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

Traitement du Signal
Vol. 41, No. 3, June, 2024, pp. 1405-1418
Journal homepage: http://iieta.org/journals/ts

Advancing Biometric Identity Recognition with Optimized Deep Convolutional Neural Networks

Hammam Alshazly[1], Hela Elmannai[2*], Reem Ibrahim Alkanhel[2], Abdelrahman Abdelnazeer[1]

[1] Faculty of Computers and Information, South Valley University, Qena 83523, Egypt
[2] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding Author Email: hselmannai@pnu.edu.sa

https://doi.org/10.18280/ts.410329

Received: 17 April 2023
Revised: 1 August 2023
Accepted: 18 October 2023
Available online: 26 June 2024

Keywords:
deep learning, ear recognition, ear biometrics, convolutional neural networks, transfer learning, visual explanation, interpretable models

ABSTRACT

Biometric identity recognition, capitalizing on unique physical attributes, represents an increasingly explored research field within the biometrics community, with implications spanning surveillance, crowd analytics, automated identity checks, and user device access. Ear images, in particular, offer a robust data source for devising effective personal identification systems. The biometric field has seen a surge in the application of machine learning algorithms, specifically deep neural network architectures such as Convolutional Neural Networks (CNNs) and transfer learning methods, to enhance ear recognition systems. This study evaluates leading deep CNN architectures - ResNet, DenseNet, MobileNet, and Inception - for their efficacy in creating ear recognition systems resilient to varying imaging conditions. The AMI and WPUT datasets, publicly accessible ear image datasets, were utilized to train and assess the proposed models. The models demonstrated substantial success, achieving rank-1 accuracies of 96% and 83% on the AMI and WPUT datasets, respectively. Additionally, the Gradient-weighted Class Activation Mapping (Grad-CAM) visualization technique was employed to elucidate the models' decision-making processes, revealing a reliance on auxiliary features like hair, cheek, or neck when available. The use of Grad-CAM not only enhances understanding of the decision-making processes within the CNNs but also highlights potential areas of improvement for the proposed ear recognition systems.

1. INTRODUCTION

Recent advancements in communications, coupled with increasing security demands, have stimulated active research within the biometric community on identity recognition based on physical characteristics. This domain is pivotal for applications such as surveillance, crowd analytics, automated identity checks, and device unlocking. Yet, the construction of a robust identity recognition system necessitates the use of physical traits that are consistent, measurable, and universal [1].

The quest for a unique biometric data representation is crucial for the success of automatic identification systems. This quest has been vigorously explored in facial recognition but remains challenged by uncontrollable image conditions and significant intraclass variations, brought about by changes in illumination, perspective, age, clutter, occlusions, or inconsistent image size or quality. The human ear, akin to the face, presents biometric data that can be utilized for individual identification [2]. Unlike face images, however, ear images are immune to physical deformations due to the absence of muscle activity. These images offer a consistent and recognizable structure, typically unaffected by movement, speech, or facial emotions. The visual appearance of the ear is generally unaltered by hygienic factors such as makeup, an added benefit. Furthermore, capturing ear images is less invasive than acquiring facial images, and in comparison to other biometric modalities such as fingerprinting, ear imaging is a contact-free procedure. For surveillance applications, ear images ensure quality and non-interchangeability that do not necessitate the subject's cooperation. These advantages posit ear images as suitable candidates for incorporation into automated identification systems or as complementary to facial recognition for enhancing the accuracy of profile images [3].

The early stages of ear recognition operated using a conventional processing pipeline, comprising feature extraction and classification. This two-step method dominated identity recognition using ear images, with manual feature extraction followed by standard classifiers [4-7]. Initially, textural information or structural descriptors of the ear were exploited to manually craft discriminative features. Subsequently, a linear or nonlinear classifier was applied to actualize identity prediction. Despite these techniques' proven success with small and mid-sized ear datasets, they are unscalable to larger datasets. Manual feature extraction is subjective, requiring domain expertise, and the extracted features often fall short in capturing complex patterns and

relationships in ear images. Moreover, conventional classifiers often underperform when applied to complex datasets with high dimensionality or nonlinear relationships between features. They are also prone to overfitting, particularly when the number of features is large relative to the number of samples in the dataset. Nonetheless, current trends in ear recognition favor deep learning-based techniques, attributed to their scalability and superior recognition performance [8-10].

Deep learning has rapidly evolved into a popular data-driven learning approach for various image recognition problems, including object detection [11, 12], deepfake detection [13, 14], and medical image analysis [15-18]. It amalgamates the feature extraction and classification processes into a single end-to-end model. Deep Convolutional Neural Networks (CNNs) offer an effective way to utilize deep learning for image recognition. As computers perceive images through pixels, the convolution process uses the relationship between image pixels to aid in image identification. An advantage of CNNs over their predecessors is the automatic detection of significant features without human intervention. Training deep neural networks also tunes the representation of the input data to the specific task, contributing to the high adaptability of deep learning techniques. However, this comes at a cost. Training such networks requires a large amount of data to counteract overfitting, a common problem in machine learning where an identification system memorizes the training images rather than learning the underlying relationships in the input data to distinguish between individuals. To mitigate this problem, this study exploits transfer learning, a potent technique in machine learning that involves reusing pretrained models and adapting them to new tasks [19].

In the realm of ear recognition, Abd Almisreb et al. [20] applied transfer learning to AlexNet [21] as a solution for ear recognition. A modest collection of 300 ear images from 10 subjects was used to fine-tune the pretrained network, with 250 images allocated for training and 50 images for validation and testing. The fine-tuned network achieved a remarkable recognition accuracy nearing 100%. Transfer learning was also employed on the renowned VGG16 model, paired with a Support Vector Machine (SVM), to fashion a hybrid algorithm for individual identification via ear images [22]. A dataset comprising 2600 ear images showcasing various poses, rotations, and illumination variations validated the proposed model. The model achieved a recognition accuracy of 98.72% in classifying ear images. In another study [23], transfer learning was used with VGG16 and ResNet50 models to accelerate model construction and enhance ear recognition performance. The IIT Delhi ear dataset, consisting of 1286 ear images from 221 subjects, was used to evaluate these models. The VGG16 model outperformed ResNet50, achieving a recognition accuracy of 89.73% based on their experiments. Alejo and Hate [24] examined the use of transfer learning to tackle the challenge of unconstrained ear recognition. Eight different pretrained CNN models were explored, and their performances were compared on a dataset of 250 ear images from 10 subjects, sourced from the Internet. The models achieved a recognition accuracy above 95%. In this study, we apply transfer learning methods to enhance the accuracy of ear recognition models by leveraging pretrained models on related image recognition tasks. One approach to using transfer learning for ear recognition is to utilize pretrained models trained on similar visual recognition tasks, such as face recognition or object recognition. These models can then be fine-tuned with ear images to improve their performance on the ear recognition task. This approach is particularly useful when there is limited labeled ear image data available for training the models from scratch [25-27].

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

- ✧ We explore and compare transfer learning and fine-tuning of deep CNN architectures to aid in the design process of robust ear identification systems.
- ✧ Various deep CNN architectures are evaluated for their robustness in recognizing individuals from their ear images.
- ✧ The proposed models are compared against a range of deep learning-based recognition methods on two publicly available benchmark datasets, aiming to enhance the overall recognition accuracy.
- ✧ We provide visual explanations for a better understanding of the proposed models, offering insights into which parts of the ear image are most critical for recognition decisions.

The remainder of the paper is organized as follows: Section 2 discusses the related previous work. Section 3 describes the different CNN architectures employed in this work. Section 4 outlines our methodology for building robust ear recognition models. The experimental settings and results are presented in Section 5. Finally, Section 6 concludes the paper and outlines the future research direction.

2. RELATED WORK

The potential of the human ear as a biometric modality for identity recognition has been the focus of numerous research studies [28-30]. Previous studies have demonstrated that ear recognition, using traditional machine learning methods, can achieve satisfactory recognition rates with carefully crafted features [31-33]. However, these methods often prove sensitive to noise, illumination, and pose variations, and may struggle to capture the subtle differences among ear images. To overcome these limitations, the introduction of deep learning-based ear recognition methods has been proposed. Researchers are exploring the potential of deep network architectures such as CNNs and transfer learning methods to enhance performance. As ear image datasets are relatively small and overfitting is a concern, deep learning techniques have only recently been applied to ear recognition tasks. Strategies such as aggressive data augmentation, model size reduction, regularization techniques, or transfer learning using pretrained models from large datasets like ImageNet are crucial to navigate these constraints. The most significant challenge for current ear recognition algorithms is the transition from controlled to uncontrolled imaging conditions, typically attributed to the lack of extensive, large-scale ear datasets encompassing highly variable ear images.

