

Age Group and Gender Classification from Facial Images Based on Deep Neural Network Fusion

Eren KARAKAŞ¹, İbrahim Yücel ÖZBEK*¹

¹ Department of Electrical and Electronic Engineering, Faculty of Engineering, Ataturk University, Erzurum, Turkey

Geliş / Received: 27/02/2021, Kabul / Accepted: 29/03/2021

Abstract

The rapid development of smart applications paved the way for systems to make automatic inferences based on artificial intelligence and caused the research to shift towards this area. The human face is important in this respect as many features such as emotion, expression, age, gender can be obtained. In this article, various experiment have been conducted on age and gender classification from facial images. Also, algorithms based on convolution neural networks have been trained with different transfer learning methods and several modifications have been provided to increase the performance of the system. Especially, the results obtained from the models are fused to increase the success of the proposed method. Adience data set was used to measure the success of the proposed method in the study. This data set was created from unfiltered real-life photographs, and therefore, it may contain images with low quality. According to the obtained experimental results, the accuracy performance of the proposed method for gender and age classification was detected as 92.29% and 60.26%, respectively.

Keywords: deep learning, gender classification, age classification, Adience, majority voting

Derin Sinir Ağlarının Füzyonu ile Yüz İmgelerinden Yaş Grubu ve Cinsiyet Sınıflandırma

Öz

Akıllı uygulamaların hızla gelişmesi sistemlerin yapay zekâya dayalı otomatik çıkarımlar yapmasının önünü açmış ve araştırmanın bu alana doğru kaymasına neden olmuştur. İnsan yüzü bu açıdan önemlidir. Çünkü duygu, ifade, yaş, cinsiyet gibi birçok özellik insan yüzünden elde edilebilir. Bu makalede yüz görüntülerinden yaş ve cinsiyet sınıflandırması üzerine birçok araştırma yapılmıştır. Evrimsel sinir ağı tabanlı birçok algoritma çeşitli transfer öğrenme yöntemleri ile eğitilmiş ve sistem performansının artması için birçok modifikasyonlar yapılmıştır. Özellikle modellerden elde edilen sonuçlar füzyon yapılarak önerilen yöntemin başarısı artırılmıştır. Çalışmada önerilen yöntemin başarısını ölçmek için Adience veri seti kullanılmıştır. Bu veri seti filtrelenmemiş gerçek fotoğraflardan oluşturulmuştur ve bu nedenle düşük görüntü kalitesine sahip görüntüler içermektedir. Elde edilen deneysel sonuçlara göre önerilen yöntemin doğruluk performansı cinsiyet ve yaş sınıflandırması için sırasıyla %92,29 ve %60,26 olarak bulunmuştur.

Anahtar Kelimeler: derin öğrenme, cinsiyet sınıflandırma, yaş sınıflandırma, adience, çoğunluk oylayıcı

1. Introduction

The increasing widespread use of deep learning applications, which is a sub-branch

of machine learning, is a natural result of the generation of big data and the increase in the speed and power performance of the processors. Applications in smart phones,

*Corresponding Author: iozbek@atauni.edu.tr

biometric security systems, personalized medication, unmanned aerial and ground vehicles, computer games are just a few of the applications that deep learning seeks to find a solution or development. Determining the age and gender of a person from the picture is one of the most basic steps of biometric security applications, and numerous studies have been published on this subject in the literature. Some of these studies are described as follows

Levi and Hassner (2015) tried to bridge the gap between the abilities of automatic facial recognition and those of the methods of estimating age and gender utilizing deep convolutional neural networks (CNN). Apart from using a lean network architecture, they have applied dropout and data augmentation methods to further limit the risk of overfitting. Wolfshaar et al. (2015) brought a different perspective with their study and designed a hybrid machine learning system. The proposed approach is based on combining a predefined convolutional neural network (CNN) with a linear support vector machine (SVM). Zhang et al. (2017) proposed a new Residual of Residual network (RoR) architecture for age and gender classification in natural life with high-resolution facial images. Recently, Agbo-Ajala and Viriri (2020) have proposed a new end-to-end CNN approach in their studies. The double-level CNN architecture includes feature extraction and classification itself. Specifically, they handle unfiltered images with a powerful image preprocessing algorithm that prepares and processes facial images before forwarding them to the CNN model. Technically, their networks were pre-trained on an IMDB-WIKI with noisy tags, followed by MORPH-II and finally fine-tuned in the training set of the OIU-Adience (original) dataset.

