam, D. Fusion of Shape and Texture for Unconstrained Periocular Authentication. *Int. J. Comput. Inf. Eng.* **2017**, *11*, 821–827.
15. Zhao, Z.; Kumar, A. Accurate Periocular Recognition under Less Constrained Environment Using Semantics-Assisted Convolutional Neural Network. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1017–1030. [CrossRef]
16. Chen, H.; Gao, M.; Ricanek, K.; Xu, W.; Fang, B. A Novel Race Classification Method Based on Periocular Features Fusion. *Int. J. Pattern Recognit. Artif. Intell.* **2017**, *31*, 1–21. [CrossRef]
17. Castrillón-Santana, M.; Lorenzo-Navarro, J.; Ramón-Balmaseda, E. On Using Periocular Biometric for Gender Classification in the Wild. *Pattern Recognit. Lett.* **2016**, *82*, 181–189. [CrossRef]
18. Tapia, J.; Aravena, C.C. Gender Classification from Periocular NIR Images Using Fusion of CNNs Models. In Proceedings of the 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis, ISBA 2018. Singapore, 11–12 January 2018; pp. 1–6. [CrossRef]
19. Kuehlkamp, A.; Bowyer, K. Predicting Gender from Iris Texture May Be Harder than It Seems. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, 7–11 January 2019; pp. 904–912. [CrossRef]
20. Kumari, P.; Seeja, K.R. Periocular Biometrics for Non-Ideal Images: With off-the-Shelf Deep CNN & Transfer Learning Approach. *Procedia Comput. Sci.* **2020**, *167*, 344–352. [CrossRef]
21. Naseer, M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H. Intriguing Properties of Vision Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23296–23308.
22. Zhang, L.; Wen, Y. A Transformer-Based Framework for Automatic COVID19 Diagnosis in Chest CTs. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 513–518. [CrossRef]
23. Guo, J.; Han, K.; Wu, H.; Xu, C.; Tang, Y.; Xu, C.; Wang, Y. CMT: Convolutional Neural Networks Meet Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1–14.
24. Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; Veit, A. Understanding Robustness of Transformers for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 10211–10221. [CrossRef]
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5999–6009.

26. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
27. Xia, Y.; He, T.; Tan, X.; Tian, F.; He, D.; Qin, T. Tied Transformers: Neural Machine Translation with Shared Encoder and Decoder. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; Volume 5, pp. 5466–5473. [CrossRef]
28. Padole, C.N.; Proenca, H. Periocular Recognition: Analysis of Performance Degradation Factors. In Proceedings of the 2012 5th IAPR International Conference on Biometrics, ICB 2012, New Delhi, India, 29 March–1 April 2012; pp. 439–445. [CrossRef]
29. Available online: www.youtube.com (accessed on 31 January 2023).
30. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
31. Wightman, R. ViT Training Details Issue #252 Rwightman/Pytorch-Image-Models GitHub. Available online: https://github.com/rwightman/pytorch-image-models/issues/252#issuecomment-713838112,2013.3,9 (accessed on 6 June 2022).
32. Liu, P.; Guo, J.M.; Tseng, S.H.; Wong, K.S.; der Lee, J.; Yao, C.C.; Zhu, D. Ocular Recognition for Blinking Eyes. *IEEE Trans. Image Process.* **2017**, *26*, 5070–5081. [CrossRef] [PubMed]
33. Kumari, P.; Seeja, K.R. A Novel Periocular Biometrics solution for authentication during COVID-19 pandemic situation. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 10321–10337. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.