Recently, the use of deep neural networks for ear recognition has led to notable improvements in recognition performance compared to traditional methods. Emeršič et al. [34] examined the challenge of training deep CNN-based models using a limited number of ear images. They explored three different network architectures and various approaches to train models, employing different degrees of image augmentation up to 100 times the original training image. A combined dataset of 2304 ear images from 166 subjects was

used for model training and evaluation. The top-performing models demonstrated the capacity to automatically learn discriminative features from raw ear images, achieving a recognition accuracy of 62%. In another study [35], the authors conducted an experimental exploration of ear recognition using various deep architectures of increasing depth, specifically, the VGG models [36]. The AMI and WPUT ear datasets were utilized for the experiments, with three different network training strategies tested. The strategy of fine-tuning emerged as the most effective, achieving a rank-1 recognition accuracy of 96.78% and 74.36% on the AMI and WPUT datasets, respectively. To further enhance accuracy, numerous model combinations were ensembled. A similar study was conducted in the study [27], where the authors experimented with various deep residual networks (ResNet [37]) of differing depths. Fine-tuning strategies were employed to achieve improved performance. Four ear datasets, containing ear images captured under both constrained and unconstrained imaging conditions, were used to evaluate the models. The proposed models achieved rank-1 recognition accuracy ranging from 67% to 99% for ear images taken under unconstrained and constrained conditions, respectively. Moreover, the best results were garnered using an ensemble of several fine-tuned ResNet models of varying depth.

To assess the effectiveness of ear recognition technology on a substantial ear dataset, the inaugural Unconstrained Ear Recognition Challenge (UERC) [21] was held in 2017. For model evaluation, the challenge considered tightly cropped ear images that exhibited various head movements (poses), lighting changes, image resolution variations, and occlusions. The ear recognition methods submitted for the challenge were evaluated and analyzed to investigate their capacity to handle these diverse variations in ear images. The study shed light on some key findings, including the need for significant performance improvements prior to deploying ear recognition technology in unconstrained environments. Additionally, the experiments revealed that the evaluated methods were sensitive to changes in head poses.

The authors [7] proposed a framework to tackle the unconstrained ear recognition problem using multiple ear image datasets. They leveraged CNN-based models for ear normalization and description, combined with a set of hand-picked handcrafted image descriptors, and then fused both handcrafted and CNN-based features. A variety of feature combinations were tested, and substantial improvements in recognition performance were reported when both feature types were combined.

The authors [26] proposed a two-stage domain adaptation strategy of fine-tuning deep CNN-based models to address the unconstrained ear recognition problem. They conducted a thorough analysis of several crucial factors such as dataset bias, illumination, aspect ratio, the impact of data augmentation, and alignment on ear recognition performance.

Alshazly et al. [38] conducted an extensive study on unconstrained ear recognition. They employed different transfer learning strategies using well-known deep CNN architectures, including AlexNet, VGG, Inception [39], ResNet, and ResNext [40], to overcome the problem of inadequate data for training deep CNNs from scratch. The experiments were conducted on the EarVN1.0 dataset [41], which comprises 28,412 ear images from 164 subjects obtained from the web. The results indicated an improved recognition performance above 93% when using the fine-tuning strategy of pretrained deep CNN architectures with custom-sized inputs to maintain the aspect ratio of the image in the EarVN1.0 dataset.

Khaldi et al. [42] proposed implementing a deep unsupervised active learning (DUAL) strategy in the field of ear recognition. They used a pretrained VGG16 model and applied it to three ear datasets. The training process was divided into two stages: an initial supervised training stage using a classification model, followed by an unsupervised active learning stage. Three ear image datasets comprising ear images captured under both constrained and unconstrained imaging conditions were used for training and performance evaluation. The proposed technique achieved superior performance, indicating a significant improvement in the recognition rate.

Our study builds upon previous research on unconstrained ear recognition and reports the results of experiments conducted on two challenging ear image datasets. Further, we evaluated the performance of new deep CNN models (DenseNet and MobileNet) in ear identification. Given the limited quantity of ear images in the datasets considered, we proposed a transfer learning approach to fine-tune the pretrained models on ear images, aided by data augmentation, to achieve improved recognition performance. We provided a detailed comparative analysis using various performance evaluation metrics and reported the results achieved by each model. Owing to the black-box nature of deep models and in an effort to make them more transparent, we applied the Grad-CAM visualization technique. Grad-CAM provides visual explanations, highlighting the significant ear image regions that are frequently considered by the models for making accurate predictions.

3. DEEP ARCHITECTURES

Deep CNN architectures are the type of deep learning algorithms particularly well-suited for image processing and recognition tasks. Over the years, CNN architectures have evolved and different variants of the CNN architectures have been developed, resulting in incredible advancements in the growing field of deep learning. The development of new architectures is to achieve comparable accuracy and address the problems related to computational efficiency, error rate, and gradient vanishing or exploding. This section provides a brief description of the deep CNN architectures employed in our study to construct the ear recognition system. Moreover, it highlights the diverse building blocks and approaches used to build these deep architectures.

3.1 ResNet

A residual network (ResNet), is a deep neural network architecture that was introduced by He et al. [37] in 2015. It is a modification of the traditional CNN architecture that overcomes the problem of vanishing gradients that occurs in very deep networks. The ResNet architecture introduces the concept of residual learning, which involves the addition of shortcut connections between layers that skip over one or more layers in the network. These shortcut connections enable the gradient to flow directly through the network, which makes it easier for the network to learn from the data. The ResNet architecture is characterized by a series of residual blocks that contain several convolutional layers and shortcut connections. The residual blocks allow the network to learn more complex

features from the data and also enable it to be trained more efficiently. In addition, the ResNet architecture includes batch normalization and ReLU activation functions, which further improve the network performance.

3.2 DenseNet

A densely connected network (DenseNet) is deep network architecture that was introduced by Huang et al. [43] in 2017. It is a modification of the traditional CNN architecture that improves the efficiency of information flow between layers. The DenseNet architecture introduces the concept of dense connectivity, which involves connecting every layer to every other layer in a feedforward fashion. This dense connectivity enables the network to extract more feature representations from the input and effectively reuse them across the network. The cornerstone of the DenseNet architecture is the Dense block, which enables the network to learn complex feature representations from the input and effectively reuse them across the network. The dense connectivity pattern ensures that each layer in the block has access to all previously learned features, and reduces the number of parameters in the network, making it more computationally efficient. The DenseNet architecture is constructed using a series of Dense blocks that contain several convolutional layers, batch normalization, and ReLU activation functions. In each dense block, the output of each layer is concatenated with the outputs of all previous layers in the block.

3.3 InceptionV3

InceptionV3 is a deep convolutional neural network architecture introduced by Google in 2015 as an extension to the Inception family of models [39]. It is a deep learning architecture designed for image classification and object detection tasks. The InceptionV3 architecture consists of multiple modules, with each module containing a combination of different convolutional layers. The main idea behind the InceptionV3 architecture is to use multiple filters of different sizes, which allows the network to capture features at different scales. This is achieved using a combination of 1×1, 3×3, and 5×5 convolutional filters within the same layer. Additionally, the architecture also incorporates the use of pooling and batch normalization layers to improve training stability and accuracy. The cornerstone of the InceptionV3 architecture is the Inception module, which consists of a combination of different convolutional layers. The Inception module is designed to maximize the use of computational resources by using parallel convolutional layers with different filter sizes. This enables the network to capture features at different scales and helps reduce the number of parameters in the network.

3.4 MobileNet

MobileNet is a convolutional neural network architecture introduced by Google in 2017 that is optimized for mobile and embedded devices [44]. The main objective of MobileNet is to provide high accuracy with low computational cost and low memory footprint, making it suitable for deployment on mobile devices with limited resources. The MobileNet architecture consists of depthwise separable convolutions, which decompose the standard convolutional operation into two separate layers: depthwise and pointwise convolution. Depthwise convolution applies a single filter to each input channel, producing a set of output channels. The output channels generated by the depthwise convolution are then combined using a pointwise convolution. This approach significantly reduces the number of parameters in the network while maintaining high accuracy. MobileNet also uses a technique called linear bottleneck, where the number of filters is reduced at the beginning of each layer and then increased again at the end of the layer. This technique reduces the computational cost of the network while preserving its accuracy.

4. PROPOSED METHODOLOGY

This section describes the proposed framework for ear recognition. As discussed in the studies [45, 46], the learned features of a deep CNN trained on large image datasets are highly transferable to other vision tasks and datasets. As the similarity between the pretraining and target tasks grows, the transferability becomes more effective. Nevertheless, the transfer of learned features, even from a remote task, is superior to learn them from scratch on the target dataset.

