IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 4, August, 2023, pp. 1629-1639
Journal homepage: http://iieta.org/journals/ts

Deep Learning Based Gender Identification Using Ear Images

Şafak Kılıç*, Yahya Doğan

[1] Department of Software Engineering, Kayseri University, Kayseri 38100, Turkey
[2] Department of Computer Engineering, Siirt University, Siirt 56100, Turkey

Corresponding Author Email: safakkilic@kayseri.edu.tr

This article is part of the Special Issue Advances of Machine Learning and Deep Learning

https://doi.org/10.18280/ts.400431

Received: 5 February 2023
Revised: 31 July 2023
Accepted: 11 August 2023
Available online: 31 August 2023

Keywords:
deep learning, gender identification, ear images, convolutional neural network (CNN), biometric identification, image processing, machine learning, facial recognition

ABSTRACT

The classification of an individual as male or female is a significant issue with several practical implications. In recent years, automatic gender identification has garnered considerable interest because of its potential applications in e-commerce and the accumulation of demographic data. Recent observations indicate that models based on deep learning have attained remarkable success in a variety of problem domains. In this study, our aim is to establish an end-to-end model that capitalizes on the strengths of competing convolutional neural network (CNN) and vision transformer (ViT) models. To accomplish this, we propose a novel approach that combines the MobileNetV2 model, which is recognized for having fewer parameters than other CNN models, with the ViT model. Through rigorous evaluations, we have compared our proposed model with other recent studies using the accuracy metric. Our model attained state-of-the-art performance with a remarkable score of 96.66% on the EarVN1.0 dataset, yielding impressive results. In addition, we provide t-SNE results that demonstrate our model's superior learning representation. Notably, the results show a more effective disentanglement of classes.

1. INTRODUCTION

Biometric identification technology is rapidly advancing, driven by its commendable security and reliability. As a consequence, it is finding widespread applications in various domains, including e-commerce, e-government, and crime detection. Biometric identification-based verification systems, such as finger-print recognition and facial identity verification systems, are constantly evolving [1, 2]. In recent years, automatic gender determination has attracted considerable interest due to its potential in a variety of applications, including human-computer interaction, banking transactions, disease diagnosis, visual surveillance, and demographic data collection. Gender determination is essential to biometric identification systems because it permits the database to be cut in half, thereby simplifying and expediting the identification process. This increases the efficacy and efficiency of these systems in their respective applications.

Ear recognition technology has emerged as a significant non-invasive personal identification method with numerous applications, comparable to facial recognition. The human ear has stable, well-structured characteristics that are unaffected by facial expressions and aging. Notably, the earlobe, a distinguishing characteristic utilized in forensic investigations, continues to change over time. The ear's visibility in images recorded by security cameras and its ease of capture in profile views, video recordings, or photographs increase its identification utility [3].

Due to its unique and relatively stable characteristics, the ear functions as a valuable biometric identifier in both biometric research and forensic medicine [4]. In contrast to the face, which can be affected by factors such as facial expressions, facial hair, and cosmetics, the ear's appearance is stable, making it an advantageous identifier [5]. Particularly, the earlobe stands out as a distinguishing characteristic frequently used in forensic investigations, and it continues to change over time, providing additional identification clues [6]. The ability of security cameras to capture the ear, whether it is partially or completely visible, aids in the identification process [4]. Moreover, the ear is simpler to capture than the face in profile views, video recordings, and photographs [7]. In Figure 1, the fundamental components of the human ear are provided.

Indeed, numerous studies have concentrated on the use of ear images for identification purposes [3-9]. In contrast, the research on extracting sensitive biometric features from ear images, such as age, has been relatively limited. Despite this, there has been considerable interest in using ear images for gender classification, as evidenced by several studies [10-13]. Previous research has investigated the use of ear images for gender classification, laying the groundwork for our research. In the study conducted by Gnanasivam and Muttan [10], they utilized ear hole measurements as a reference point and calculated Euclidean distances between the detected ear hole from masked ear images and seven distinct ear features. They used the Bayes classifier, the KNN classifier, and neural networks as classifiers. The KNN classifier yielded the best results, achieving a remarkable classification accuracy of 90.42%. Zhang and Wang [11] investigated gender classification using profile face images and ear images separately. They employed support vector machines (SVM) with a histogram intersection kernel for classification. By

performing score-level fusion based on Bayesian analysis, they were able to improve accuracy. In the study conducted by Khorsandi and Abdel-Mottaleb [12], they employed Gabor filters for feature extraction and performed classification based on features extracted by dictionary learning. This approach resulted in an accuracy of 93.5%. The study utilized 2D images from the UND biometrics dataset collection F [14]. The fusion approach achieved an impressive accuracy of 97.65%, whereas the accuracy for face-only classification was 95.43% and ear-only classification was approximately 91.7%.

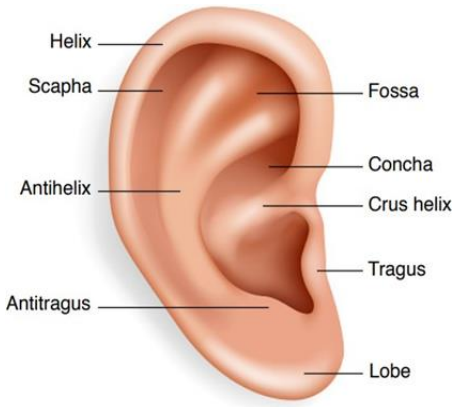

Figure 1. Morphological components of the human ear [6]

Despite the fact that these studies have demonstrated a high rate of accuracy for gender classification based on ear images, they have primarily relied on conventional machine-learning techniques. Our research extends beyond these prior works by proposing a novel hybrid recognition architecture, MobileNetV2 with ViT, for gender classification tasks. Unlike the traditional methods employed in prior studies, our model leverages deep learning techniques, particularly CNN and ViT, which have shown success in computer vision tasks. The combination of these models provides superior performance, computational efficiency, and reduced parameter count compared to individual approaches. Moreover, we provide t-SNE results to assess how well our proposed model separates gender-specific characteristics in a low-dimensional space. This analysis provides helpful insights into the interpretability and feature representations of the model, which can be used to better comprehend its decision-making process.

The main contributions of this work can be summarized as follows:

- We propose a hybrid recognition architecture called MobileNetV2 with ViT, which combines the strengths of CNN and ViT, to perform gender classification tasks.
- Recently, the efficacy of models like MLP-Mixer and ViT, which are alternatives to CNNs, has been evaluated. In terms of performance, computational efficiency, and the number of parameters, a hybrid model architecture that incorporates the benefits of these models has been identified as the best option.
- We provided t-SNE results to observe how well the proposed model disentangles features compared to other models in a low-dimensional space.

