INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Classifying Gender Based on Face Images Using Vision Transformer

Ganjar Gingin Tahyudin [a], Mahmud Dwi Sulistiyo [a,b,*], Muhammad Arzaki [a], Ema Rachmawati [a,b]

[a] School of Computing, Telkom University, Jl. Telekomunikasi No.1 Terusan Buah Batu, Bandung, 40257, Indonesia
[b] Artificial Intelligence Laboratory, Telkom University, Jl. Telekomunikasi No.1 Terusan Buah Batu, Bandung, 40257, Indonesia
*Corresponding author: *mahmuddwis@telkomuniversity.ac.id*

***Abstract*— Due to various factors that cause visual alterations in the collected facial images, gender classification based on image processing continues to be a performance challenge for classifier models. The Vision Transformer model is used in this study to suggest a technique for identifying a person's gender from their face images. This study investigates how well a facial image-based model can distinguish between male and female genders. It also investigates the rarely discussed performance on the variation and complexity of data caused by differences in racial and age groups. We trained on the AFAD dataset and then carried out same-dataset and cross-dataset evaluations, the latter of which considers the UTKFace dataset. From the experiments and analysis in the same-dataset evaluation, the highest validation accuracy of 0.9676 happens for the image of size 160×160 pixels with eight patches. In comparison, the highest testing accuracy of 0.9843 occurs for the image of size 224×224 pixels with 28 patches. Moreover, the experiments and analysis in the cross-dataset evaluation show that the model works optimally for the image size 224×224 pixels with 14 patches, with the value of the model's accuracy, precision, recall, and F1-score being 0.8174, 0.8188, 0.8189, and 0.8189, respectively. Furthermore, the misclassification analysis shows that the model works optimally in classifying the gender of people between 21-70 years old. The findings of this study can serve as a baseline for conducting further analysis on the effectiveness of gender classifier models considering various physical factors.**

***Keywords*— Gender; classification; face image; vision transformer.**

Manuscript received 21 Jun. 2023; revised 3 Oct. 2023; accepted 6 Nov. 2023. Date of publication 31 Mar. 2024.

I. INTRODUCTION

In the current era of advanced computer vision, a system that can carry out automated monitoring activities has become an inseparable aspect of daily life. Some examples include gesture recognition, body tracking, face recognition, age estimation, and gender classification. Gender classification benefits several applications, such as limiting access to a particular building/room to specific genders and collecting some demographic data [1].

There have been numerous earlier investigations regarding gender classification based on facial images. Liew et al. used the Convolutional Neural Network to classify a gender [2]. Asmara et al. succeeded in classifying gender using the Naïve Bayes method [3]. Mohamed et al. [4] successfully classified gender using several facial features and the K-Nearest Neighbor method. Azzopardi et al. [5] developed a technique to classify genders by combining feature extraction of the eyes, cheeks, and mouth and the Support Vector Machine. Tianyu et al. [6] successfully created a gender classification algorithm using the Multi-Block Local Binary Pattern method to perform feature extraction and the Support Vector Machine for its classification.

Furthermore, Deep Learning via Convolutional Neural Networks is prevalently used in computer vision cases (see, e.g., [7]). According to Dosovitskiy et al. [8], the Vision Transformer model, inspired by the Transformer model originally introduced in 2017 for a Natural Language Processing task [9], can also be used to perform image classification tasks. The Vision Transformer method works by self-attention mechanism, i.e., by looking at the relationship between one element and another [9].

Even though it is only a binary classification, the task of distinguishing gender based on facial image data is currently still an exciting challenge due to the data's high variability and complexity. For broader applications such as security and monitoring systems [10], user personalization [11], product marketing [12], and many more [13], the system is demanded to have high accuracy and handle many variations regarding the differences in ethnicity, age, hairstyle, expression, and lighting. However, the current study that optimizes a gender recognition model and analyzes the data variability and complexity is still lacking.

This paper proposes the Vision Transformer method to classify gender as suggested by Dosovitskiy et al. [8], who conjectured that this method likely outperforms previously state-of-the-art techniques. In addition to observing the use of Vision Transformers, this study also analyzes the reliability of the proposed method for variations in dataset sources, racial differences, and age groups.

The rest of the paper is organized as follows. Section II discusses related works and some literature regarding Transformer, Vision Transformer, and data augmentation for image datasets. Section III presents the description of the proposed method. The results of the model evaluation and its corresponding analysis are discussed in Section IV. Finally, this paper is concluded in Section V.

II. Materials and Method

A. Related Works

Studies on gender classification and its applications have been extensively conducted in the past few decades. Some are based on facial images [14, 15, 16, 17, 18, 19, 20, 21, 22, 23], while others use different modalities such as gait [24, 25, 26], text [27, 28], speech [29, 30], and others. This research focuses on gender recognition using facial images with various variations. Several articles [13, 31, 32, 33] have conducted comprehensive reviews related to this study. Various methods have been applied, and comparative studies have been among them [34, 35]. The proposed solutions can be divided into two main groups: those based on neural networks [14, 36] and those based on other machine learning approaches [3, 4, 5, 6, 15, 17, 20, 23].