4.1 Transfer learning

Transfer learning is a machine learning approach in which a pretrained model is used as the starting point for a model on a new task. In the context of deep neural networks, two commonly used transfer learning methods are applicable: fine-tuning and feature extraction. Fine-tuning is a technique used in deep learning to adapt a pretrained model for a new task or domain. Pretrained models are trained on large datasets and can often learn general features useful for other tasks beyond their original training objectives. Fine-tuning involves taking a pretrained model and training it on a new task or dataset, often with a smaller amount of training data than the original training data. This approach can significantly improve the performance of a model on a new task, while reducing the amount of data and training time required compared to training a new model from scratch.

The process of fine-tuning involves the following steps:

1. Load the pretrained model: Load the weights of a pretrained model, such as a deep CNN, from a publicly available source or one that you have trained yourself.
2. Replace the last layer: Replace the last layer of the pretrained model with a new layer that has the number of outputs required for the new task. For example, if the used pretrained models were trained for image classification with 1,000 categories, and our ear datasets have only 100 and 474 subjects, then the last layer needs to be replaced with a new layer that has 100 and 474 outputs, respectively.
3. Freeze some layers: Freeze some earlier layers in the pretrained model to prevent them from being updated during training. This is because the earlier layers in the model have learned more general features, such as edges and curves, that are likely to be useful for the ear recognition task.
4. Train the model: Train the modified model on the ear recognition task or dataset, often with a smaller learning rate than the pretrained model, to avoid overfitting.
5. Fine-tune the model: Gradually unfreeze more layers in the pretrained model and continue training until the model achieves the desired performance on the new task.

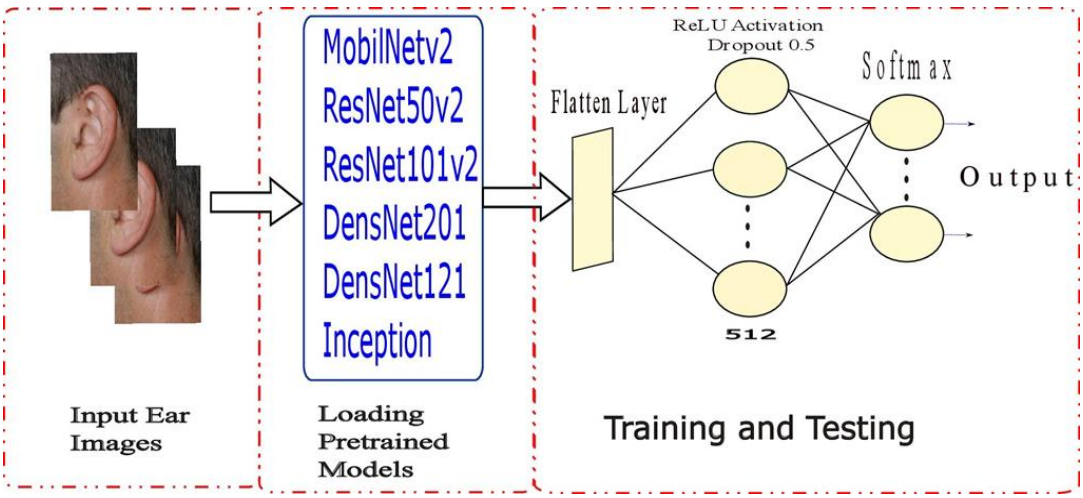


Figure 1. Schematic flow of fine-tuning pretrained CNN models on ear images from the WPUT ear dataset

4.2 Ear recognition framework

The ear recognition framework is based on the concepts of fine-tuning pretrained deep CNNs. The models included MobileNet, ResNet, Inception, and DenseNet that were trained on the ImageNet dataset. As we know, the final set of layers for these models are fully-connected (FC) layers with a softmax classifier. In fine-tuning, we actually remove the original set of FC layers and add a new set of FC layers, which are placed on top of the original architecture. These new FC layers can then be trained and adjusted to the specific ear image dataset. Usually, the newly added FC layers have fewer parameters than the original ones; however, this really depends on the particular dataset. The new FC layers are randomly initialized and connected to the body of original network. However, if we start training the entire network we face the problem of modifying the already learned and discriminating filters of the convolutional layers. The new FC layers are brand new and totally random and if the gradient backpropagates from these random values through the structure of the network, we encounter eliminating these discriminative features. To circumvent this, we freeze all layers in the network and allow only the newly attached layers to be adjusted. The network forward propagates the training data, but backpropagation is stopped after the FC layers, allowing the new layers to begin learning patterns from the discriminating convolutional layers. Training is then allowed to continue until sufficient accuracy is obtained. Figure 1 illustrates the schematic diagram of the fine-tuning process.

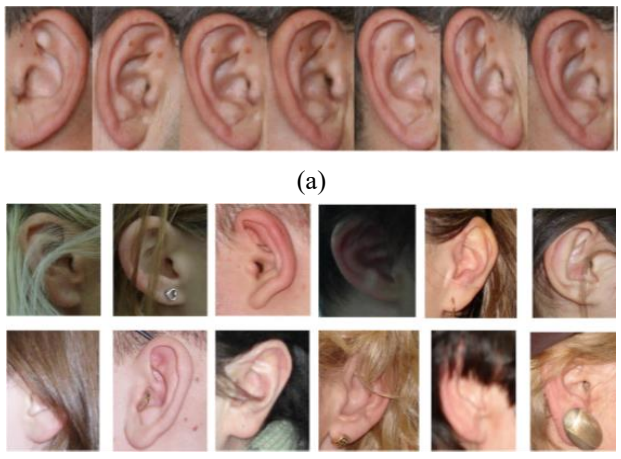
5. EXPERIMENTAL SETUP

This section discusses the experimental settings for the conducted experiments. The ear datasets, model settings, preprocessing steps including, data augmentation, and evaluation metrics are mentioned in the succeeding subsections.

5.1 Datasets

Human ears have various textures, colors, and shapes, making ear images distinctive for people and allowing the identity prediction. Moreover, occlusions, changes in illumination, perspective, and resolution are a few additional variables that contribute to the diversity of ear images. Generally, the selection of the ear dataset and the degree to which these variations are present and controlled have a significant impact on the overall difficulty of conducting identity prediction. Although there are numerous ear datasets, the number of images they offer is still constrained. Because there are few data to account for the significant intraclass variation, uncontrolled datasets pose a problem for deep learning. Two representative datasets were used to train and evaluate our proposed models. The first dataset is the Mathematical Analysis of Images (AMI) dataset [47], which contains 700 ear images acquired from 100 subjects, where each subject has six images for the right ear and a single image for the left ear. The second dataset is the West Pomeranian University of Technology (WPUT) dataset [48], which contains 1960 cropped ear images taken from 474 subjects, where each subject has some images between 4 and 8 images.

Example images from (a) AMI and (b) WPUT datasets are shown in Figure 2. Moreover, a summarized description of these datasets is given in Table 1.



(a)

(b)

Figure 2. Example images from (a) AMI and (b) WPUT datasets

Table 1. Description of ear datasets

Dataset	Subjects	Images	Resolution
AMI [47]	100	700	492×702
WPUT [48]	474	1960	380×500

5.2 Experimental settings

We divided each dataset into two disjoint sets, training and test, each of which contained 60% and 40% of the ear images, respectively. The training set is used to adjust the weights of the various networks, while the test set is used to evaluate the models and report the results. Table 2 describes the different setup configurations for each considered deep CNN architecture on the used ear datasets. The CNN models are trained for a number of epochs ranging from 100 to 250 with a batch size of 16. The ReLU activation function is used in the newly added layers. The adaptive moment optimizer (Adam) is used for optimizing all models. For the AMI dataset, all models are trained for 250 epochs except DenseNet201, which converges after 100 epochs. However, for the WPUT dataset, all models are trained for 250 epochs except InceptionV3, which requires 200 epochs to converge. The convergence of a neural network is the moment in training a model after which adjusting the learning rate becomes less significant and the errors produced by the model reach the lowest level of tolerable error. All models were trained on a machine with Intel (R) Core (TM) i7 CPU, 16 MB RAM, and Nvidia RTX until convergence.

Table 2. Parameter configuration for the deep CNN models on two ear datasets

Deep Network Architecture	Dataset	
	AMI	WPUT
InceptionV3		
The number of epochs	250	200
Batch size	16	16
ResNet50V2		
The number of epochs	250	250
Batch size	16	16
DenseNet121		
The number of epochs	250	250
Batch size	16	16
DenseNet201		
The number of epochs	100	250
Batch size	16	16
MobileNetV2		
The number of epochs	250	250
Batch size	16	16

6. DATA AUGMENTATION

Data augmentation is a technique used in deep learning to increase the size of a training dataset by creating new variations of the existing data. This technique is particularly useful when the available dataset is small, and the model must generalize well to unseen data. The process of data augmentation involves creating new training examples by applying various transformations to the original data. These transformations may include:

1. Flipping: flipping the image horizontally or vertically.
2. Rotation: rotating the image to a certain degree.
3. Zooming: cropping and scaling the image to create a zoomed-in or zoomed-out version.
4. Translation: shifting the image horizontally or vertically.
5. Brightness adjustment: increasing or decreasing the brightness of the image.
6. Contrast adjustment: increasing or decreasing the contrast of the image.
7. Adding noise: adding random noise to the image.
8. Shearing: distorting the image by shearing it in one or more directions.