The rest of the article is structured as follows: Section 2 provides a brief overview of relevant prior research. Section 3 delves into the methodology and materials used in the study. The comparison results of the methods are presented in Section 4. Finally, the article concludes with potential future directions.

2. RELATED WORKS

Ear recognition has garnered a great deal of interest, with applications in fields such as security, surveillance, and forensic science. The uniqueness and stability of ear characteristics make it a valuable biometric for individual identification. One of the crucial steps in ear recognition is the extraction of relevant features from ear images. The ability to distinguish accurately between different ears is significantly dependent on the efficiency of this feature extraction process. Nevertheless, it is frequently regarded as one of the most difficult aspects of ear-based identification systems.

In recent years, the field of ear recognition has witnessed a shift from traditional handcrafted methods to deep learning-based approaches, mainly due to their superior performance in various recognition tasks. Multiple studies have explored the potential of deep learning in ear recognition, each proposing novel architectures and techniques to achieve accurate and robust results. Dodge et al. [15] introduced a deep learning-based ear recognition system that utilized CNNs and transfer learning for feature extraction. The extracted features were then fed to a shallow classifier for identification. Alshazly et al. [16] proposed a combination scheme for an ensemble of deep learning models. The study compared models trained with random weights, pre-trained models, and fine-tuned pre-trained models. In their evaluation, the efficacy of finely-tuned models was deemed to be superior. Ahila Priyadharshini et al. [17] developed a simple CNN architecture for ear recognition and evaluated its performance on ear images obtained under both controlled and uncontrolled environmental conditions. Khaldi and Benzaoui [18] introduced a new framework for ear recognition using generative adversarial networks in unconstrained conditions, showcasing the versatility of deep learning methods in various scenarios. Additionally, the authors proposed a deep unsupervised active learning-based ear recognition system [19], which was tested in both controlled and uncontrolled conditions, further demonstrating the adaptability of deep learning techniques to diverse environments. Mewada et al. [20] proposed a spectral-spatial feature based on CNN for describing ear images and an embedding algorithm for fusing multilevel spectral information from the CNN network. The performance of the proposed system was evaluated on ear images captured under uncontrolled conditions.

The selection and evaluation of features play a vital role in ear recognition, but this process remains difficult. To address this issue, the study by Omara et al. [21] proposed a novel method for extracting features using CNN models. For classification, they then utilized the large margin distance learning metric (LDMLT) learning algorithm to calculate the Mahalanobis distance based on KNN. This method aimed to improve the accuracy and efficiency of ear recognition. However, one of the major limitations of deep learning techniques in ear recognition is the need for large amounts of data and the time required for models to acquire meaningful ear features. To overcome these limitations, Korichi et al. [22] proposed a computationally efficient and straightforward deep neural network model called TR-ICANet for ear recognition. Despite its simplicity, the TR-ICANet achieved an accuracy of 51.25% when tested on the AWE (Audio-Visual Event) dataset. In the study by Emeršič et al. [23], a CNN-based pipeline was proposed, showcasing its effectiveness in both ear detection and recognition tasks. The results obtained from the AWE and UERC ear databases demonstrated high

accuracy, with 99.8% for ear detection and 92.6% for ear recognition. Similarly, in the studies by Alshazly et al. [24] and Radhika et al. [25], various deep learning algorithms and models were evaluated to optimize the accuracy of ear recognition. Alshazly et al. [24] employed an ensemble of ResNeXt101 models, achieving improved performance in ear recognition. Furthermore, Alshazly et al. [26] conducted further analysis by exploring different datasets, contributing to the ongoing efforts to enhance ear recognition performance using deep learning approaches.

In conclusion, recent research in the field of ear recognition has shown promising results in using ear images for gender recognition tasks with high accuracy rates. In this study, a hybrid model has been proposed by combining state-of-the-art deep learning architectures to enhance the performance of ear gender recognition problem.

3. MATERIAL AND METHOD

3.1 Dataset

In the field of ear recognition, datasets collected under unrestricted conditions are scarce. These datasets differ in terms of ear morphology, the number of individuals included, and the data collection techniques. One of the earliest datasets employed for ear recognition research is the WPUT dataset [27], which consists of 2071 ear images from 501 individuals of varying ages and includes variables such as illumination, head position, and occlusions. The AWE dataset [3] comprises 1000 ear images of 100 celebrities collected from the internet, whereas the UERC dataset [3] is an extension of AWE, containing 11,804 ear images from various individuals. The In-the-wild ear dataset [28] comprised of 2,058 ear images cropped from a larger dataset originally intended for face recognition, containing data from 231 individuals. On the other hand, the EarVN1.0 dataset [29] is one of the largest public ear datasets, containing 28,412 RGB ear images from 164 Asian individuals. This dataset was constructed by cropping Internet images of ears, capturing various camera and lighting conditions. The presence of pose, scale, and illumination variations in these datasets makes them suitable for building models adaptable to real-life scenarios; however, it also presents training challenges. We utilized the EarVN1.0 dataset for our research, which includes gender information. Figure 2 depicts example images of the ear from the EarVN1.0 dataset.

Figure 2. Sample ear images taken EarVN1.0 dataset; top: left ear images; down: right ear images for the same person. The images' resolutions vary because they were captured under unconstrained conditions

3.2 Method

In this section, we aim to devise an architecture for the ear gender recognition problem that achieves both low-parameter complexity and state-of-the-art performance. To accomplish this, we propose a hybrid architecture that incorporates three ground-breaking deep learning techniques: CNN, MLP-Mixer, and ViT, as well as their derivatives.

3.2.1 Convolutional neural networks

Within the domain of CNN, the central operation is convolution. This process involves applying a *KxK* sized convolution kernel to an input image with dimensions *HxW*, where H and W represent the height and width of the feature map, respectively. The input image consists of N channels, and the terms M and N denote the number of convolution kernels and output feature-map channels, respectively. Figure 3 illustrates this process. As shown, standard convolution entails convolving the input data with multiple convolution kernels of the same depth, and the final result is computed by summing up the results corresponding to each channel. In recent times, CNN-based models that have demonstrated exceptional performance in the Large Scale Visual Recognition Challenge (ILSVRC) [30] have also proven to be effective in various other problem domains. Several examples of these models are described below.