There have been numerous investigations related to gender classification based on face images. Previously Liew et al. obtained 99.38% accuracy using the Convolutional Neural Network for such a classification [2]. Here, the authors used the AT&T face database dataset containing 400 facial images, each represented in 32×32 pixels. Asmara et al. achieved 80% accuracy in classifying the gender of facial images using the Naïve Bayes method. Their study considers a dataset of 300 images [3]. Mohamed et al. obtained 99.3% accuracy in gender classification based on facial images using several facial features and the K-Nearest-Neighbor method. Their research considers the FERET dataset and ESSEX database containing 485 and 153 images of 32×32 pixels, respectively [4]. Azzopardi et al. used the SVM method for the gender classification of facial images on GENDER-COLOR-FERET datasets containing 836 images. This method extracted several parts of the face, such as eyes, cheeks, and mouths, and obtained a 96.4% accuracy [5]. Finally, Tianyu et al. [6] performed gender classification of facial images using the combination of Multi-Block Local Binary Pattern and Support Vector Machine. This classification is performed on the Face Fowl library dataset and obtained 94.7% accuracy.

B. Transformer

Transformer was initially a machine learning model designed for natural language cases. It was introduced in 2017 and quickly became one of the state-of-the-art models for classification related to natural language processing [37]. Transformer follows the encoder-decoder concept. The encoder layer works as a continuous input receiver, denoted by $(x_1, x_2, \dots, x_n)$, and transforms an *input mapping operation* to produce $(z_1, z_2, \dots, z_n)$. In the decoder block, $(z_1, z_2, \dots, z_n)$ is transformed into $(y_1, y_2, \dots, y_n)$ in continuous time fashion. Each transformation step is autoregressive, which means each step's output becomes the input for the subsequent iteration [9]. In other words, a Transformer is a sequence representing a deep-learning model that uses a stacked self-attention and pointwise method. Every self-attention block is linked to the fully connected layers on each encoder and decoder sequence [9].

C. Vision Transformer

The Transformer model is not limited to the natural language processing domain. For example, the application of the Transformer model in computer vision was first discussed by Guo et al. [37]. A digital image usually consists of several tuples (x, y) where each tuple has a corresponding intensity value. These tuples are called pixels, the smallest unit in a digital image. In the Transformer model, instead of processing the self-attention of an image in a pixel-by-pixel fashion, the image is initially transformed into a one-dimensional vector. This process is conducted to simplify the computation process [8]. Fig. 1 illustrates converting a two-dimensional image array into a one-dimensional vector.

According to [8], the Vision Transformer works by reshaping the input image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ of size $H \times W$ with C channels into its corresponding sequence of flattened two-dimensional patches $\boldsymbol{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ where (P, P) is the resolution of each image *patch* (a patch is a collection of several pixels, an image is divided into several patches) and $N = HW/P^2$. The value N represents the number of patches of the image.

D. Data Augmentation

Data augmentation is a method to construct a proper deep-learning model by continuously reducing the number of validation errors during the training phase. This method is used to overcome classical problems such as *overfitting* [38, 39]. Overfitting may happen when the accuracy produced in the training phase is considerably higher than that yielded in the validation phase [39].

Data augmentation involves the creation of new data items from the original datasets. There are several possible data augmentation procedures for image datasets, such as geometric and color transformations, random erasing, random zoom, and random flip (see [39] for an extensive bibliography). This study uses rotation, flip, and zoom to obtain augmented datasets for the training and validation phases.

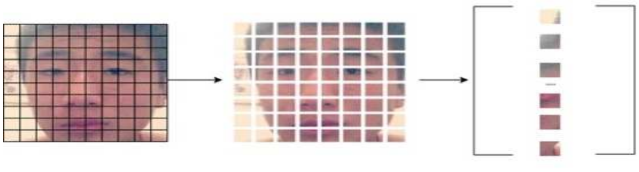

Fig. 1 Conversion from a two-dimensional image array into a one-dimensional vector.

E. Overview of the Proposed Method

In this section, we describe our proposed model for gender classification based on facial images using Visual Transformer. First, we explain the dataset used in the training phase of our model. Second, we discuss some data preprocessing techniques to ensure that each item in the dataset can be processed using the Visual Transformer. Third, we describe the construction of our model and classify dataset items into train and test data. Here, we also discuss the pixels and patch size of the images. Finally, we discuss the model evaluation for measuring the performance of our proposed system.

Our proposed model uses the Vision Transformer to classify the gender of a given facial image. Each image in the previous training data is labeled. This model is summarized in the diagram Fig. 2. Performance satisfaction is quantitatively assessed using some standard evaluation metrics.

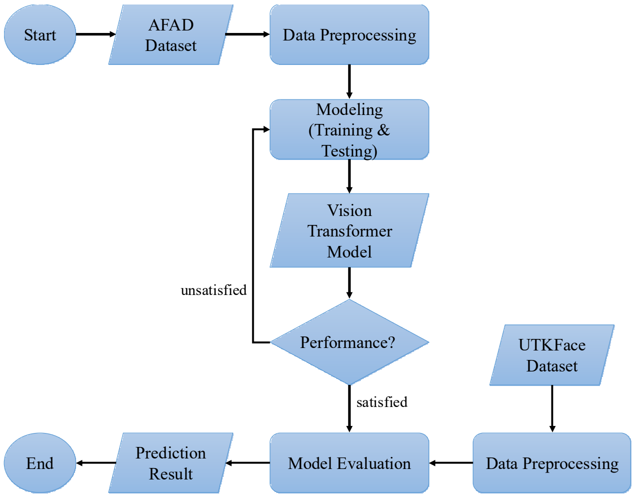

Fig. 2 General flowchart of our proposed model.