These transformations create new variations of the original image that can help the model generalize better. For example, flipping an ear image horizontally creates a new ear image of the same person from a different angle, which can help the model learn to recognize ear images from different angles. For our experiments, we applied simple transformations, which include horizontal and vertical flipping, and shifting the height and width with a small range (i.e., 0.3 in case of AMI dataset and 0.2 in case of WPUT dataset).

6.1 Performance evaluation

Evaluation metrics for an ear recognition system can include accuracy, precision, recall, and F1-score. Here are the definitions and formulas for each metric. Here, TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively.

1. Accuracy: The proportion of correctly classified ear images over the total number of images.

Accuracy=(TP+TN)/(TP+FP+TN+FN)

2. Precision: The proportion of correctly classified positive cases (ear images) over the total number of positive cases predicted by the system.

Precision=TP/(TP+FP)

3. Recall (or sensitivity): The proportion of correctly classified positive cases (ear images) over the total number of positive cases in the dataset.

Recall=TP/(TP+FN)

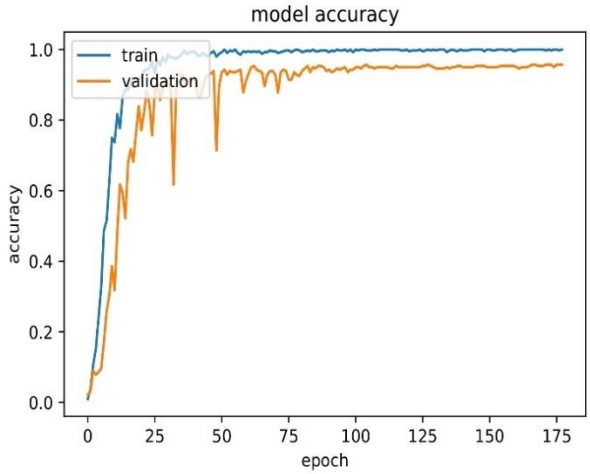
4. F1-score: The harmonic mean of precision and recall, providing a balanced measure of the system's performance.

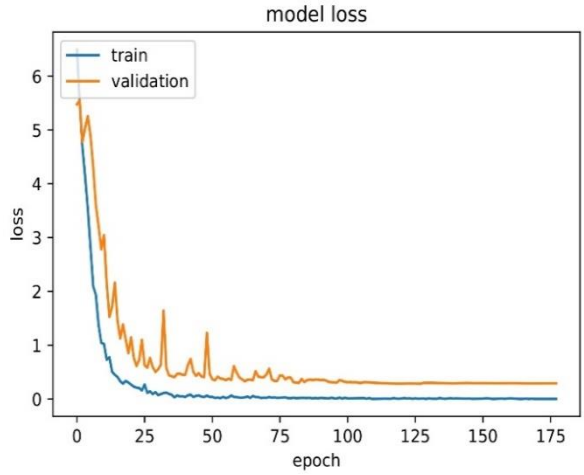
F1-Score=2*(Precision*Recall)/(Precision+Recall)

6.2 Experimental results

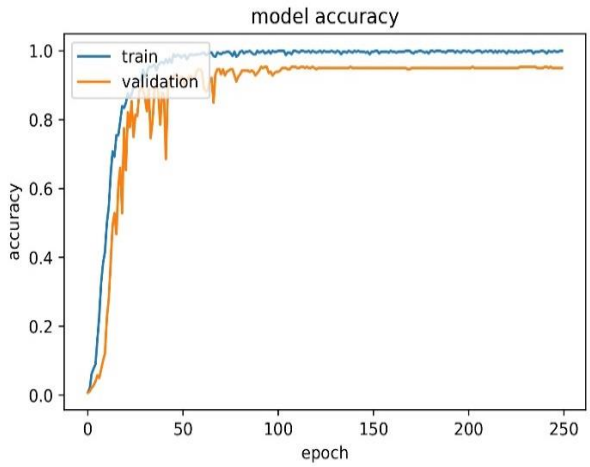
In an ear recognition system, accuracy and loss curves are used to evaluate the performance of the system during training and testing. The accuracy curve shows how well the recognition system is able to correctly classify input, as well as to evaluate the overall performance of the recognition system and to determine if it is improving or getting worse over time. The loss curve indicates how well the system minimizes the difference between predicted and true outputs. Figure 3 shows the accuracy and loss curves for the CNN models used when training and validating on AMI ear dataset. It can be seen from the accuracy curves that the performance of the models increases with time, indicating that they are learning. We also observe that they improve at the beginning, but over time they reach a plateau, which means that they are not able to learn anymore. On the other hand, the loss curve over time measures the models' error or how our models are doing. Despite the slight ups and downs, in the long run, the loss gets smaller, indicating that the models are improving and learning.