In the proposed study, it was aimed to increase the success rate of automatic age and gender classification and decrease the margin of error. For this purpose, Adience (Eidinger, Enbar, & Hassner, 2014) data set, which is frequently used in age and gender classification, was used.

Multiple deep neural network architectures were tested and various network parameters were adjusted to achieve the best result in the study. These deep neural networks were finally fused by the majority vote method, and thus, a very high age and gender classification success rate was obtained.

The details of the study are given as follows. In Chapter 2, the deep neural network architectures used for the training algorithm, the data set and performance criteria used in the study are explained. A detailed description of the tests and the obtained results in this work are given in Chapter 3. Finally, Chapter 4 outlines the conclusion of this study and gives future work plan.

2. Material and Methods

The human face contains many features such as emotion, expression, gender, identity and age. The rapid increase of smart applications today has revealed the importance of automatic detection of these facial features. Age estimation has started to be used in areas such as access control, human-computer interaction, authorization and surveillance, search, tracking by law enforcement. (Tivive and Bouzerdoum 2006)

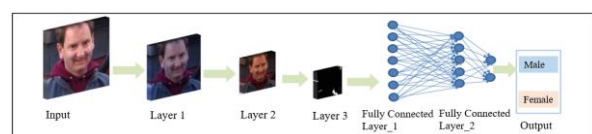


Figure 1. Example of gender classification CNN network

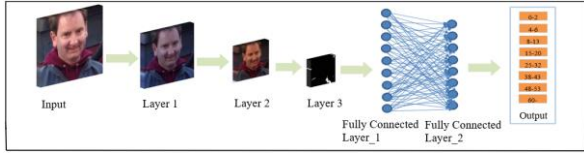


Figure 2. Example of age classification CNN network

In this study, it is aimed to find high-performance solutions to the problem of age prediction and gender classification by using face images. The main challenge here is the design and implementation of highly accurate systems using different feature extractors or even combining them together. For this purpose, various deep neural architectures based on CNN and Pre-RoR are used and their individual performance are combined to increase the performance with the majority voting method. The deep neural architectures used in this work are described below.

Convolutional neural network (CNN) architecture

The CNN structure used in this study was designed based on the architectural structure proposed by Levi and Hassner (2015). Here, the number of layers of the network was increased and various hyper parameters were changed. Examples of age and gender classification CNN network can be seen in Figure 1 and Figure 2. In the convolutional neural network architecture, the network structures in age and gender classification are almost the same to each other as seen in Figure 3. While the last fully connected layer has eight output neurons in age classification, that of in gender classification has two. Among the parameters used in training, initial learning rate, mini-batch size and maximum number of epochs was defined as 0.001, 32 and 150,

respectively. For training and validation, the 256×256 input images were cropped to a random size of 227×227 pixels while for testing same size center crop was used.