F. Datasets for the Training Phase

Our model uses the AFAD (Asian Face Age Dataset) dataset as a training dataset. This dataset contains 165,432 facial images divided into 63,680 images of the female class and 1100,752 $images$of the male category. Fig. 3 illustrates some samples in the AFAD datasets.

G. Data Preprocessing

Our model performs some data preprocessing techniques for the training data. These techniques include data augmentation methods such as random rotation, random zoom, and random horizontal flip. According to Gonzalez and Woods, data preprocessing can enhance the performance of the image classification model [37].

H. Model Construction

Our model uses the Vision Transformer architecture. Initially, the model classifies images into training and testing data. The training process in this study is conducted in several scenarios according to pixel and patch sizes.

Fig. 3 Some images in the AFAD datasets.

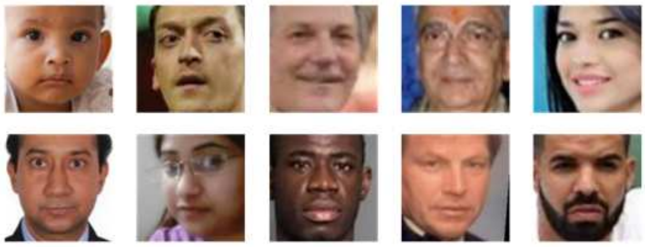

Fig. 4 Some images in the UTKFace dataset.

There are two possible pixel sizes, namely 160 × 160 and 224 × 224. In addition, each pixel size may have different patch sizes, namely 8, 10, 14, 16, 20, 28, and 32. For training purposes, our model uses images of sizes 160 × 160 pixels with $1ten$atches and 20 patches as well as images of sizes 224 × 224 pixels with 14 and 16 patches.

I. Cross-dataset Evaluation and Evaluation Metrics

After training is performed using the AFAD dataset, the proposed model is tested using cross-dataset evaluation against the UTKFace dataset containing 11,316 male and 12,392 female facial images, respectively. Fig. 4 provides some examples of images in the UTKFace dataset.

Both AFAD and UTKFace datasets contain facial images of female and male types. The cross-dataset evaluation is performed using a modified confusion matrix depicted in, adapted from [40]. Here, the actual value is the exact label of an image, while the predicted value is the value obtained from our model prediction. We define a *true male* as a condition when both actual and predicted classes of an image are male. A *true female* is defined analogously; both actual and predicted classes are female. A *false male* is a condition when a female-labeled image is incorrectly predicted as male. In contrast, a *false female* occurs when a male-labeled image is incorrectly predicted as female.

		Actual Value	
		Male	Female
Predicted Value	Male	True Male	False Male
	Female	False Female	True Female

Fig. 5 Modified confusion matrix for our proposed model

Our evaluation metrics use accuracy, precision, recall, and F1-score. These metrics are adapted from the standard binary category classification model described in [41]. In the following formulas, TM, TF, FM, and FF correspondingly denote the condition of true male, true female, false male, and false female.

The model's accuracy is a ratio between the correct prediction (the total proportion of actual males and true females in Fig. 5) and the entire prediction result. Mathematically, it is expressed as follows.

$$accuracy = \frac{TM+TF}{TM+TF+FM+FF} \tag{1}$$

The precision of the model is the ratio between the number of correct results and the proportion of the positive predictive value. For the male class, the precision is defined as follows:

$$precision_{male} = \frac{TM}{TM+FM} \tag{2}$$

While for the female class, the precision is described as follows.

$$precision_{female} = \frac{TF}{TF+FF} \tag{3}$$

In other words, the precision of a particular gender class x is defined as the number of correct predictions for gender class x divided by the number of all predictions related to gender class x. The precision of the system is the arithmetic mean of $precision_{male}$ and $precision_{female}$.

The recall value of a class in our model is defined as the ratio between the number of correct results for that class and the total number of actual items for such class. For the male class, the recall is defined as follows.

$$recall_{male} = \frac{TM}{TM+FF} \tag{4}$$

While for the female class, the recall is defined as follows.

$$recall_{female} = \frac{TF}{TF+FM} \tag{5}$$

In other words, the recall of a particular gender class x is defined as the number of correct predictions for gender class x divided by the total number of actual data related to gender class x. The recall of the system is the average value of $recall_{male}$ and $recall_{female}$.

Finally, the F1-score of the system is defined as the harmonic mean of the previously described precision and recall, which is as follows.

$$\begin{aligned} \text{F1-Score} &= \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \\ \text{F1-Score} &= \frac{2 \cdot precision \cdot recall}{(precision + recall)} \end{aligned} \tag{6}$$

III. Results and Discussion

This section describes the experimental results and discusses them based on several scenarios. In our experiments, we test our model using same-dataset and cross-dataset. The cross-dataset differs from the dataset used in the training phase but has the same domain as the training dataset. We perform both quantitative and qualitative evaluations from the experimental results. The quantitative assessment uses the previously mentioned evaluation metrics such as accuracy, precision, recall, and F1-score. In contrast, the qualitative evaluation visually compares the images associated with the highest validation accuracy and those corresponding to the lowest testing accuracy.