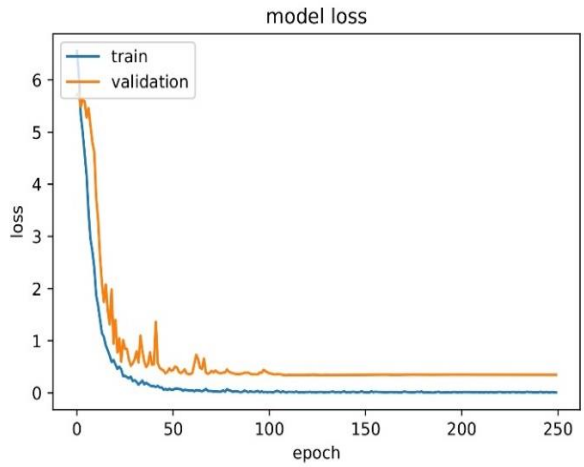
DenseNet201



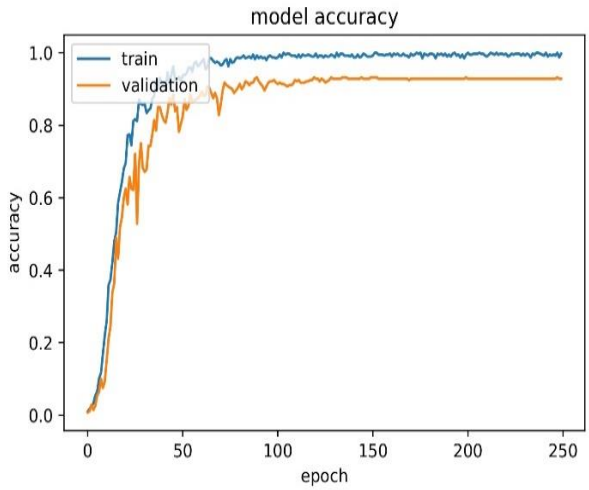


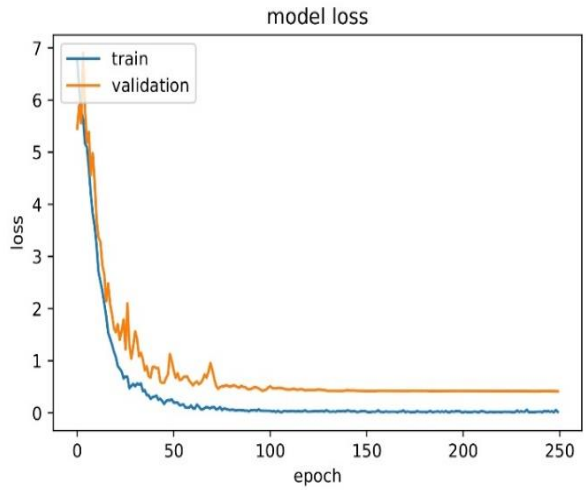
DenseNet121



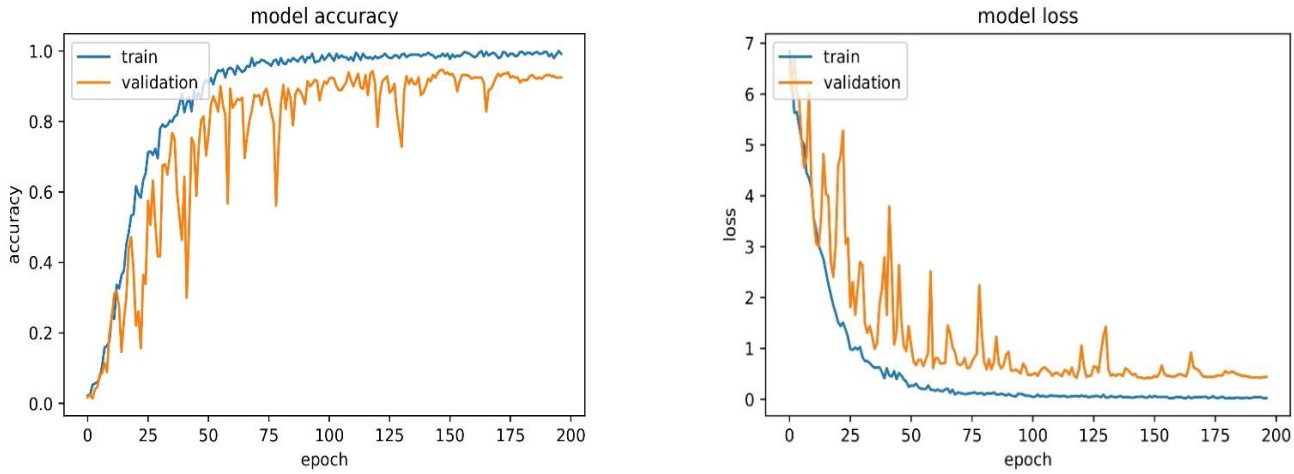


InceptionV3





ResNet50V2



ResNet101V2

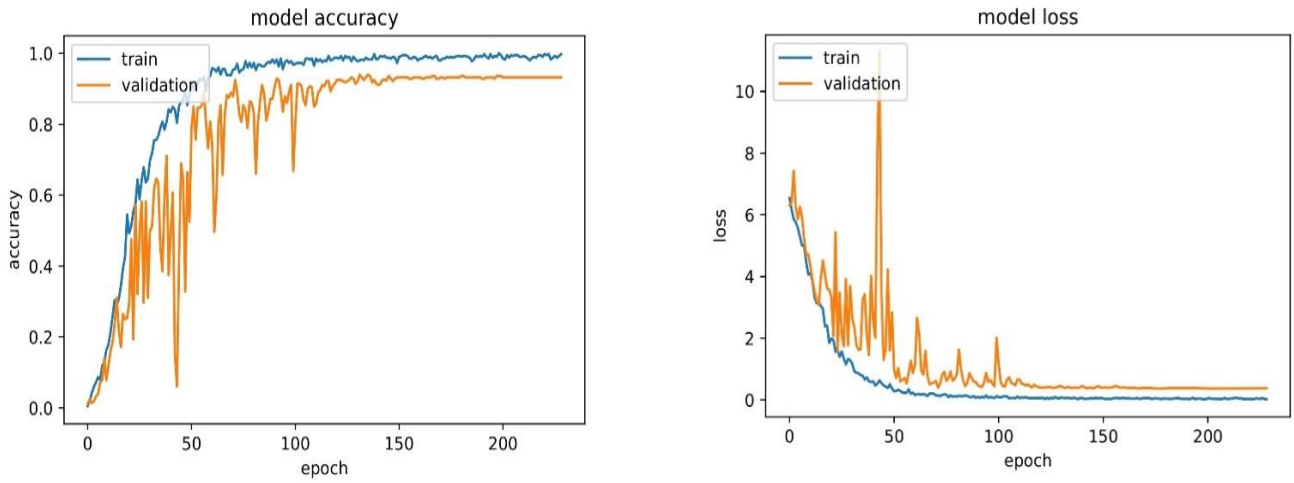
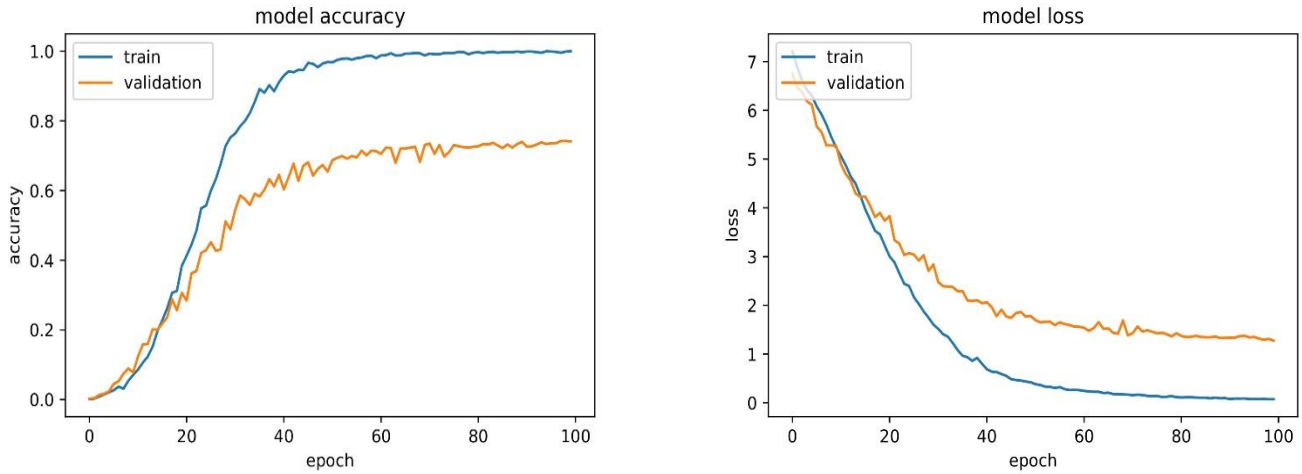
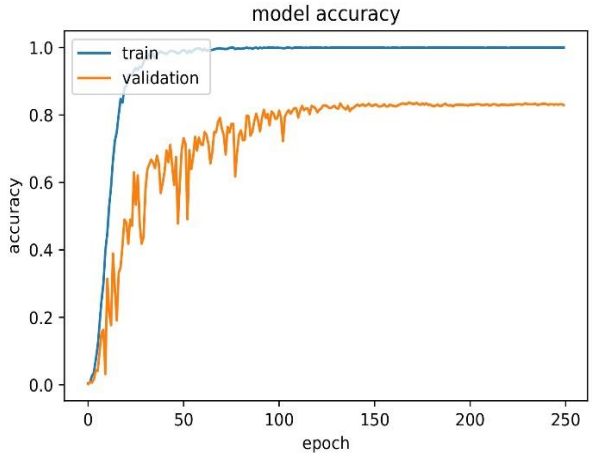


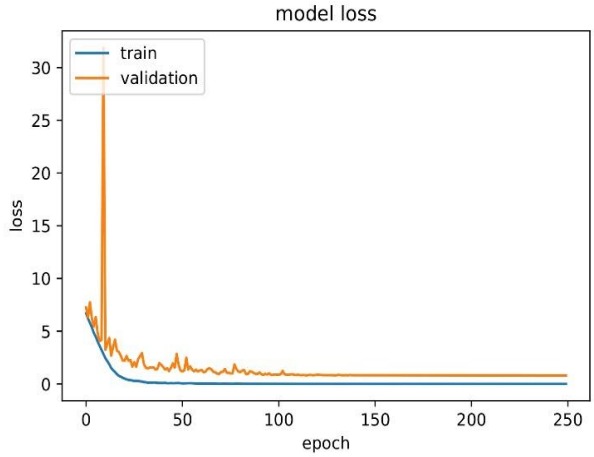
Figure 3. Accuracy and loss curves for the different CNN architectures on the AMI ear dataset