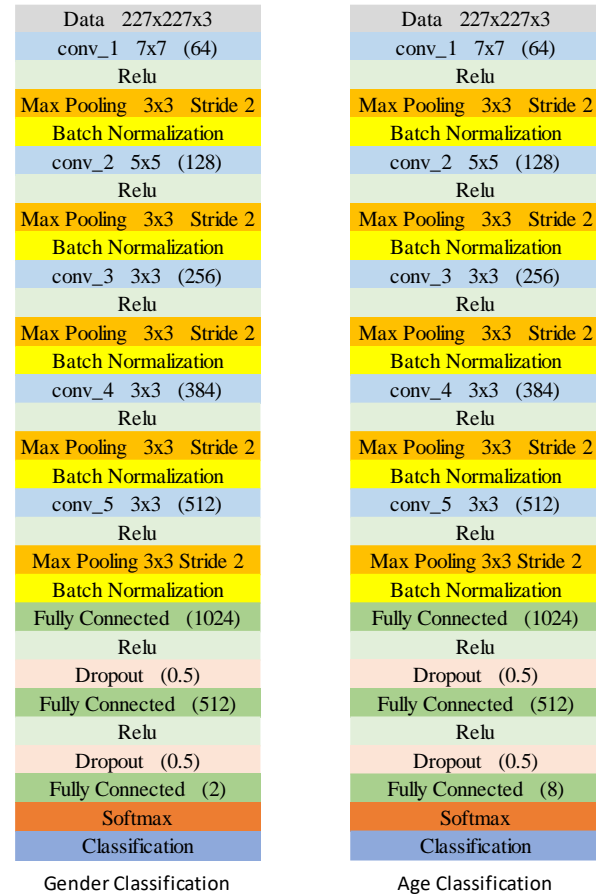


Figure 3. Proposed CNN network for gender and age classification

Pre-RoR architecture

The Pre-RoR network structure utilized in this investigation is described by Zhang et al. (2017) and it was constructed in MATLAB environment. Since the Pre-RoR architecture is not included in the ready-to-use networks library in MATLAB, this network architecture, shown in Figure 4, has been developed and used in the current study. There is a total of four levels in the proposed Pre-RoR network architecture.

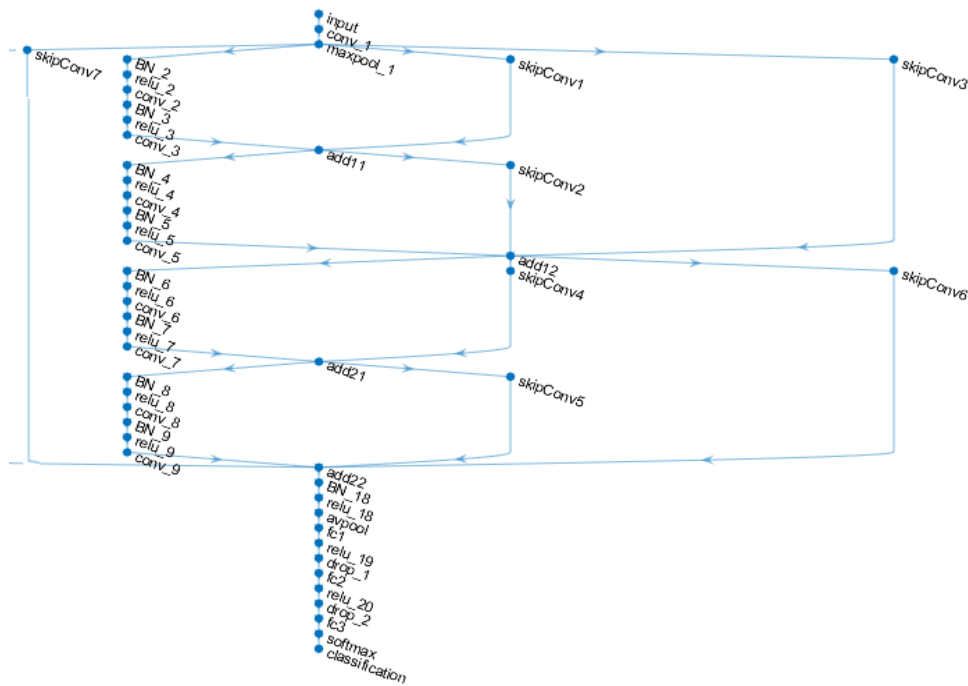


Figure 4. Pre-RoR architecture studied

In Figure 4, only two levels of proposed Pre-RoR network architecture are shown so that they can be clearly observed. There are two "conv" layers in each of levels. And four addition layers were created for the architecture. "conv_3" and "pool_1" were added to each other in the addition layer called *add11*. The layer called *add12* was the sum of the conv_5 and *add11* layer outputs, and similarly the *add21* layer was the sum of conv_7 and *add12* layer outputs. Finally, the outputs of layer conv_9, *maxpool_1*, *add12*, *add21* were added to the sum that was called *add22*.