A. Results of Same-datasets Evaluation

In the same-dataset testing, the accuracy of the proposed system is measured after the training phase is completed. There are ten scenarios involving two image sizes, namely, 160×160 pixels and 224×224 pixels. The patches of the images of 224×224 pixels are 8, 14, 16, 28, and 32, whereas the patches of the images of 160×160 pixels are 8, 10, 16, 20, and 32. Fig. 6 depicts an example of an image with 224×224 pixels and 14 patches.

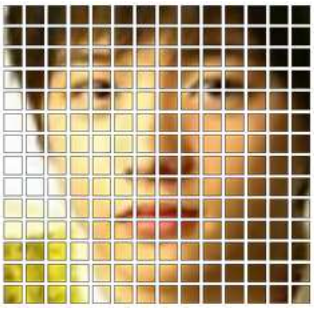

Fig. 6 A 224 × 224 pixels image with 14 patches.

The hyperparameters used in this test have a learning rate of 0.001 with 100 epochs and a transformer layer depth of 12. We used the previously mentioned AFAD dataset and divided it into around 90% for the training phase (containing 91,300 images of male labels and 57,700 images of female labels), around 5% for the validation purpose (containing 5,190 images of male labels and 3,086 images of female labels), and around 5% for the model testing (containing 5,067 images of male labels and 3,208 images of female labels).

TABLE I
Experimental Results for Training Phase (Same-dataset Testing)

Image Size (pixels)	Patch size	Validation Accuracy	Testing Accuracy
160 × 160	8	**0.9676**	0.9661
160 × 160	10	0.9628	0.9598
160 × 160	16	0.9609	0.9610
160 × 160	20	0.9485	0.9460
160 × 160	32	0.9196	0.9154
224 × 224	8	0.9639	0.9631
224 × 224	14	0.9667	0.9644
224 × 224	16	0.9633	0.9622
224 × 224	28	0.9492	**0.9843**
224 × 224	32	0.9381	0.9362

Table I summarizes the quantitative evaluation results for the same dataset testing based on image and patch sizes. We infer that the highest validation phase accuracy of 0.9676 occurred when the image size is 160×160 pixels and the patch size is 8, while the lowest validation phase accuracy of 0.9196 happened when the image is 160×160 pixels and the patch size is 32. For the testing phase, the highest accuracy of 0.9843 occurred when the image size is 224×224 pixels and the patch size is 28, while the lowest testing accuracy of 0.9154 happened when the image size is 160×160 pixels and the patch size is 32.

Fig. 7 provides the qualitative evaluation involving five images, four labeled male. The target describes the actual label of images and $h \times w$ (p) describes the size (in pixel) and the patch numbers of the images used. Here, we visually compare the actual labels of five images and their prediction results when the training considers images of size 160 × 160 pixels with 8 and 32 patches, respectively. From Table I, the validation accuracy for images of size 160 × 160 pixels with 8 patches is the highest among the results in the validation phase. On the other hand, the testing accuracy for images of size 160 × 160 pixels with 32 patches is the lowest among the results in the testing phase. Therefore, we can deduce that increasing the number of patches does not necessarily ensure the correctness of the prediction result.

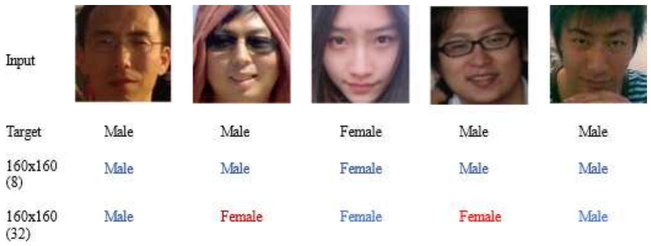

Fig. 7 Samples of qualitative evaluation involving images of 160 × 160 pixels with patch sizes 8 and 32.