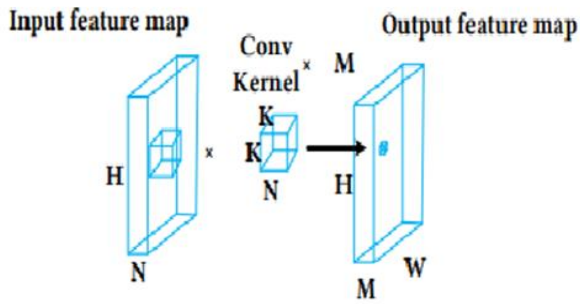

Figure 3. The overall process of standard convolution [8]

3.2.2 AlexNet

The AlexNet model, proposed by Krizhevsky et al. [31], achieved significantly improved recognition accuracy on the ImageNet dataset [32], resulting in its victory in the ILSVRC-2012 competition. This model, which consists of eight layers, has demonstrated unique advantages in image classification tasks [33]. The data source requires input in the format of 227×227×3 pixels, where the dimensions 227×227 represent the height and width of the input image, and the value 3 indicates that the data is in RGB mode with three channels. The first two layers of the model perform convolution, followed by activation (using ReLU), max-pooling, and normalization operations. Subsequently, the output of the second layer undergoes convolution with 256 feature maps using a kernel size of 5×5, a stride of 1, and other parameters matching the first layer. The third and fourth layers only perform convolution and activation operations, while the fifth layer combines convolution, activation, and normalization operations. The output of the fifth layer is then reshaped and flattened into a long vector, which is fed into a traditional neural network comprising three fully connected layers. The first two fully connected layers contain 4,096 neurons each, and the final layer has output nodes corresponding to the number of classes, i.e., 1,000.

3.2.3 VGGNet

Researchers from the Oxford University Visual Geometry Group and Google DeepMind devised VGGNet, a deep CNN consisting of six models with varying depths, ranging from 11 to 19 layers [34]. The models with the greatest number of layers, namely 16 and 19, have proven to be the most effective for image classification and localization tasks. The architecture of VGGNet is based on five convolutional layers, each utilizing a kernel size of 3×3, a stride of 1, and a padding of 1, followed by a max-pooling layer with a size of 2×2 and a stride of 2. Subsequent to the final maxpooling layer, the features within the image feature map are integrated through three fully connected layers, with the final layer employing Softmax for image classification and normalization. In comparison to traditional CNNs, VGGNet's significant contributions include a smaller size of convolution and pooling kernels, an increased number of convolutional layers, the use of pre-trained data for parameter initialization [35], and a method for converting fully connected layers into convolutional layers during the testing phase.

3.2.4 Inception

InceptionV1, also known as GoogLeNet, is a deep convolutional neural network architecture proposed by Szegedy et al. [36]. It achieved a remarkable accuracy of 93.3% in the ILSVRC competition and stood out for its significantly reduced number of parameters compared to earlier models like AlexNet and VGG.

The unique feature of this architecture is its departure from the conventional sequential process. Instead, it employs a combination of network layers, pooling layers, and parallel computations of both large and small convolutional layers. It also utilizes 1×1 convolutions for dimensionality reduction. This approach of parallelism and dimensionality reduction significantly reduces the number of parameters and computational cost, making it more memory and computation efficient. Various versions of Inception exist, including InceptionV1/GoogLeNet, InceptionV2, and InceptionV3, among others.

3.2.5 ResNet

ResNet, a deep learning architecture proposed by He et al. [37], addresses the challenges of training deep neural networks, which include high computational costs and limitations on the number of layers. ResNet tackles these issues by introducing skip-connections or shortcuts. Unlike other architectural models, ResNet's performance does not degrade as the depth of the architecture increases, and it significantly improves computational efficiency, enabling better training of networks. The ResNet model incorporates skip-connections, ReLU, and batch normalization in its architectures, typically between two to three layers. The exceptional image classification performance of ResNet, as demonstrated by He and his colleagues [37], highlights its efficacy in extracting image features.

3.2.6 DenseNet

DenseNet, introduced by Huang et al. [38], is a deep learning architecture that employs direct connections between all layers, facilitating an efficient flow of information. Each layer in the DenseNet architecture receives input from all preceding layers and shares its feature maps with all subsequent layers. The feature maps generated by a given layer are concatenated with those from the preceding layer, leading to a design known as DenseNet.

3.2.7 MobileNet

MobileNet is a deep learning architecture designed to improve accuracy by minimizing the number of convolutional layers. This reduction can however cause the issue of gradient vanishing. MobileNetV2 was developed as an enhancement over MobileNetV1 to address this issue. MobileNetV2 incorporates the residual structure from ResNet to enable better information flow between layers and mitigate gradient vanishing during backward propagation. MobileNetV2's fundamental building block is a depthwise separable convolution block with linear bottleneck and inverted residuals, which transforms features from N to M channels. The bottleneck consists of a 1×1 convolutional layer with a linear activation function, followed by a depthwise convolutional layer with subsampling using the s parameter. The network structure of MobileNetV2 consists of 19 layers and is depicted in Table 1, where conv2d represents standard convolution, avgpool represents average pooling, c denotes the number of output channels, and n signifies the number of repetitions. The intermediate layers are in charge of feature extraction, whereas the final layer is responsible for classification.

Table 1. The overall network structure of MobileNetV2, where 'k' signifies the number of classes

Input Shape	Operator	t	c	n	s
224*224*3	conv2d	-	32	1	2
112*112*32	bottleneck	1	16	1	1
112*112*16	bottleneck	6	24	2	2
56*56*24	bottleneck	6	32	3	2
28*28*32	bottleneck	6	64	4	2
14*14*64	bottleneck	6	96	3	1
14*14*96	bottleneck	6	160	3	2
7*7*160	bottleneck	6	320	1	2
7*7*320	conv2d 1×1	-	120	1	1
7*7*1280	avgpool 7×7	-	-	1	-
1*1*320	conv2d 1×1	-	k	-	

3.3 MLP-mixer

MLP-mixer [39]: With the help of state-of-the-art models, MLP-mixer offers a fairly straightforward architecture that performs competitively on benchmarks for image clasification. The MLP-mixer is focused only on multi-layer perceptrons (MLPs) that do not employ convolutions or self-attention. The Mixer layer is made up of two separate MLP layers: one that applies MLPs to individual picture patches independently, "mixing" location-specific characteristics and another that applies MLPs to many patches simultaneously, "mixing" spatial information. Only simple matrix multiplication operations, data layout adjustments (reshapes and transpositions), and scalar nonlinearity are used by Mixer.