The proposed RoR network structure is described layer by layer below.

- Image data in the size of $227 \times 227 \times 3$ was applied to the input layer of the network.
- The data, applied to the input layer, are processed by convolution process. The size and number of kernel filters used in this step were selected as 3×3 and 16, respectively. Among the hyper parameters specific to the convolution

process, stride and padding were determined as one and zero, respectively. Convolution output, the data size is $225 \times 225 \times 16$ at this point, is applied to the pooling layer. The size of the kernel filter applied here is 3×3 and the stride is 2. After the pooling, the size of the data becomes $112 \times 112 \times 16$. This point is called "pool_1". It will be used in the aggregation layer in the next step.

- Following the "pool_1" operation, batch normalization (BN) and rectified linear unit layers (ReLU) are applied. Afterwards, convolution process is used again for certain parameters (kernel size: 7×7 , number of filter: 16, stride: 2, padding: 0). This yields to an output image of $56 \times 56 \times 16$. This convolution layer is called as "conv_2".
- Similar layers of BN, ReLU and "conv_3" (kernel size: 7×7 , number of filters: 16, Stride 1, padding same) are applied to the output conv_2. Output of "conv_3" the data size is $56 \times 56 \times 16$.
- Then, "conv_3" and "pool_1" were added to each other in the addition layer called *add11*. In order to perform this addition, "conv_3" and "pool_1" must have same sizes. Since "Conv_3" size is $56 \times 56 \times 16$ and "pool_1" size is $112 \times 112 \times 16$, this addition process

cannot be performed directly. For this reason, the dimensions are equalized with the use of skip convolutions called as *skipconv* in MATLAB. This is actually a convolution operation and operationally, it is not different from other convolutions. By applying *skipconv* (kernel size: 1x1, number of filter: 16, stride: 2 and padding: 0) to the output of pool_1, the size is reduced to 56x56x16, and addition process can be applied.

These processes, which are explained in detail for first layer, are repeated in the following layers. The only difference is that the size of the applied filters and pooling layer are different.

Transfer learning

In this study, ResNet-101, ResNet-50, ResNet-18, ResNet-101-RoR and InceptionV3 network architectures are used for transfer learning. InceptionV3, ResNet-18, ResNet-50 and ResNet-101 had been trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network had been learned rich feature representations for a wide range of images.

In all networks used for transfer learning, the hyper parameters are set to the specific values described as follows. The number of epoch for age classification and gender classification were set as 50 and 100, respectively, and initial learning rate and mini-batch size were chosen as 0.0001 and 64, respectively. Besides these, the weight values of the first 5 layers were frozen.

After the training of the ResNet-101 model, the ResNet-101-RoR model was created by adding skip convolution layers to obtain different architecture from literature. With this model, the weight values, obtained

by pre-training in ResNet-101, were selected as initial weights and these parameters were adjusted for better performance. ResNet-101-RoR architecture generated in this study given in Figure 5.

Five different models were trained using InceptionV3, ResNet-18, ResNet-50, ResNet-101, ResNet-101-RoR architectures for transfer learning.