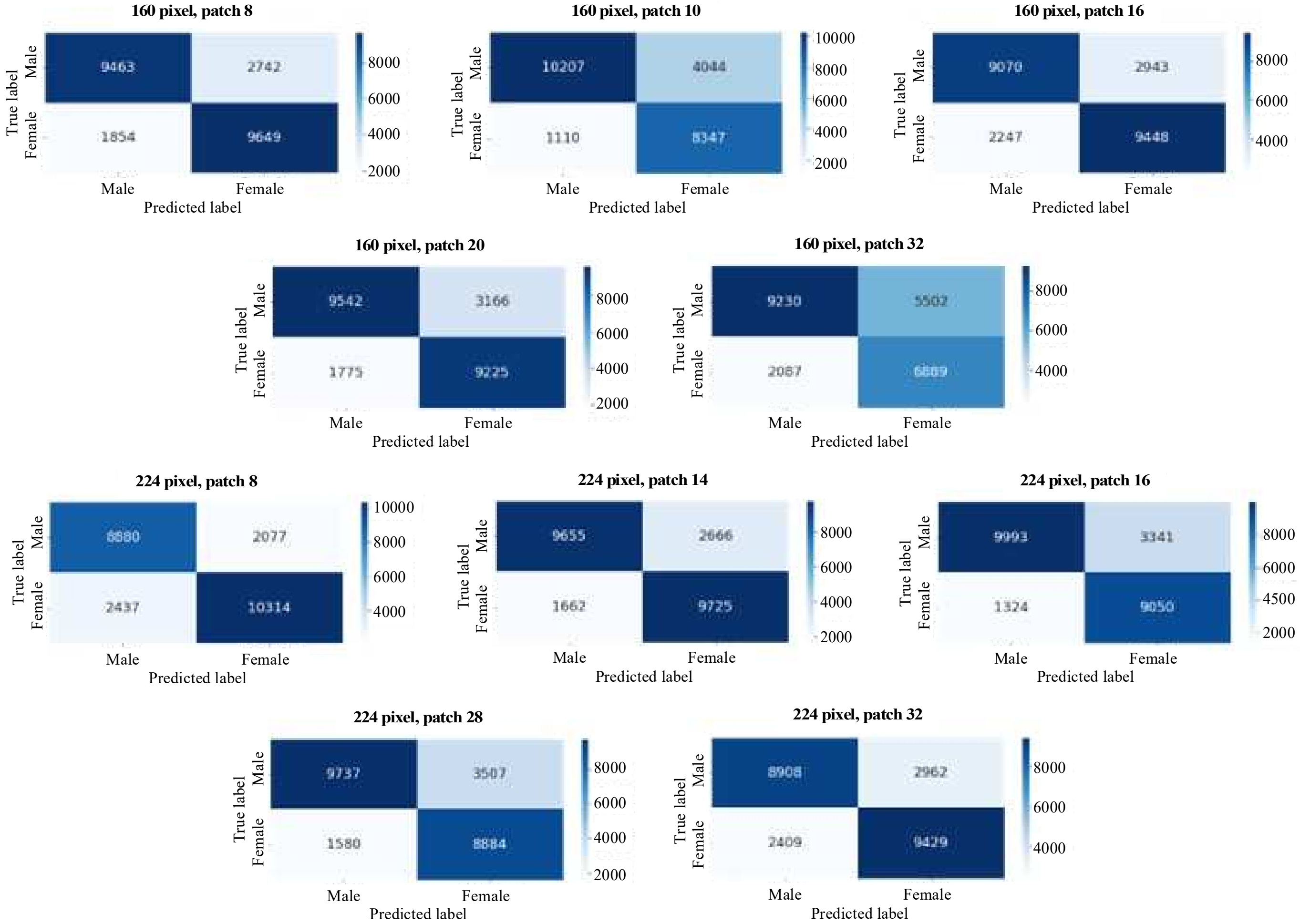

Fig. 8 Confusion matrix from cross-dataset evaluation against the UTKFace dataset.

B. *Results of Cross-datasets Evaluation*

After the model is trained and evaluated using the AFAD dataset, it is then tested using the cross-dataset evaluation technique. Our cross-dataset evaluation uses UTKFace datasets containing 26,132 facial images. This cross-dataset evaluation is measured quantitatively using the confusion matrix described in Fig. 5. The experiment uses two types of pixel sizes, 160 × 160 and 224 × 224. The images in 160 × 160 pixels consider five patch sizes, namely 8, 10, 16, 20, and 32, while the 224 × 224 pixels images employ five patch sizes, namely 8, 14, 16, 28, and 32. The result of this cross-dataset evaluation is summarized in Fig. 8.

The highest level of misclassification of the model when predicting the male gender from the images in which the actual label is female happens for the images of size 224 × 224 pixels with 32 patches. On the other hand, the highest level of misclassification of the model when predicting the female gender in which the actual label is male occurs for the images of size 160 × 160 pixels with 10 patches. The quantitative evaluation for each previously mentioned combination of pixel and patch sizes is summarized in Table II.

From Table II, we infer that the highest accuracy of 0.8174 occurs when the image size is 224 × 224 pixels with 14 patches. This configuration of image size and patches also yields the highest precision, recall, and F1-score values of 0.8188, 0.8189, and 0.8189,, respectively.

In addition, from Table II, we deduce that the lowest accuracy of 0.6789 happens when the image size is 160 × 160 with 32 patches. This combination of image size and patches also produces the lowest precision, recall, and F1-score values of 0.6970, 0.6857, and 0.6913.

TABLE II
CROSS-DATASET EVALUATION RESULTS

Pixel Size	Patch Size	Accuracy	Precision	Recall	F1-Score
160	8	0.8061	0.8070	0.8074	0.8072
160	10	0.7820	0.8050	0.7999	0.8024
160	16	0.7810	0.7814	0.7819	0.7817
160	20	0.7915	0.7947	0.7938	0.7942
160	32	0.6789	0.6970	0.6857	0.6913
224	8	0.8096	0.8096	0.8085	0.8089
224	14	**0.8174**	**0.8188**	**0.8189**	**0.8189**
224	16	0.8032	0.8109	0.8066	0.7983
224	28	0.7854	0.7921	0.7886	0.7903
224	32	0.7734	0.7734	0.7740	0.7737

As in the same-dataset evaluation, we also conducted qualitative analysis for the cross-dataset assessment. Fig. 9 visually compares the actual label of five images with their corresponding prediction results according to the configuration that gives the highest and lowest accuracy. Recall that the highest accuracy occurs when the image size is 224 × 224 and the number of patches is 14, while the lowest accuracy occurs when the image size is 160 × 160 with 32 patches. As in the qualitative evaluation for the same dataset, the higher the number of patches does not correspond to the improvement of the prediction.