DenseNet201

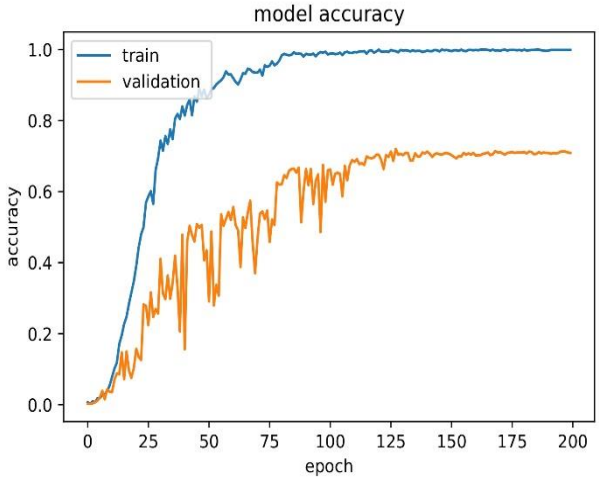


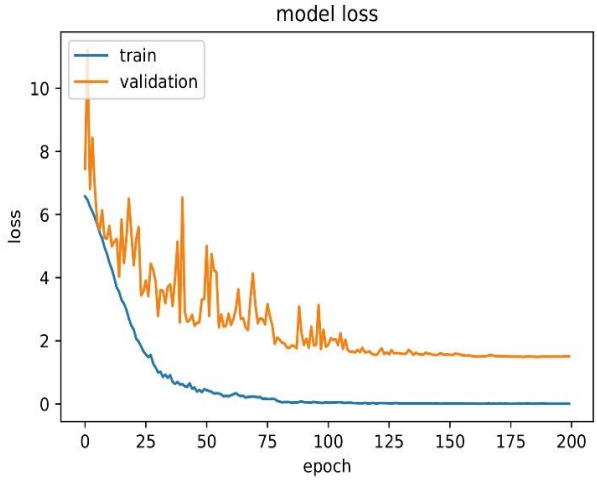
DenseNet121



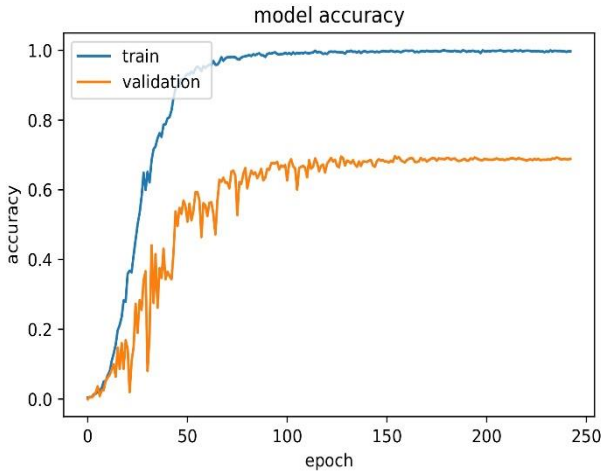


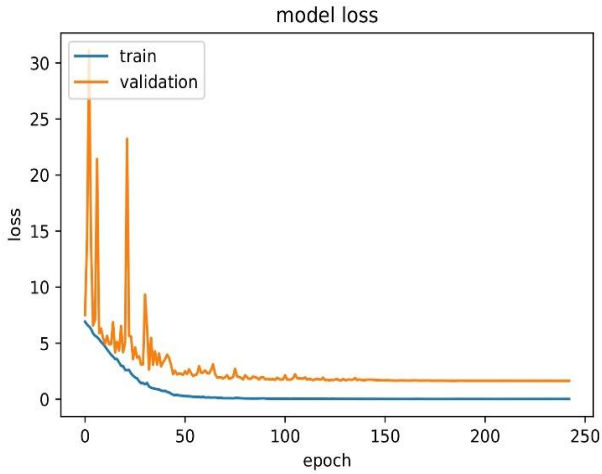
InceptionV3





ResNet50V2





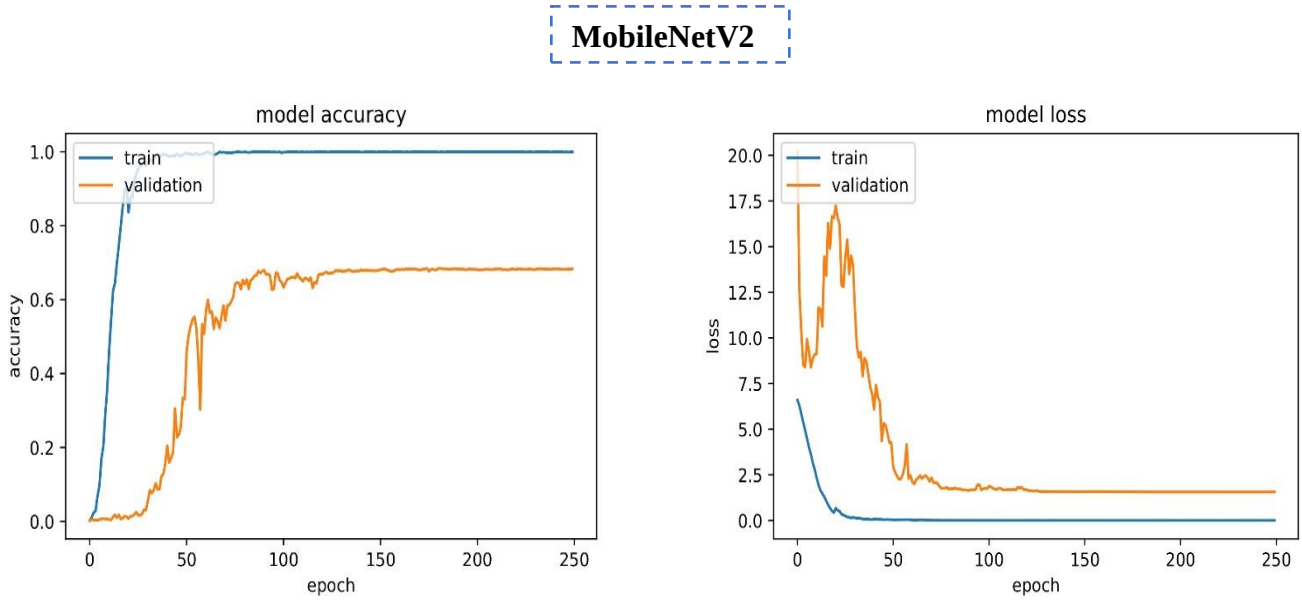


Figure 4. Accuracy and loss curves for the different CNN architectures on the AMI ear dataset

Table 3 summarizes the obtained results for the different CNN models on the AMI ear dataset. As can be seen from the table, all models obtain accuracy, precision, and F1-score above 92% and a recall score of 94%. In addition, the highest performance with respect to all evaluation metrics was achieved by the DenseNet variants.

Figure 4 illustrates the accuracy and loss curves for the CNN models used when training and validating on the WPUT ear dataset. Similarly, the performance of the models increases with time, indicating that the models are learning. We also observe that they improve at the beginning, but over time they reach a plateau. We notice the gap between the training and validation curves is a bit wider compared to those of the AMI dataset due to the wide variations encountered in the WPUT dataset. Moreover, the loss curves show a slight ups and downs at the begging of the learning process; however, over time, the loss gets smaller, indicating that the models are learning.

Table 4 presents the results for the different CNN models when conducting the experiments on the WPUT ear dataset. From the table, we observe that the best performance with respect to all evaluation metrics is again achieved by the DenseNet architecture.

Table 3. Results obtained from different deep CNN models on AMI dataset

	Accuracy	F1-Score	Recall	Precision
DenseNet121	95	95	96	95
ResNet50V2	93	92	94	93
InceptionV3	93	92	94	93
MobileNetV2	92	92	94	92
DenseNet201	96	95	97	95
ResNet101V2	93	92	94	93

Table 4. Results obtained from different deep CNN models on WPUT dataset

	Accuracy	F1-Score	Recall	Precision
DenseNet121	83	82	84	82
ResNet50V2	69	64	70	65
InceptionV3	71	67	72	68
MobileNetV2	68	65	71	67
DenseNet201	74	71	77	72

6.3 Grad-CAM visualization

In recognition systems, it is crucial to comprehend and interpret the performance and logic behind network decisions. In this section, we deal with the interpretation of what the CNN models have learned by highlighting the regions that the models consider for prediction. The main goal is to check whether the models trained for ear recognition actually focus on the ears or whether they also use other auxiliary textures or details such as hair or skin parts. This can be tackled with the help of Grad-CAM (Gradient-weighted Class Activation Mapping) [49] visualization technique. Grad-CAM highlights image regions that strongly contribute to making a specific decision, by providing a heatmap. The Grad-CAM algorithm uses the gradients of the final convolutional layer of the CNN with respect to the output class score to generate a heatmap that indicates the importance of each pixel in the input image. By visualizing the Grad-CAM heatmaps, we can gain insights into which parts of the ear image are most important for the recognition decision. For example, we may find that the network focuses on specific regions of the ear, such as the helix or the lobule, that contain distinctive features for different individuals. This can help us design more effective ear recognition systems by focusing on the most informative regions of the ear and improving the network's ability to capture these features.

To apply Grad-CAM to ear recognition systems, we first trained a CNN on a large dataset of ear images, and then used Grad-CAM to generate heatmaps that highlight the regions of the ear image that contribute the most to the network's decision. Tables 5 and 6 illustrate Grad-CAM visualization for various examples of ear images from the AMI and WPUT datasets, where the models correctly identified the subjects. Similarly, we show some cases of misclassified ear images to gain insights into false predictions. It can be seen that, making a correct recognition decision when the model concentrates on the ear's geometrical structure as the most discriminative region when the model focuses on the geometrical structure of the ear as the most discriminative region. However, when the models focus on textures at the ear boundary or hairstyle, it leads to a wrong identification decision.