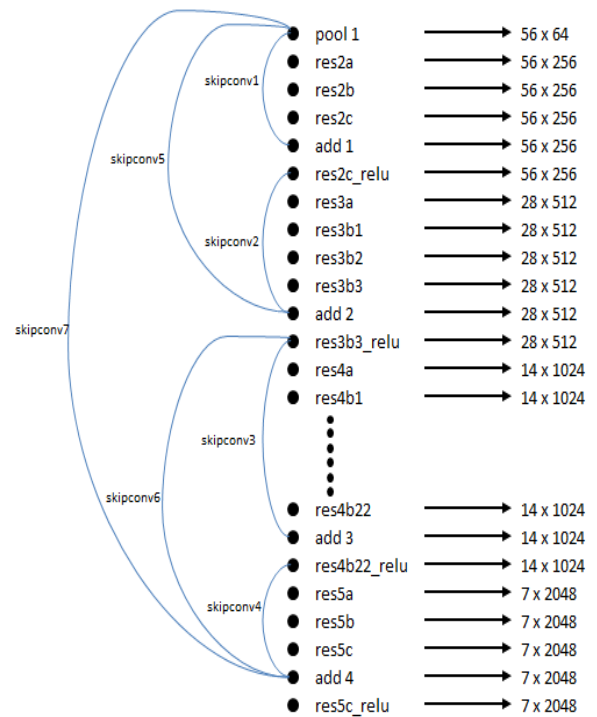


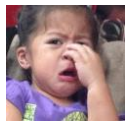

Figure 5. ResNet-101-RoR skip convolution connections

Classifier Fusion by Majority Voting

The results obtained from seven models, such as Convolutional Neural Network, ResNet-101, ResNet-50, ResNet-18, Pre-RoR, ResNet-101-RoR and InceptionV3, were fused by majority voting to compute the estimation result more accurately. The number of models were determined to be an odd number to avoid from an even decision. As an example, Table 1 denotes age class estimations performed by networks for two given sample pictures, and the corresponding majority voting result are given in the last

column. For the first row, five of the seven models estimated the sample in the 4-6 years old class. Hence, most models predict in the same class, and therefore, the final estimated age of the individual is decided as 4-6 years old class. Similarly, the sample in the second row is estimated in the 15-20 years-old class.

Table 1. Examples for majority voting decision

Images		
CNN	0-2	8-13
Resnet-101-RoR	4-6	25-32
Resnet-101	4-6	8-13
Resnet-50	4-6	15-20
InceptionV3	4-6	15-20
Pre-RoR	0-2	25-32
Resnet-18	4-6	15-20
Majority Voting	4-6	15-20

Dataset

In the study, Adience (Eidinger, Enbar, & Hassner, 2014) data set was used. In this dataset, it is aimed to make the sample data as close as possible to the challenges of real-life conditions. All pictures were automatically collected from smartphones and uploaded by Flickr without any prior filtering. Due to the lack of filtering, there are big differences in the pictures in terms of exposure, lighting conditions and image quality. The dataset samples consist of two groups, namely faces and aligned faces. There are total of 26850 photographs from 2284 people. There are 2 classes (female / male) for gender classification, 8 classes (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-50, 60-) for age classification. In the proposed study, data in the group of aligned faces were used for both age and gender classification. Images of

people labeled as unidentified gender were removed, and a total of 17492 images were used for gender classification. Likewise, 17523 images are used for age classification. The distribution of the data used for age and gender classification are given in Table 2 and Table 3. Moreover, these figures also show the number of data in each cross-validation fold.

Table 2. Five-fold cross-validation image distribution by gender

Gender	Fold0	Fold1	Fold2	Fold3	Fold4
Female	1948	1998	1774	1804	1848
Male	2047	1611	1363	1502	1597
Sum	3995	3609	3137	3306	3345

Table 3. Five-fold cross validation image distribution by age

Age	Fold0	Fold1	Fold2	Fold3	Fold4
0-2	960	84	810	151	483
4-6	494	480	358	238	570
8-13	213	600	475	497	340
15-20	152	525	270	468	227
25-32	1646	635	774	970	1056
38-43	555	485	276	522	507
48-53	219	146	120	104	241
60 -	139	156	202	118	257
Sum	4378	3111	3285	3068	3681

Performance measures

The confusion matrix is a summary of the results of predictions in a classification problem. A class can be predicted correctly or incorrectly as a member of another class, and these results are clearly shown in the confusion matrix.

For this reason, this study utilizes the confusion matrix as the performance criterion, similar to others in the literature. Five-fold cross validation was also utilized. The classification process was performed dividing

the data set into five folds, using four folds for training, one fold for testing and shifting the test fold through the all available folds. The final success rate was obtained averaging 5 tests (Refaeilzadeh et al. 2009). Both training and testing processes were carried out using the Matlab program on a computer with 128GB ram and 3 GeForce RTX 2080 Ti (12GB) graphic cards.