MLP is a type of neural network architecture that process input data through various layers and produce output data as a result. The architecture of an MLP includes an input layer, several hidden layers, and an output layer. The hidden layers process the data and learn higher level features. The output layer perceptrons then produce the output data. An MLP includes weight and bias values for each perceptron. These values are learned during training and used to process the data. The weight values represent the connections between perceptrons and the bias values represent the threshold values of the perceptrons. This architecture is typically feed-forward, meaning that the data flows from input to output. It is also typically trained using the back-propagation algorithm, which updates the weight values to reduce the error rate between the output data and the real data. MLP is a powerful architecture that can be used for a variety of applications such as natural language processing, image recognition and audio recognition. With its ability to process data through multiple layers and learn advanced features, it has proven to be a valuable tool in the field of machine learning. An MLP can be represented mathematically using the following equation:

$$y = f\,(Wn * f\,(Wn-1 *\ldots* f\,(W2 * f\,(W1 * x + b1) + b2)\ldots + bn - 1) + bn) \quad (1)$$

The equation, where y is the output of the Multi-layer Perceptron (MLP), x is the input data, f is the activation function (e.g., sigmoid, ReLU, etc.), *W1, W2, ...*, *Wn* are the weight matrices for each layer, *b1, b2, ..., bn* are the bias vectors for each layer and n is the number of layers in the MLP, describes how the input data is transformed through multiple layers of perceptrons. The input data is multiplied by the weight matrix of the first layer, passed through the activation function to introduce non-linearity, and then added to the bias vector of the first layer. This process is repeated for each subsequent layer, with the output of one layer serving as the input for the next layer. The final output is obtained by passing the output of the final hidden layer through the activation function and adding the bias vector of the output layer. It should be noted that the above equation depicts a feed-forward process, where information flows from the input layer to the output layer without feedback. The back-propagation algorithm can be utilized to adjust the weight and bias values during training to minimize the error between the predicted output and the actual output.

The MLP-mixer is a method that uses a sequence of linearly projected image patches, referred to as tokens, as input. The input data is arranged in a table with dimensions "patches x channels". The mixer has two types of MLP layers to mix spatial information: channel-mixing MLPs and token-mixing MLPs. Channel-mixing MLPs allow communication between different channels by processing each token independently and using individual rows of the input table. Token-mixing MLPs facilitate communication between different spatial locations by independently processing each channel and using individual columns of the input table. By alternating between these two types of layers, the MLP-Mixer is able to effectively mix both input dimensions, resulting in mixed feature maps.

The MLP-mixer takes as input a sequence of S non-overlapping image patches, each projected to a hidden dimension of C. This produces a two-dimensional input table, X; ϵ; RSXC. The mixer is made up of multiple layers, each of the same size, and each layer consists of two blocks of MLPs. The first block is the token-mixing MLP, which operates on the columns of XT and maps the input to the same dimension as the output. The second block is the channel-mixing MLP, which operates on the rows of X and maps the input to the same dimension as the output. Each MLP block has two fully-connected layers and a nonlinearity operation. The nonlinearity is applied to each row of the input data tensor individually. In Table 2, we have provided the specifications used in the MLP-mixer architecture.

Table 2. Specifications of the mixer architectures

Specification	Value
Number of layers	6
Patch size	4
Number of channel	128
Hidden size	128
Output neuron size	100

3.4 Vision transformer (ViT)

Vision Transformers, initially introduced by Dosovitskiy et al. [40], are a deep learning architecture that has demonstrated superior performance in image classification applications when trained on large-scale datasets compared to CNNs. However, their reliance on vast quantities of training data and computational resources poses a challenge. To address this issue, Touvron et al. [41] introduced a data-efficient version of ViT by using commonly used data augmentation and manipulation techniques for CNNs. They also improved the performance of ViT through a transformer-based teacher-student approach. The high performance of ViT has prompted further research into its use for various vision tasks [42].

ViT is a deep learning architecture designed to categorize images by modeling a series of image patches into a semantic label. Unlike traditional CNN designs, the ViT utilizes the encoder module of the transformer to allow for the interpretation of information throughout the entire image through its attention mechanism. The architecture of the ViT typically includes (1) an embedding layer, (2) an encoder, and (3) a final classifier head.

The first step in the process is to divide the training set images into non-overlapping patches. Each patch is evaluated as a separate token by the transformer. A [*c, h, w*] dimensional image results in *n* series of [*c, p, p*] patches, where c represents the number of channels, h represents the height, *w* represents the width, and p is the size of the patch. The number of patches, *n*, is calculated by dividing *h w* by p^2. Typically, *a* patch size of 16 or 32 is chosen, as a smaller patch size results in a wider array and vice versa.

3.4.1 Embedding layer

The patches obtained from dividing the image into non-overlapping sections are transformed into a 1-dimensional vector via a trainable linear projection (embedded matrix E) prior to being fed into the encoder. The embedded patches are then combined with a learnable embedding classification indicator, x_{class}, necessary for performing the classification task.

The position of each patch within the image is incorporated into its representation through the addition of position embeddings. The position embeddings, denoted by E_{pos}, have a dimension of $(n+1)\times D$, where n is the number of patches and D is the dimension of the vector representation. The combined representation of each embedded patch and its position embedding is represented by z_0 in Eq. (2), with E being the embedding matrix of size (p_2c) and Epos$\in R^{(n+1)\times D}$.

$$z_0 = \left[x_{class}\, x_p^1 E_i x_p^2 E_i \ldots; x_p^n E\right] + E_{pos} \quad (2)$$

3.4.2 Vision transformer encoder

The encoder in ViT is constructed from L identical layers, each layer comprised of a Multi-Head Self-Attention (MSA) block and a Multi-Layer Perceptron (MLP) block. These blocks are separated by a normalization layer, and there are skip connections present after each block. The MLP block is composed of two layers utilizing the GELU activation function.

$$z_l' = MSA\left(LN(z_{l-1})\right) + z_{l-1}, l = 1 \ldots L \quad (3)$$

$$z_l = MLPLNz_l' + z_k' l = 1 \ldots L \quad (4)$$

$$y = LN_L^0 z \quad (5)$$

The first component in the sequence is taken from the last layer of the encoder and fed to the head classifier as shown in (5) to predict the class label.

3.4.3 Compact convolutional transformer (CCT)

ViT architectures have been demonstrated to achieve superior performance in image classification tasks when trained on large-scale datasets, however, they come with the requirement for significant amounts of data and computational resources [38]. To overcome this challenge, researchers have explored the combination of transformers and convolutions to benefit from their respective strengths. Hassani et al. [43] introduced the Compact Convolutional Transformer (CCT) architecture, which combines a patch-based approach to preserve local information, addressing some of the limitations of ViT, while still achieving improved performance. This model can encode the relationships between patches differently from the original ViT. CCTs are efficient to kenizers that preserve local spatial relationships while having short receptive fields. Additionally, the transformer encoder provides a sequential pooling approach called SeqPool which gathers sequential information from the encoder. SeqPool eliminates the need for an additional Classification Symbol. Figure 4 shows the entire end-to-end architecture of the model.