Table 5. Grad-CAM localization to illustrate the important ear image regions for making a recognition decision on the AMI dataset

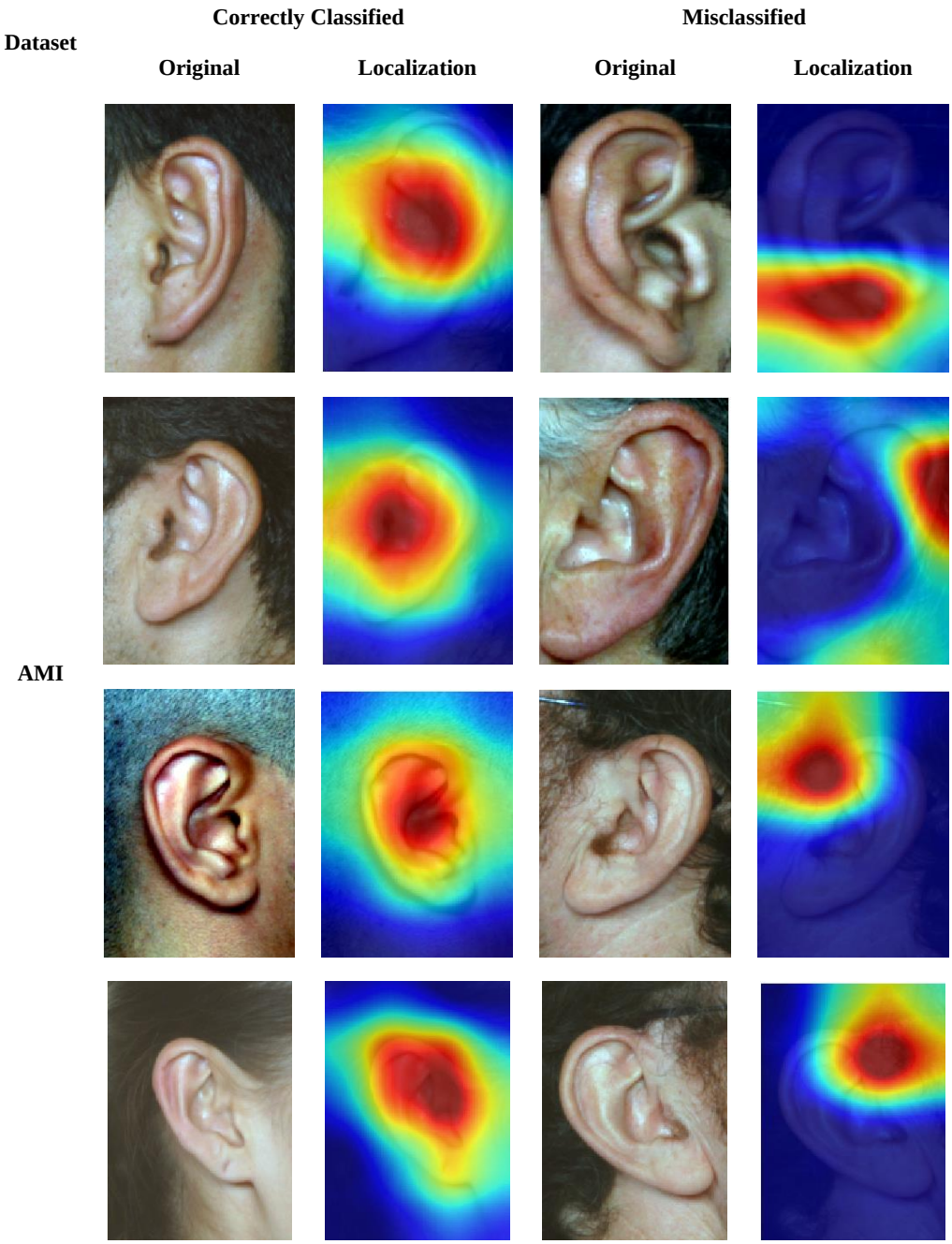
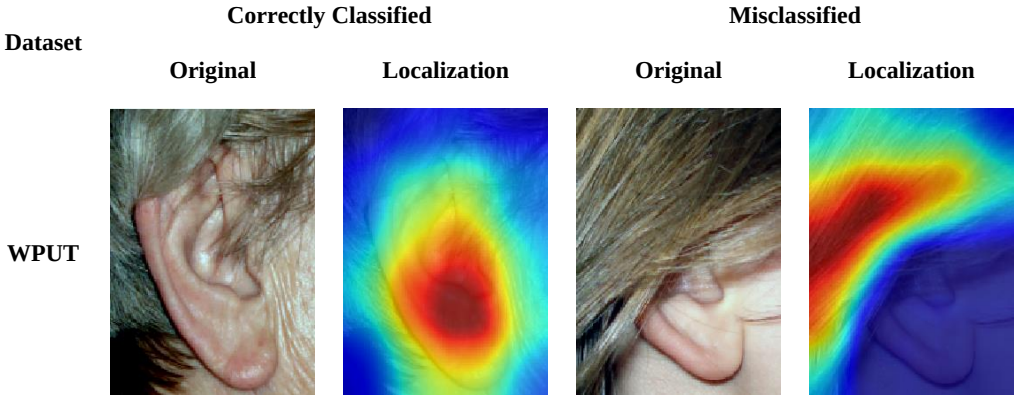
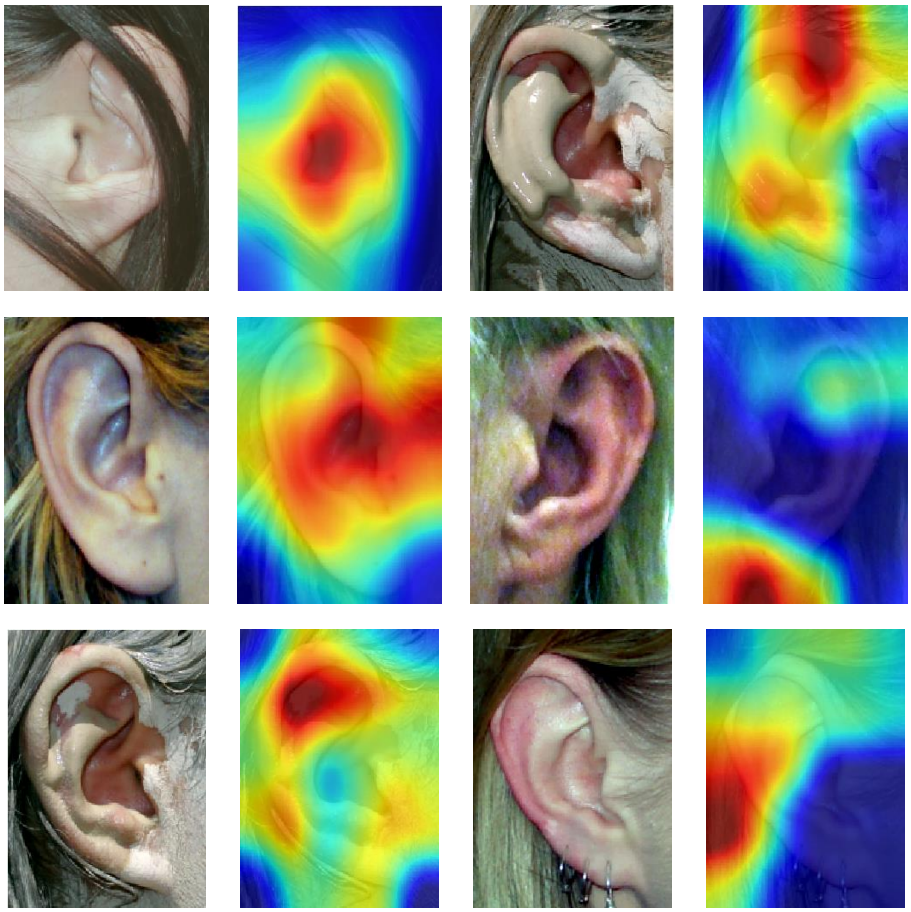


Table 6. Grad-CAM localization to illustrate the important ear image regions for making a recognition decision on the WPUT dataset





7. CONCLUSION

This work presented an ear recognition system based on state-of-the-art deep learning models. Different deep CNN architectures are utilized to improve the previous state-of-the-art results on the AMI and WPUT ear datasets. Due to the limited number of ear images required to train the models from scratch, we adopted a transfer learning strategy and fine-tuned a set of pretrained deep architectures to overcome the limited training ear images. The DenseNet architecture yielded the highest recognition rate on both AMI and WPUT datasets. To increase the interpretability of the proposed models and gain some insights on what the models have learned, Grad-CAM visualizations are provided, which highlight the important ear image regions the models consider for prediction. The models emphasize the pinna when making correct decisions. It also appears a convenient phenomenon that models can make correct predictions when focusing on the geometric structure of the ear, even though they are not constrained to utilize only these features. However, using deep neural networks for feature learning has enabled the automatic and robust feature extraction from raw ear images, and can capture subtle differences among individuals. However, there are still some challenges and limitations in this field, such as the lack of large-scale annotated datasets, the vulnerability to adversarial attacks, and the generalization to unseen domains. In our future research, we will focus on improving the recognition performance even further on the considered ear datasets especially the WPUT dataset. This will be addressed by exploring different learning strategies and building specific and more effective deep CNN models. Moreover, we plan to address the problem of unconstrained ear recognition using large-scale ear image datasets.