3. Research Findings

The comparison of the accuracy rate of majority voting along with that of other methods are given in Table 4. Among the individual models, the best accuracy performance in gender classification problem was obtained as 90.29% with ResNet-101. Also, InceptionV3 model showed the best accuracy performance with 55.28% for age classification. Using majority voting method as suggested in the proposed work, on the other hand, improves the over-all accuracy rate to 92,29 % for gender classification and 60.26% for age classification problems.

Table 4. Comparison of test results

Model	Gender Classification Accuracy	Age Classification Accuracy
CNN	% 85,93	% 52,39
ResNet-101	% 90,29	% 54,39
ResNet-50	% 89,12	% 55,16
ResNet-18	% 89,17	% 54,85
Pre-RoR	% 86,30	% 39,42
ResNet-101-RoR	% 88,76	% 54,44
InceptionV3	% 90,18	% 55,28
Proposed Method	%92,29	% 60,26

Confusion matrices of the fusion by majority voting are also given for detailed examination. They are shown in Figure 6 for gender and Figure 7 age classification experiments. Here, cross-descending cells represent accurately classified classes. The bottom row of the graphs shows the percentage of all samples of a particular class that are estimated correctly and incorrectly. The cell in the lower right corner of the chart, on the other hand, shows the overall accuracy.

Confusion Matrix (Gender Classification)

Predicted Class	female	8795 50.3 %	766 4.4 %	92.0 % 8 %
	male	577 3.3 %	7354 42.0 %	92.7 % 7.3 %
		93.8 % 6.2 %	90.6 % 9.4 %	92.3 % 7.7 %
		female	male	
		Real Class		

Figure 6. Confusion matrix (Gender Classification)

Confusion Matrix (Age Classification)

		0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-		
Predicted Class	0-2	1966 11,2%	287 1,6%	28 0,2%	2 0,0%	17 0,1%	7 0,0%	3 0,0%	1 0,0%	85,1%	
	4-6	485 2,8%	1460 8,3%	317 1,8%	17 0,1%	21 0,1%	3 0,0%	6 0,0%	0 0,0%	63,2%	
	8-13	22 0,1%	331 1,9%	1208 6,9%	114 0,7%	71 0,4%	12 0,1%	2 0,0%	2 0,0%	68,6%	
	15-50	2 0,0%	17 0,1%	268 1,5%	533 3,0%	540 3,1%	73 0,4%	18 0,1%	7 0,0%	36,6%	
	25-32	10 0,1%	40 0,2%	282 1,6%	921 5,3%	3873 22,1%	1162 6,6%	145 0,8%	46 0,3%	59,8%	
	38-43	3 0,0%	2 0,0%	15 0,1%	49 0,3%	536 3,1%	975 5,6%	414 2,4%	252 1,4%	43,4%	
	48-53	0 0,0%	2 0,0%	4 0,0%	2 0,0%	20 0,1%	67 0,4%	135 0,8%	155 0,9%	56,6%	
	60-	0 0,0%	1 0,0%	3 0,0%	4 0,0%	3 0,0%	46 0,3%	107 0,6%	409 2,3%	35,1%	
			79,0%	68,2%	56,8%	32,5%	76,2%	41,6%	16,3%	46,9%	60,3%
			21,0%	31,8%	43,2%	67,5%	23,8%	58,4%	83,7%	53,1%	39,7%
		Real Class									

Figure 7. Confusion matrix (Age Classification)

The results are also compared with the results given in the literature as shown in Table 5. It is clear that the proposed majority voting method gives the highest accuracy rates among all except Zhang et al. (2017) They are pre-trained dataset samples and fine-tuned their network, and therefore, obtained the highest rates. Although their work is

considered in a different method, the accuracy rates of the proposed study, without pre-training and fine-tuning, is promisingly very close to their scores.