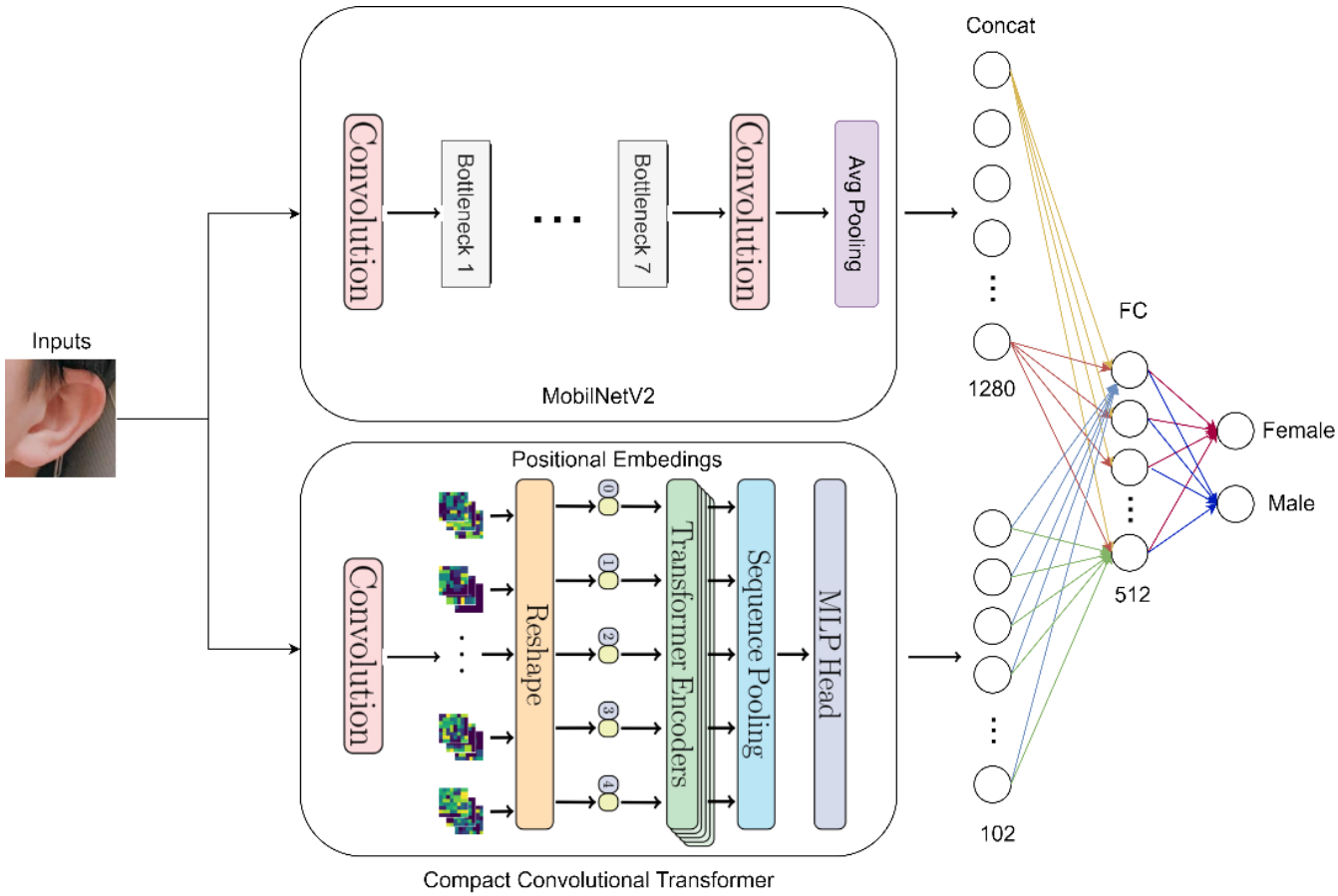

Figure 4. The proposed method for ear gender recognition

4. PROPOSED METHOD

In this section, we introduce an end-to-end approach for ear gender recognition. Our proposed method seeks to develop a hybrid model that is low-parameter, trainable quickly, and state-of-the-art. In order to achieve this, we have developed a hybrid architecture that takes into account two approaches that are presently competing in the image classification space: CNNs and ViT. Previous research has demonstrated that the VGG19 CNN model obtained a high level of performance for the ear recognition problem, achieving 91.2% accuracy [8]. However, this model had too many parameters (see Table 3) and thus did not meet our objective of developing a model with few parameters. As a result, we used the CNN model MobileNetV2, which has low accuracy but a limited number of parameters (see Table 3). For the ViT part, we used one of the vision transformer's derivatives, i.e., CCT. CCT architecture introduces compact transformers by incorporating convolutions in place of patching and utilizing sequence pooling. This reduces the number of parameters while maintaining high accuracy. Figure 4 shows our proposed hybrid architecture. In this architecture, the last layer of MobileNetV2 is removed and its feature vector of 1280 is concatenated with the feature vector of CCT model. Then, an additional fully connected layer is added to blend the feature vectors from both models and attach them to the class layer. Our proposed hybrid model takes advantage of the strengths of CNN and CCT to achieve high performance for the ear gender recognition problem.

Table 3. Parameter size of CNN-based models

Model Name	Parameter Size
AlexNet	45.342.082
VGG16	134.268.738
VGG19	137.218.626
InceptionV3	21.806.882
ResNet50	23.538.690
ResNet101	42.540.482
DenseNet201	17.721.960
MobileNetV2	2.243.490

Table 4. Hyperparameters of the proposed model

Batch size	16
Learning rate	0.001
Input_size	224×224×3
Optimizer	RMSProp
Loss function	cross entropy loss
Momentum	0.9
Epoch	200
Epsilon	1e-08
Learning rate decay factor	0.5

4.1 Training details

We utilize the RMSProp optimizer [44] with a batch size of 16 and train all models from scratch for 200 epochs, setting the learning rate to 0.001, rho to 0.9, and epsilon to 1e 08. When the minimum validation loss stops improving after 2 epochs and the best model is saved using model checkpoint monitoring validation loss, we cut the learning rate by a factor of 0.5%. The cross entropy loss function was utilized to update the model weights during training. The hyper parameters used in the training of the overall model are displayed in Table 4.