DATA AVAILABILITY

The AMI dataset to support the findings of this work is publicly available for download from the website: https://ctim.ulpgc.es/research_works/ami_ear_database. The clean WPUT dataset is available from the corresponding author upon request.

ACKNOWLEDGEMENTS

This research project was funded by the Deanship of Scientific Research, Princess Nourah bint Abdulrahman University, through the Program of Research Project Funding After Publication (Grant No.: 43-PRFA-P-29).

REFERENCES

[1] Jain, A.K., Ross, A., Prabhakar, S. (2004). An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology, 14(1): 4-20. https://doi.org/10.1109/TCSVT.2003.818349

[2] Hassaballah, M., Alshazly, H.A., Ali, A.A. (2019). Ear recognition using local binary patterns: A comparative experimental study. Expert Systems with Applications, 118: 182-200. https://doi.org/10.1016/j.eswa.2018.10.007

[3] Chang, K., Bowyer, K.W., Sarkar, S., Victor, B. (2003). Comparison and combination of ear and face images in appearance-based biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(9): 1160-

1165. https://doi.org/10.1109/TPAMI.2003.1227990
[4] Pflug, A., Paul, P.N., Busch, C. (2014). A comparative study on texture and surface descriptors for ear biometrics. In 2014 International Carnahan Conference on Security Technology (ICCST), Rome, Italy, pp. 1-6. https://doi.org/10.1109/CCST.2014.6986993
[5] 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
[6] Benzaoui, A., Hezil, N., Boukrouche, A. (2015). Identity recognition based on the external shape of the human ear. In 2015 International Conference on Applied Research in Computer Science and Engineering (ICAR), Beiriut, Lebanon, pp. 1-5. https://doi.org/10.1109/ARCSE.2015.7338129
[7] Hansley, E.E., Segundo, M.P., Sarkar, S. (2018). Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics, 7(3): 215-223. https://doi.org/10.1049/iet-bmt.2017.0210
[8] Emeršič, Ž., Meden, B., Peer, P., Štruc, V. (2020). Evaluation and analysis of ear recognition models: Performance, complexity and resource requirements. Neural Computing and Applications, 32: 15785-15800. https://doi.org/10.1007/s00521-018-3530-1
[9] Alshazly, H., Linse, C., Barth, E., Martinetz, T. (2019). Handcrafted versus CNN features for ear recognition. Symmetry, 11(12): 1493. https://doi.org/10.3390/sym11121493
[10] Oyebiyi, O.G., Abayomi-Alli, A., Arogundade, O.T., Qazi, A., Imoize, A.L., Awotunde, J.B. (2023). A systematic literature review on human ear biometrics: Approaches, algorithms, and trend in the last decade. Information, 14(3): 192. https://doi.org/10.3390/info14030192
[11] Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X.D. (2019). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11): 3212-3232. https://doi.org/10.1109/TNNLS.2018.2876865
[12] Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M., Lee, B. (2022). A survey of modern deep learning based object detection models. Digital Signal Processing, 126: 103514. https://doi.org/10.1016/j.dsp.2022.103514
[13] Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H. (2022). Deepfake detection: A systematic literature review. IEEE Access, 10: 25494-25513. https://doi.org/10.1109/ACCESS.2022.3154404
[14] Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., Alshehri, A.H. (2023). Deep fake detection and classification using error-level analysis and deep learning. Scientific Reports, 13(1): 7422. https://doi.org/10.1038/s41598-023-34629-3
[15] Alshazly, H., Linse, C., Abdalla, M., Barth, E., Martinetz, T. (2021). COVID-Nets: Deep CNN architectures for detecting COVID-19 using chest CT scans. PeerJ Computer Science, 7: e655. https://doi.org/10.7717/peerj-cs.655
[16] Alshazly, H., Linse, C., Barth, E., Martinetz, T. (2021). Explainable COVID-19 detection using chest CT scans and deep learning. Sensors, 21(2): 455. https://doi.org/10.3390/s21020455
[17] Kaushik, H., Singh, D., Kaur, M., Alshazly, H., Zaguia, A., Hamam, H. (2021). Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models. IEEE Access, 9: 108276-108292. https://doi.org/10.1109/ACCESS.2021.3101142
[18] Zahid, U., Ashraf, I., Khan, M.A., Alhaisoni, M., Yahya, K.M., Hussein, H.S., Alshazly, H. (2022). BrainNet: Optimal deep learning feature fusion for brain tumor classification. Computational Intelligence and Neuroscience, Article ID: 1465173. https://doi.org/10.1155/2022/1465173
[19] Pan, S.J., Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10): 1345-1359. https://doi.org/10.1109/TKDE.2009.191
[20] Abd Almisreb, A., Jamil, N., Din, N.M. (2018). Utilizing AlexNet deep transfer learning for ear recognition. In 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), Kota Kinabalu, Malaysia, pp. 1-5. https://doi.org/10.1109/INFRKM.2018.8464769
[21] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. https://doi.org/10.1145/3065386
[22] Mehta, R., Sheikh-Akbari, A., Singh, K.K. (2023). A noble approach to 2D ear recognition system using hybrid transfer learning. In 2023 12th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, pp. 1-5. https://doi.org/10.1109/MECO58584.2023.10154993
[23] Singh, S., Suman, S. (2022). Transfer learning: a way for ear biometric recognition. In 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India, pp. 1-6. https://doi.org/10.1109/I2CT54291.2022.9824374
[24] Alejo, M., Hate, C.P. (2019). Unconstrained ear recognition through domain adaptive deep learning models of convolutional neural network. International Journal of Recent Technology and Engineering, 8(2): 3143-3150. https://doi.org/10.35940/ijrte.B2865.078219
[25] Eyiokur, F.I., Yaman, D., Ekenel, H.K. (2018). Domain adaptation for ear recognition using deep convolutional neural networks. IET Biometrics, 7(3): 199-206. https://doi.org/10.1049/iet-bmt.2017.0209
[26] 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
[27] 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
[28] Abaza, A., Ross, A.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
[29] Wang, Z.B., Yang, J., Zhu, Y. (2021). Review of ear biometrics. Archives of Computational Methods in Engineering, 28: 149-180. https://doi.org/10.1007/s11831-019-09376-2
[30] Benzaoui, A., Khaldi, Y., Bouaouina, R., Amrouni, N., Alshazly, H., Ouahabi, A. (2023). A comprehensive survey on ear recognition: databases, approaches, comparative analysis, and open challenges.

Neurocomputing, 537: 236-270. https://doi.org/10.1016/j.neucom.2023.03.040
[31] Alshazly, H.A., Hassaballah, M., Ahmed, M., Ali, A.A. (2019). Ear biometric recognition using gradient-based feature descriptors. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Springer International Publishing, pp. 435-445. https://doi.org/10.1007/978-3-319-99010-1_40
[32] Benzaoui, A., Adjabi, I., Boukrouche, A. (2017). Experiments and improvements of ear recognition based on local texture descriptors. Optical Engineering, 56(4): 043109. https://doi.org/10.1117/1.OE.56.4.043109
[33] Hassaballah, M., Alshazly, H.A., Ali, A.A. (2020). Robust local oriented patterns for ear recognition. Multimedia Tools and Applications, 79: 31183-31204. https://doi.org/10.1007/s11042-020-09456-7
[34] Emeršič, Ž., Štepec, D., Štruc, V., Peer, P. (2017). Training convolutional neural networks with limited training data for ear recognition in the wild. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG), Washington, DC, USA, pp. 1-8. https://doi.org/10.1109/FG.2017.123
[35] 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
[36] 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
[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), Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90
[38] 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
[39] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308
[40] Xie, S.N., Girshick, R., Dollár, P., Tu, Z.W., He, K.M. (2017). Aggregated residual transformations for deep neural networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1492-1500. https://doi.org/10.1109/CVPR.2017.634
[41] 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
[42] 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
[43] 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), Honolulu, HI, USA, pp. 4700-4708. https://doi.org/10.1109/CVPR.2017.243
[44] Howard, A.G., Zhu, M.L., Chen, B., Kalenichenko, D., Wang, W.J., Weyand, T., Andreetto, M., Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv Preprint arXiv: 1704.04861. https://doi.org/10.48550/arXiv.1704.04861
[45] Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks?. Advances in Neural Information Processing Systems, 27: 1-9. https://doi.org/10.48550/arXiv.1411.1792
[46] Zhuang, F.Z., Qi, Z.Y., Duan, K.Y., Xi, D.B., Zhu, Y.C., Zhu, H.S., Xiong, H., He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1): 43-76. https://doi.org/10.1109/JPROC.2020.3004555
[47] Gonzalez, E. (2008). AMI ear dataset. https://webctim.ulpgc.es/research_works/ami_ear_database/.
[48] 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, Springer Berlin Heidelberg, pp. 227-234. https://doi.org/10.1007/978-3-642-13775-4_23
[49] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 618-626. https://doi.org/10.1109/ICCV.2017.74