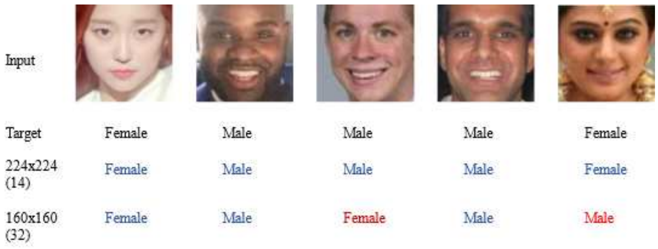

Fig. 9 Samples of qualitative evaluation involving images of 224 × 224 pixels with 14 patches and 160 × 160 pixels with 32 patches.

C. *Misclassification Analysis for the Best Scenario of Cross-dataset Evaluation*

According to the confusion matrix in Fig. 8 and the summary in Table II, the best scenario occurs when the model considers images of 224 × 224 pizes with 114 *patches* This configuration yields the highest accuracy, precision, recall, and F1-score values. Consequently, the misclassification level for this scenario is also the lowest among other combinations. To further analyze the model with this configuration, we analyze the misclassification level for four racial categories, namely white (Caucasian), black (African), Asian (particularly East and Southeast Asian), and Indian. The summary of the misclassification levels for these categories is explained in Fig. 10.

From Fig. 10, we infer that the misclassification rate among the Asian group is the lowest at 20.9%, followed by the African group at 33.9%, then by the Indian group at 44.2%, and finally by the Caucasian group at 50.4%. These results are unsurprising, considering the dataset used in the training phase is the AFAD dataset specifically built from Asian facial images. Moreover, unexpectedly, the system can classify the genders by the facial images of other racial groups with correct classification rates as high as 49%.Some misclassified images for the Caucasian group are depicted in Fig. 11.

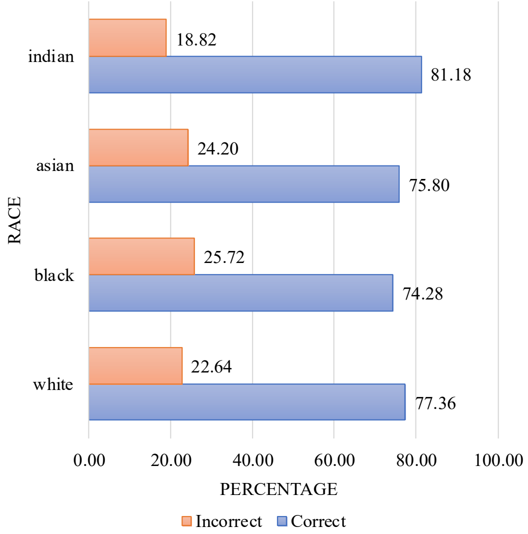

Fig. 10 Misclassification percentage for four different racial groups, i.e., white (Caucasian), black (African), Asian, and Indian.

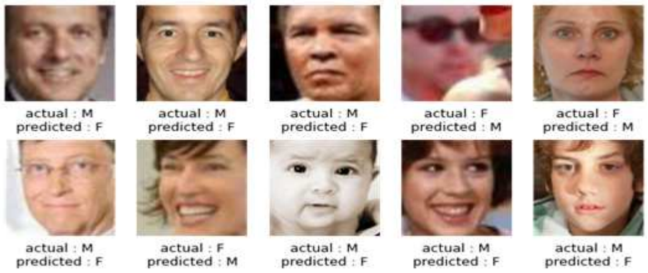

Fig. 11 Examples of misclassified images for the Caucasian group in the best scenario evaluation (pixel size: 224 × 224, number of patches: 14).

We also analyze the misclassification level for different age groups. Here, we divide the test data into twelve different age categories, namely age 0-10, age 11-20, age 21-30, age 31-40, age 41-50, age 51-60, age 61-70, age 71-80, age 81-90, age 91-100, age 101-110, and age 111-116. Fig. 12 summarizes the misclassification level for these age groups.

From Fig. 12, we deduce that our proposed model is most suitable for people in the 41-50 years group. Moreover, the model classification rates for all age groups between 21-70 years are always greater than 80% . The highest misclassification level occurs for the age group 111-116 years, followed by the age groups 91-100, 0-10, and 81-90. For other age groups, the misclassification rate is lower than 27%. One of the reasons is the insufficient training data in the AFAD dataset for those groups. Furthermore, it is visually challenging, even for a human, to distinguish the genders of very young or very old people solely based on their facial images. Some examples of misclassified images for the age group 0-10 in the best scenario evaluation are presented in Fig. 13.

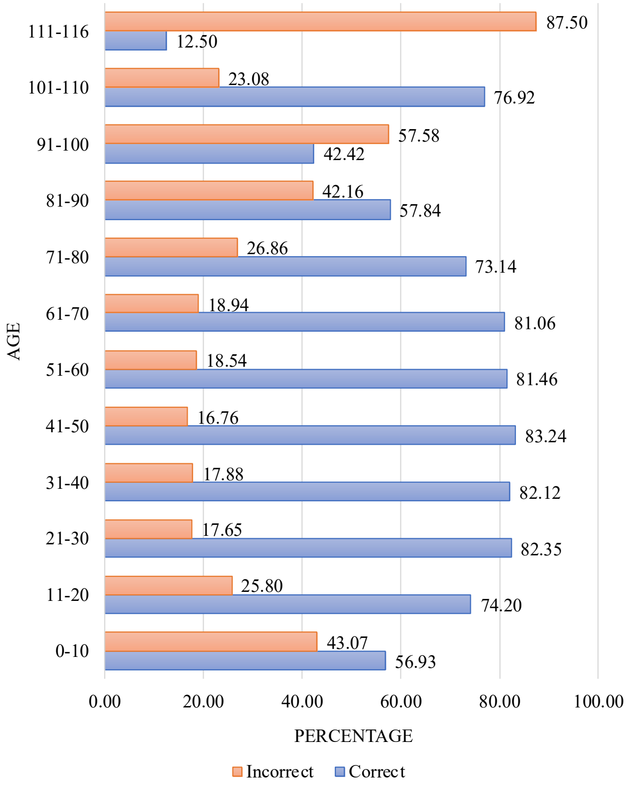

Fig. 12 Misclassification rate for twelve different age groups.