5. EXPERIMENT RESULTS

In this section, we discuss the steps we took to attain state-of-the-art results for the problem of ear gender recognition. Our objective is to develop a hybrid model by combining three approaches that have recently made significant advancements in image classification: CNNs, MLP-Mixer, and ViT. In this context, upon examination of the literature, no studies have been observed that used MLP-Mixer and ViT models, while many models such as VGG, ResNet, etc., that performed well on the ImageNet dataset have been used with CNN models. Nguyen-Quoc and Hoang [8] examined the performance of popular CNN models on the EarVN1.0 dataset, and the highest test accuracy of 91.12% was obtained with the VGG19 model. However, the VGG19 model is a dense network and contains approximately 137 million parameters (see Table 3). In contrast, the MobileNetV2 model has a modest number of parameters, approximately 2 million, but a performance of 85.55%. Comparing the training periods of the two models reveals that MobileNetV2 is trained approximately three times faster (see Table 5). Since our aim is to construct a model with few parameters, the MobileNetV2 model was chosen as the CNN model, and experimental studies were conducted to improve its performance.

Table 5. Comparison of model results in terms of test accuracy, run time, and parameter size

Model	Test Acc (%)	Run-Time	Parameter Size
MobileNetV2	85.55	6h 32min	2.243.490
VGG19	91.12	18h 48min	137.218.626
MLP-Mixer	88.45	9h 11min	16.171.778
ViT	90.49	6h 32min	4.573.941
Our MLP-Mixer with ViT	91.80	10h 45min	20.968.845
Our MobileNetV2 with MLP-Mixer	96.34	10h 57min	18.871.618
Our MobileNetV2 with ViT	96.66	9h 7min	7.506.729

In Table 5, we compare the accuracy, run-time, and parameter size of various models. Initially, using the MLP-Mixer model on the EarVN1.0 dataset and training it from scratch yielded a test accuracy of 88.45%. This model performed better than MobileNetV2 but was behind VGG19 in terms of accuracy. In terms of parameter size, it was roughly eight times larger than MobileNetV2 and 0.12 times smaller han VGG19 (see Table 5). Similarly, by training from scratch using the ViT model, a test accuracy of 90.49% was achieved. This model was found to be comparable to VGG19 and MobileNetV2 in terms of efficacy and parameter size, respectively. In addition, the performance of hybrid models with MobileNetV2, MLP-Mixer, and ViT architectures was investigated. As seen in Table 5, test accuracy of 91.80%, 96.34% and 96.66% was obtained by using MLP-Mixer with ViT, MobileNetV2 with MLP-Mixer and MobileNetV2 with ViT respectively. These models were trained from scratch and end-to-end. In terms of both efficacy and parameter size, they surpass VGG19. Specifically, the suggested MobileNetV2 with ViT model could be preferred in terms of performance and parameter size; this model increased test accuracy by 5.54 percent compared to VGG19 and attained state-of-the-art results in the ear gender recognition field. In terms of parameter size, the magnitude of the suggested model was

approximately 1/18 that of VGG19.

Table 6 illustrates the results of state-of-the-art models proposed for the ear gender recognition problem. Karasulu et al. [45] created a hybrid model by combining CNN and RNN architectures and achieved an accuracy of 85.16% on the test set. Nguyen-Quoc and Hoang [8] conducted experimental studies using popular CNN models with their default values and obtained the highest score of 91.1% with the VGG19 model. In terms of test accuracy, number of parameters, and training time, our proposed MobileNetV2 with ViT model provides a superior model architecture.

Table 6. Comparison of models

	Network	Accuracy (%)
Karasulu et al. [45]	CNN with RNN	85.16
Nguyen-Quoc and Hoang [8]	AlexNet	87.65
	VGG16	91.04
	VGG19	91.12
	InceptionV3	87.62
	ResNet50	89.37
	ResNet101	89.39
	DenseNet201	88.30
	MobileNetV2	85.55
Our	MobileNetV2 with ViT	96.66

The hybrid model generated by integrating CNN, MLP-Mixer, and ViT has demonstrated superior performance in comparison to existing methods. The factors contributing to this performance improvement are as follows: (1) Complementary Feature Extraction: The combination of CNN, MLP Mixer, and ViT provides for complementary feature extraction capabilities. CNNs are renowned for their ability to identify local and spatial features in images, whereas ViT excels at identifying global context and long-distance dependencies. MLP Mixer complements these methods by enhancing the representation of features through token mixing. Using the strengths of each component, hybrid models produce a more comprehensive representation of the input data. (2) Hierarchical Information Processing: CNNs inherently capture hierarchical information through their stacked convolutional layers. MLP Mixers, on the other hand, introduce token mixing operations that enable the model to explicitly model relationships between different parts of the image. ViT uses self-attention mechanisms to capture global dependencies. By integrating these techniques, hybrid models are able to process both local and global data efficiently, resulting in enhanced performance. (3) Enhanced Representation Learning: The combination of different architectures allows for enhanced representation learning. CNNs, MLP Mixers, and ViT each have their own unique mechanisms for learning representations from data. By integrating these mechanisms, the hybrid models can capture a wider range of features and patterns, potentially leading to better discrimination and classification capabilities. (4) Adaptability to Data Characteristics: The hybrid models provide flexibility in adapting to different types of data characteristics. CNNs have been widely used for image-related tasks and are effective in capturing spatial information. MLP Mixers, with their token mixing operations, can handle various input sizes and effectively model relationships between tokens. ViT, with its attention mechanism, is capable of handling both image and sequence data. The combination of these models allows for a more versatile and adaptable approach to the task, accommodating different data characteristics that may arise. In summary, the proposed hybrid models leverage the strengths of CNN, MLP Mixer, and ViT to capture complementary features, process hierarchical information, enhance representation learning, and adapt to various data characteristics. These factors contribute to the observed performance improvements when compared to current methods.

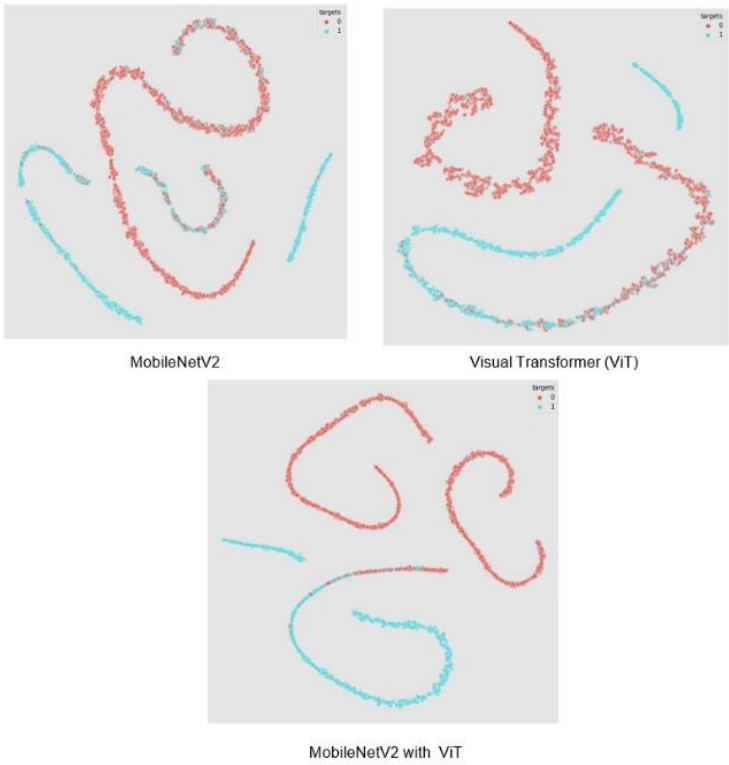

Figure 5. Evaluating the models using the t-SNE method. The graphs represent the 0 (red): male and 1 (blue): female class