Table 5. Success rates of previous studies in the literature

Model	Gender Classification Accuracy	Age Classification Accuracy
Levi and Hassner	%86,8	% 50,7
Wolfshaar et al.	%86,2	No study
Ekmekji	%80,8	% 50,2
Rodríguez et al.	%92,4	% 57,8
Hosseini et al.	%88,9	% 61,3
Akbulut vd.	%87,13	No study
Zhang et al.	%93,24	% 67,34
Proposed Method	%92,29	% 60,26

4. Results

In this study, age and gender classification are performed by using various models of convolutional neural networks. Since the images in the data set are natural in the form of unfiltered images taken from real life, the accuracy of the test data is very close to real life conditions. Tests show that mini-batch parameter selection should be performed based on the network model in order to achieve the best results. Keeping the mini-batch size as small as possible ensures that packets are processed very slowly in the course of learning. Although this is good for training, selecting the mini-batch size too small causes over-fitting. Also, small sized mini-batches negatively affects the training time. Performing transfer learning with the use of pre-trained ready networks ensures that

required long training times can be shortened with good classification performance. Furthermore, it is better to work on datasets with classes similar to the ones used in the pre-trained network in order to get the best out of transfer learning. In that case, all weights can be frozen as performed in the proposed study. Also, the use of too deep network does not necessarily yield to a very high performance. Age classification problem reveals such an example. Although the ResNet-101 model is deeper than the ResNet-50 model, the average success rate of ResNet-50 model is better than the other. In majority vote method, pictures which are incorrectly predicted by some models can be corrected with the aid of some other models. Thus, the number of incorrectly predicted images is reduced. When the results are compared, the success rate of the majority method is higher than the success rates in all tests.

5. References

- Sharma, A. (tarih yok). *Confusion Matrix in Machine Learning*. from GeeksforGeeks: <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>
- Agbo-Ajala, O., & Viriri, S. (2020). Deeply Learned Classifiers for Age and Gender Predictions of Unfiltered Faces. *Hindawi The Scientific World Journal*.
- Akbulut, Y., Şengür, A., & Ekici, S. (2017). Gender recognition from face images with deep learning. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*.
- Eidinger , E., Enbar , R., & Hassner, T. (2014). Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security*, 2170 - 2179.

- Ekmekji, A. (2016). Convolutional Neural Networks for Age and Gender Classification. *Stanford University*.
- Hosseini, S., Lee, S. H., Kwon, H. J., Koo, H. I., & Cho, N. I. (2018). Age and Gender Classification Using Wide Convolutional Neural Network and Gabor Filter. *2018 International Workshop on Advanced Image Technology*. Chiang Mai: IEEE.
- Lapuschkin, S., Binder, A., Müller, K. R., & Samek, W. (2017). Understanding and Comparing Deep Neural Networks for Age and Gender Classification. *ICCV'17 Workshop on Analysis and Modeling of Faces and Gestures*.
- Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
- Liu, X., Li, J., Pan, J.-S., & Hu, C. (2017). Deep convolutional neural networks-based age and gender classification with facial images. *2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS)*.
- Mingxing, D., Li, K., Yang, C., & Li, K. (2018). A hybrid deep learning CNN-ELM for age and gender classification. *Neurocomputing*, 448-461.
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. *Encyclopedia of Database Systems* (p. 532-538).
- Rodríguez, P., Cucurull, G., Gonfaus, J., Roca, F., & González, J. (2017). Age and gender recognition in the wild with deep attention. *Pattern Recognition*, 563-571.
- Wolfshaar, J., Karaaba, M., & Wiering, M. (2015). Deep Convolutional Neural Networks and Support Vector Machines for Gender Recognition. *2015 IEEE Symposium Series on Computational Intelligence*. Cape Town.
- Zhang, K., Sun, M., Han, T., Yuan, X., Guo, L., & Liu, T. (2016). Residual Networks of Residual Networks: Multilevel Residual Networks. *IEEE TRANSACTIONS ON LATEX CLASS FILES*.
- Zhang, K., Gao, C., Guo, L., Sun, M., Yuan, X., Han, T., Li, B. (2017). Age Group and Gender Estimation in the Wild With Deep RoR Architecture. *IEEE Access*, 22492 - 22503.
- Tivive, F., & Bouzerdoum, A. (2006). A Gender Recognition System using Shunting Inhibitory Convolutional Neural Networks. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 5336-5341.