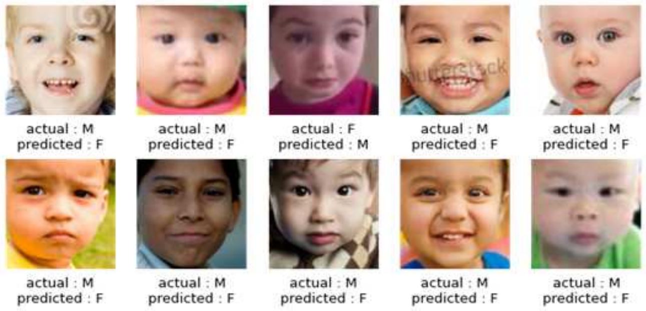

Fig. 13 Examples of misclassified images for the age group 0-10 in the best scenario evaluation (pixel size 224×224, number of patches: 14).

IV. Conclusion

This paper has successfully implemented the Vision Transformer model to classify genders into male and female categories. Observations and analyses regarding the proposed method's reliability for variations in dataset sources, racial differences, and age groups are also provided in this paper. Our experiment and analysis show that the validation and testing accuracies from the same-dataset evaluation for all scenarios are always more than 90%. The highest validation accuracy level of 0.9676 occurs for the image of size 160 × 160 pixels with 8 patches. In comparison, the highest testing accuracy of 0.9843 happens for the image of size 224 × 224 pixels with 28 patches.

The testing accuracy of 0.9843 in the best image configuration for the same-dataset testing is higher than those obtained by Asmara et al. [3], Azzopardi et al. [5], and Tianyu et al. [6]. On the other hand, the results and analysis from the cross-dataset evaluation show that the optimal configuration for the image is 224 × 224 pixels with 14 patches. In this setting, the system's accuracy, precision, recall, and F1-score are respectively 0.8174, 0.8188, 0.8189, and 0.8189, which is lower than those obtained from the previous work except that obtained by Asmara et al. [3].

Our proposed model may not outperform the previous state-of-the-art technique. Still, the result of our experiments provides important insight for further research, i.e., the Vision Transformer model can be used to classify genders based on facial images. We conjecture that combining our proposed model and other techniques likely outperformed the current state-of-the-art methods for gender classification based on face images.

Our model is specifically designed for Asian facial images. However, the misclassification analysis shows that this model can also be used to classify genders of other races' facial images with a correct classification rate as high as 49%. Moreover, the misclassification analysis shows that the model works optimally in classifying the gender of people between 21-70 years old. As an opportunity for further development, the misclassification rate of the model can be made lower by enhancing the datasets with other data items, such as facial images of other racial groups and facial images of very young and very old people.

References

[1] B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 707-711, 2002.

[2] S. S. Liew, M. K. Hani, S. A. Radzi and R. Bakhteri, "Gender classification: a convolutional neural network approach," *Turkish Journal of Electrical Engineering and Computer Sciences,* vol. 24, no. 3, pp. 1248-1264, 2016.

[3] R. A. Asmara, B. S. Andjani, U. D. Rosiani and P. Choirina, "Klasifikasi Jenis Kelamin Pada Citra Wajah Menggunakan Metode Naive Bayes," *Jurnal Informatika Polinema,* vol. 4, no. 3, pp. 212-217, 2018.

[4] S. Mohamed, N. Nour and S. Viriri, "Gender identification from facial images using global features," in *2018 Conference on Information Communications Technology and Society (ICTAS)*, 2018.

[5] G. Azzopardi, P. Foggia, A. Greco, A. Saggese and M. Vento, "Gender recognition from face images using trainable shape and color features," in *2018 24th International conference on pattern recognition (ICPR)*, 2018.

[6] L. Tianyu, L. Fei and W. Rui, "Human face gender identification system based on MB-LBP," in *2018 Chinese Control And Decision Conference (CCDC)*, 2018.

[7] J. Singh and S. Shekhar, "Road damage detection and classification in smartphone captured images using mask R-CNN," in *IEEE BigData Cup 2018 Workshop*, 2018.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," in *2021 International Conference on Learning Representations* , 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[10] S. Poornima, N. Sripriya, B. Vijayalakshmi and P. Vishnupriya, "Attendance monitoring system using facial recognition with audio output and gender classification," in *the 2017 International Conference on Computer, Communication and Signal Processing*, 2017.

[11] Y. Yu and T. Yao, "Gender classification of Chinese Weibo users," in *International Conference on E-commerce, E-Business and E-Government*, 2017.

[12] W. Kim, C. Di Benedetto and R. A. Lancioni, "The effects of country and gender differences on consumer innovativeness and decision processes in a highly globalized high-tech product market," *Asia Pacific Journal of Marketing and Logistics,* vol. 23, no. 5, pp. 714-744, 2011.