In Figure 5, the results of the t-Distributed Stochastic Neighbor Embedding (t-SNE) method are presented for evaluation of the models. t-SNE is a dimensionality reduction technique that is particularly effective for visualizing high-dimensional data sets and was first introduced by Van der Maaten and Hinton [46]. The t-SNE algorithm calculates a measure of similarity between pairings of samples in high-dimensional space and low-dimensional space; it then attempts to optimize these two similarity measures using a loss function. This method discovers a non-parametric mapping and provides an intuitive understanding of how the data is structured in high-dimensional space. The entanglement or disentanglement of features reveals the model's performance. Examining Figure 5, it is observed that the results of the MobileNetV2 and ViT models are intertwined, making it challenging to separate the classes with a simple discriminative boundary. In our proposed MobileNetV2 with ViT model, the features obtained in the low-dimensional space are disentangled, and it is observed that the two classes can be easily separated. These results show that our proposed model is more successful for the ear gender recognition problem.

6. CONCLUSION

In recent years, deep learning techniques have been extensively researched for their potential in various applications, including biometrics. One such application is the use of ear images for gender recognition. Our research significantly contributes to the field of biometric identification technology, specifically in the area of ear gender recognition. The main contributions of this work can be summarized as follows:

(1) Hybrid Recognition Architecture: We introduce a novel hybrid recognition architecture named "MobileNetV2 with ViT." By combining the strengths of CNN and ViT, our proposed model achieves remarkable performance in gender classification tasks.

(2) Advantages over Alternative Models: Through comprehensive evaluations, we compare our hybrid architecture with alternative models like CNN models, MLP-Mixer and ViT. Our findings indicate that the hybrid model outperforms these alternatives in terms of both performance and computational efficiency. Additionally, it stands out with a favorable number of parameters, making it a superior choice for gender recognition tasks.

(3) Effective Feature Disentanglement: We provide t-SNE results to demonstrate how well our proposed model disentangles features in a low-dimensional space compared to other models. This visualization highlights the model's ability to effectively represent gender-related features, contributing to its accuracy in classification.

The proposed model presents a strong approach for gender recognition using ear images, achieving an impressive success rate of 96.66% on the EarVN1.0 dataset. This achievement opens up promising applications in real-world scenarios, including human-computer interaction, secure banking transactions, gender-based disease diagnosis, and demographic data collection. Notably, the study's reliance on a single dataset is a limitation, necessitating further research to enhance the model's ability to generalize across diverse ear datasets.

REFERENCES

[1] Ataş, M., Özdemir, C., Ataş, İ., Ak, B., Özeroğlu, E. (2022). Biometric identification using panoramic dental radiographic images withfew-shot learning. Turkish Journal of Electrical Engineering and Computer Sciences, 30(3): 1115-1126. https://doi.org/10.55730/1300-0632.3830

[2] Ozdemir, C., Gedik, M.A., Kaya, Y. (2021). Age estimation from left-hand radiographs with deep learning methods. Traitement du Signal, 38(6): 1565-1574. https://doi.org/10.18280/ts.380601

[3] Emeršič, Ž., Štruc, V., Peer, P. (2017). Ear recognition: more than a survey. Neurocomputing, 255: 26-39. https://doi.org/10.1016/j.neucom.2016.08.139

[4] Kumar, A., Wu, C.Y. (2012). Automated human identification using ear imaging. Pattern Recognition, 45(3): 956-968. https://doi.org/10.1016/j.patcog.2011.06.005

[5] Yavuz, M.S., Tatlısumak, E., Özyurt, B., Aşırdizer, M. (2013). The investigation of the effects of observers' gender in personal identification from auricle morphology. Turkish Journal of Forensic Medicine, 27(3): 173-181. https://doi.org/10.5505/adlitip.2013.08216

[6] Nixon, M.S., Bouchrika, I., Arbab-Zavar, B., Carter, J.N. (2010). On use of biometrics in forensics: Gait and ear. In 2010 18th European Signal Processing Conference, IEEE, pp. 1655-1659.

[7] Abaza, A., Ross, A., Hebert, C., Harrison, M.A.F., Nixon, M.S. (2013). A survey on ear biometrics. ACM Computing Surveys (CSUR), 45(2): 1-35. https://doi.org/10.1145/2431211.2431221

[8] Nguyen-Quoc, H., Hoang, V.T. (2020). Gender recognition based on ear images: A comparative experimental study. In 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), IEEE, pp. 451-456. https://doi.org/10.1109/ISRITI51436.2020.9315366

[9] Emeršič, Ž., Štepec, D., Štruc, V., Peer, P., George, A., Ahmad, A., Omar, E., Boult, T.E., Safdaii, R., Zhou, Y.X., Zafeiriou, S., Yaman, D., Eyiokur, F.I., Ekenel, H.K. (2017). The unconstrained ear recognition challenge. In 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, pp. 715-724. https://doi.org/10.1109/BTAS.2017.8272761

[10] Gnanasivam, P., Muttan, S. (2013). Gender classification using ear biometrics. In Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP), Springer India, 2: 137-148. https://doi.org/10.1007/978-81-322-1000-9_13

[11] Zhang, G.P., Wang, Y.H. (2011). Hierarchical and discriminative bag of features for face profile and ear based gender classification. In 2011 International Joint Conference on Biometrics (IJCB), IEEE, pp. 1-8. https://doi.org/10.1109/IJCB.2011.6117590

[12] Khorsandi, R., Abdel-Mottaleb, M. (2013). Gender classification using 2-D ear images and sparse representation. In 2013 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, pp. 461-466. https://doi.org/10.1109/WACV.2013.6475055

[13] Lei, J.J., Zhou, J.D., Abdel-Mottaleb, M. (2013). Gender classification using automatically detected and aligned 3D ear range data. In 2013 International Conference on

Biometrics (ICB), IEEE, pp. 1-7. https://doi.org/10.1109/ICB.2013.6612995
[14] Yan, P., Bowyer, K. (2005). Empirical evaluation of advanced ear biometrics. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, IEEE, pp. 41-41. https://doi.org/10.1109/CVPR.2005.450
[15] Dodge, S., Mounsef, J., Karam, L. (2018). Unconstrained ear recognition using deep neural networks. IET Biometrics, 7(3): 207-214. https://doi.org/10.1049/iet-bmt.2017.0208
[16] Alshazly, H., Linse, C., Barth, E., Martinetz, T. (2019). Ensembles of deep learning models and transfer learning for ear recognition. Sensors, 19(19): 4139. https://doi.org/10.3390/s19194139
[17] Ahila Priyadharshini, R., Arivazhagan, S., Arun, M. (2021). A deep learning approach for person identification using ear biometrics. Applied Intelligence, 51: 2161-2172. https://doi.org/10.1007/s10489-020-01995-8
[18] Khaldi, Y., Benzaoui, A. (2021). A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions. Evolving Systems, 12(4): 923-934. https://doi.org/10.1007/s12530-020-09346-1
[19] Khaldi, Y., Benzaoui, A., Ouahabi, A., Jacques, S., Taleb-Ahmed, A. (2021). Ear recognition based on deep unsupervised active learning. IEEE Sensors Journal, 21(18): 20704-20713. https://doi.org/10.1109/JSEN.2021.3100151
[20] Mewada, H.K., Patel, A.V., Chaudhari, J., Mahant, K., Vala, A. (2020). Wavelet features embedded convolutional neural network for multiscale ear recognition. Journal of Electronic Imaging, 29(4): 043029. https://doi.org/10.1117/1.JEI.29.4.043029
[21] Omara, I., Hagag, A., Ma, G.Z., Abd El-Samie, F.E., Song, E. (2021). A novel approach for ear recognition: Learning mahalanobis distance features from deep CNNs. Machine Vision and Applications, 32: 1-14. https://doi.org/10.1007/s00138-020-01155-5
[22] Korichi, A., Slatnia, S., Aiadi, O. (2022). TR-ICANet: A fast unsupervised deep-learning-based scheme for unconstrained ear recognition. Arabian Journal for Science and Engineering, 47(8): 9887-9898. https://doi.org/10.1007/s13369-021-06375-z
[23] Emeršič, Ž., Križaj, J., Štruc, V., Peer, P. (2019). Deep ear recognition pipeline. Recent Advances in Computer Vision: Theories and Applications, 333-362. https://doi.org/10.1007/978-3-030-03000-1_14
[24] Alshazly, H., Linse, C., Barth, E., Martinetz, T. (2020). Deep convolutional neural networks for unconstrained ear recognition. IEEE Access, 8: 170295-170310. https://doi.org/10.1109/ACCESS.2020.3024116
[25] Radhika, K., Devika, K., Aswathi, T., Sreevidya, P., Sowmya, V., Soman, K.P. (2020). Performance analysis of NASNet on unconstrained ear recognition. Nature Inspired Computing for Data Science, 57-82. https://doi.org/10.1007/978-3-030-33820-6_3
[26] Alshazly, H., Linse, C., Barth, E., Idris, S.A., Martinetz, T. (2021). Towards explainable ear recognition systems using deep residual networks. IEEE Access, 9: 122254-122273. https://doi.org/10.1109/ACCESS.2021.3109441
[27] Frejlichowski, D., Tyszkiewicz, N. (2010). The west pomeranian university of technology ear database-a tool for testing biometric algorithms. In Image Analysis and Recognition: 7th International Conference (ICIAR), Springer Berlin Heidelberg, pp. 227-234. https://doi.org/10.1007/978-3-642-13775-4_23
[28] Zhou, Y.X., Zaferiou, S. (2017). Deformable models of ears in-the-wild for alignment and recognition. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG), IEEE, pp. 626-633. https://doi.org/10.1109/FG.2017.79
[29] Hoang, V.T. (2019). EarVN1.0: A new large-scale ear images dataset in the wild. Data in Brief, 27: 104630. https://doi.org/10.1016/j.dib.2019.104630
[30] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z.H., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Li, F.F. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115: 211-252. https://doi.org/10.1007/s11263-015-0816-y
[31] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. https://doi.org/10.1145/3065386
[32] Smirnov, E.A., Timoshenko, D.M., Andrianov, S.N. (2014). Comparison of regularization methods for imagenet classification with deep convolutional neural networks. Aasri Procedia, 6: 89-94. https://doi.org/10.1016/j.aasri.2014.05.013
[33] Kiliç, Ş., Askerzade, I., Kaya, Y. (2020). Using ResNet transfer deep learning methods in person identification according to physical actions. IEEE Access, 8: 220364-220373. https://doi.org/10.1109/ACCESS.2020.3040649
[34] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv: 1409.1556. https://doi.org/10.48550/arXiv.1409.1556
[35] Saxe, A.M., Koh, P.W., Chen, Z.H., Bhand, M., Suresh, B., Ng, A.Y. (2011). On random weights and unsupervised feature learning. In ICML, 2(3): 1-9.
[36] Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594
[37] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. https://doi.org/10.1109/CVPR.2016.90
[38] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700-4708. https://doi.org/10.1109/CVPR.2017.243
[39] Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X.H., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A. (2021). MLP-mixer: an all-mlp architecture for vision. Advances in Neural Information Processing Systems, 34: 24261-24272.
[40] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X.H., Unterthiner, T., Dehghani, M., Minderer,

M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16×16 words: transformers for image recognition at scale. arXiv Preprint arXiv: 2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[41] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. arXiv Preprint arXiv: 2012.12877. https://doi.org/10.48550/arXiv.2012.12877

[42] Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N.J., Li, Y.X., Liu, H.R., Zheng, Y.F. (2021). MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification. In Medical Image Computing and Computer Assisted Intervention (MICCAI 2021): 24th International Conference, Springer International Publishing, pp. 45-54. https://doi.org/10.1007/978-3-030-87237-3_5

[43] Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J.C., Shi, H. (2021). Escaping the big data paradigm with compact transformers. arXiv Preprint arXiv: 2104.05704. https://doi.org/10.48550/arXiv.2104.05704

[44] Tieleman, T., Hinton, G. (2017). Divide the gradient by a running average of its recent magnitude. Coursera: Neural networks for machine learning. Technical Report.

[45] Karasulu, B., Yücalar, F., Borandağ, E. (2022). A hybrid approach based on deep learning for gender recognition using human ear images. Journal of the Faculty of Engineering and Architecture of Gazi University, 37(3): 1579-1594. https://doi.org/10.17341/gazimmfd.945188

[46] Van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11): 2579-2605.