[13] F. Lin, Y. Wu, Y. Zhuang, X. Long and W. Xu, "Human gender classification: A review," *International Journal of Biometrics,* vol. 8, no. 3-4, pp. 275-300, 2016.

[14] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[15] T. Jabid, M. H. Kabir and O. Chae, "Gender classification using local directional pattern (LDP)," in *20th International Conference on Pattern Recognition*, 2010.

[16] S. Gutta, H. Wechsler and P. J. Phillips, "Gender and ethnic classification of face images," in *Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

[17] N. Sun, W. Zheng, C. Sun, C. Zou and L. Zhao, "Gender classification based on boosting local binary pattern," in *Third International Symposium on Neural Networks*, Chengdu, 2006.

[18] A. B. A. Graf and F. A. Wichmann, "Gender classification of human faces," in *Biologically Motivated Computer Vision: Second International Workshop*, Tübingen, 2002.

[19] M. Afifi and A. Abdelhamed, "Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces," *Journal of Visual Communication and Image Representation,* vol. 62, pp. 77-86, 2019.

[20] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognition Letters,* vol. 33, no. 4, pp. 431-437, 2012.

[21] D. Yaman, F. I. Eyiokur and H. K. Ekenel, "Multimodal age and gender classification using ear and profile face images," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[22] C. Chen and A. Ross, "Evaluation of gender classification methods on thermal and near-infrared face images," in *International Joint Conference on Biometrics*, 2011.

[23] H.-C. Lian and B.-L. Lu, "Multi-view gender classification using local binary patterns and support vector machines," in *International Symposium on Neural Networks*, 2006.

[24] S. Yu, T. Tan, K. Huang, K. Jia and X. Wu, "A study on gait-based gender classification," *IEEE Transactions on Image Processing,* vol. 18, no. 8, pp. 1905-1910, 2009.

[25] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón and M. Castrillón-Santana, "Gait analysis for gender classification in forensics," in *Dependability in Sensor, Cloud, and Big Data Systems and Applications: 5th International Conference*, Guangzhou, 2019.

[26] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón and M. Castrillón-Santana, "Gender classification on 2D human skeleton," in *3rd International Conference on Bio-engineering for Smart Technologies*, 2019.

[27] H. Mubarak, S. A. Chowdhury and F. Alam, "Arabgend: Gender analysis and inference on arabic twitter," *arXiv preprint arXiv:2203.00271,* 2022.

[28] F. Rangel, P. Rosso, M. Montes-y-Gómez, M. Potthast and B. Stein, "Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter," Working notes papers of the CLEF 192, 2018.

[29] K. Rakesh, S. Dutta and K. Shama, "Gender Recognition using speech processing techniques in LABVIEW," *International Journal of Advances in Engineering & Technology,* vol. 1, no. 2, p. 51, 2011.

[30] A. Tursunov, Mustaqeem, J. Y. Choeh and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors,* vol. 21, no. 17, p. 5892, 2021.

[31] P. Rai and P. Khanna, "Gender classification techniques: A review," in *Advances in Computer Science, Engineering & Applications: Proceedings of the Second International Conference on Computer Science, Engineering and Applications*, New Delhi, 2012.

[32] B. Jaeger, W. W. A. Sleegers and A. M. Evans, "Automated classification of demographics from face images: A tutorial and validation," *Social and Personality Psychology Compass,* vol. 14, no. 3, p. e12520, 2020.

[33] C.-B. Ng, Y.-H. Tay and B.-M. Goi, "A review of facial gender recognition," *Pattern Analysis and Applications,* vol. 18, pp. 739-755, 2015.

[34] S. A. Khan, M. Ahmad, M. Nazir and N. Riaz, "A comparative analysis of gender classification techniques," *International Journal of Bio-Science and Bio-Technology,* vol. 5, no. 4, pp. 223-244, 2013.

[35] A. Hadid, J. Ylioinas, M. Bengherabi, M. Ghahramani and A. Taleb-Ahmed, "Gender and texture classification: A comparative analysis using 13 variants of local binary patterns," *Pattern Recognition Letters,* vol. 68, pp. 231-238, 2015.

[36] S. Amilia, M. D. Sulistiyo and R. N. Dayawati, "Face image-based gender recognition using complex-valued neural network," in *3rd International Conference on Information and Communication Technology*, 2015.

[37] P. Guo, Z. Xue, L. R. Long and S. Antani, "Cross-dataset evaluation of deep learning networks for uterine cervix segmentation," *Diagnostics,* vol. 10, no. 1, p. 44, 2020.

[38] X. Ying, "An Overview of Overfitting and its Solutions," in *Journal of physics: conference series*, 2019.

[39] T.-C. Pham, C.-M. Luong, M. Visani and V.-D. Hoang, "Deep CNN and data augmentation for skin lesion classification," in *Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018*, Dong Hoi, 2018.

[40] B. Xia, H. Zhang, Q. Li and T. Li, "PETs: a stable and accurate predictor of protein-protein interacting sites based on extremely-randomized trees," *IEEE transactions on nanobioscience,* vol. 14, no. 8, pp. 882-893, 2015.

[41] T. Anwar and S. Zakir, "Deep learning based diagnosis of COVID-19 using chest CT-scan images," in *2020 IEEE 23rd International Multitopic Conference*, 2020.