Journal of Electrical Engineering, Vol. 76, No. 6, 2025, pp. 557-564
https://reference-global.com/journal/JEE

A multiple datasets deep learning approach for kinship recognition from ear images

Veronika Kurilova[1,2], Martin Bartos[1], Milos Oravec[1], Jarmila Pavlovicova[1]

In kinship recognition tasks, the face has traditionally been the primary biometric modality; in contrast, kinship recognition using ear images remains a largely unexplored area. However, ear width remains relatively constant throughout adulthood. Therefore, the ear can be considered a more stable biometric modality than the face. In our method, we used a Siamese network with convolutional sub-networks to extract the features from the ear images (VGG-16, ResNet-50, ResNet-152, and ConvNeXt-T). The extracted features are subsequently passed to a Siamese network classifier, which determines whether a kinship relation exists between the two input images. In our work, we used multiple datasets, including KinEar, Meng, EarKinshipVN, as well as modifications of the hyperparameters. Our results outperformed the method of Dvoršak et al., with the leading ResNet-152 and ConvNeXt-T as convolutional sub-networks.

Keywords: kinship recognition, computer vision, Siamese network, image recognition, deep learning, multiple-datasets

1 Introduction

In kinship recognition tasks, the face has traditionally been the primary biometric modality, as individuals tend to focus on distinct facial features. Challenges in kinship recognition often arise due to variations in age and gender between individuals, making it generally easier for humans to identify kinship between siblings, especially those of the same sex [1]. The face is also a more easily detectable modality, supported by a larger number of publicly available datasets, as well as a greater volume of data in terms of family structures, relationships, and captured images. In contrast, kinship recognition using ear images remains a largely unexplored area.

The structure of the human ear begins to develop between the fifth and seventh week of gestation and exhibits linear growth during the first four months after birth. Between four months and eight years of age, ear growth occurs at a quadrupled rate, then slows down, with a subsequent increase observed from approxi-mately the age of seventy. However, ear width remains relatively constant throughout adulthood [2, 3]. The ear can therefore be considered a more stable biometric modality than the face, as it is not affected by facial expressions, which tend to change significantly over time [4]. However, it is important to understand how different ear recognition models behave when applied to challenging data captured in unconstrained environ-ments [5].

In the image of the ear, we observe the external ear, which consists of the auricle and the external auditory canal. The auricle, except for the lobule, is composed of cartilage. The outer rim of the auricle curves into a helical structure known as the helix, beginning approximately in the middle of the ear at the point called the crus helicis and ending at the lobule (lobulus). On the outer surface of the auricle, the antihelix runs parallel to the helix. The area separating the crus helicis from the inner auricular cartilage structure, the antihelix, is called the concha (Fig. 1). These individual cartilaginous formations of the auricle vary in shape, relative position, and appearance among individuals. In a single person, the left and right auricles are only negligibly different and, in some cases, nearly symmetric. A certain degree of simi-larity is also assumed among biological relatives.

In 2000, the ear was first evaluated as a biometric modality in the field of computer vision. The authors of the study examined the ear in terms of measurability, uniqueness, and consistency. Ear detection is a funda-mental component of an automatic ear kinship recognition system. This process is challenging due to variations in lighting conditions, ear position, and size within the image. In addition, the ear can be covered with long hair, head coverings, or partially covered by earrings or piercings [4]. Detection methods vary and can be categorized into traditional template matching methods, geometric approaches, morphological trans-formations, face geometry-based techniques, and, more

[1] Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava, Ilkovicova 3, 812 19, Bratislava, Slovak Republic
[2] Faculty of Medicine, Slovak Medical University, Limbova 12, 833 03 Bratislava, Slovak Republic
Corresponding author: veronika.kurilova@stuba.sk

https://doi.org/10.2478/jee-2025-0058, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

recently, machine learning-based and deep learning methods.

The next section, Related works, describes current solutions in this field, followed by Methodology describing our proposed method. Finally, the Results and analysis section and the Conclusion contain our achievements and future plans.

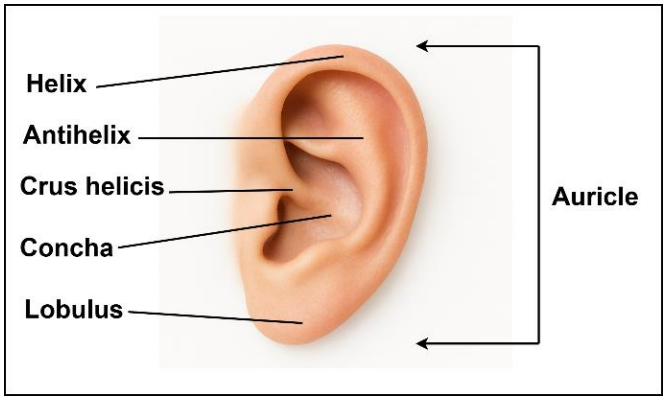

Fig. 1. Parts of the ear

2 Related works

The study by Meng et al. [6] was the first in the literature to address the recognition of kinship based on ear images. The authors developed a geometric model of the auricle, based on the distances between nine distinct landmarks on the ear. For ear detection, they employed the Hough transform, using the Canny edge detector to identify the highest point on the helix and the lowest point on the earlobe. For classification into gender and kinship categories, they applied the k-nearest neighbors (kNN) algorithm. The best-performing model achieved an accuracy of 67.2%. The authors also created their own publicly available dataset of 142 images of 71 people, consisting of 21 families.

The study by Dvoršak et al. [7] was the first to apply convolutional neural networks for kinship recognition based on ear images. The authors introduced the first larger publicly available ear dataset focused on kinship, called KinEar. For parallel processing of ear pairs, a Siamese architecture was employed, utilizing various networks as feature extractors: VGG-16, ResNet-152, ResNet-50, Attentional Feature Fusion (AFF), and Contextual Transformer Network (CoT-Net). The best performance was achieved using VGG-16 as the feature extractor, with an accuracy of 64.01% and an AUC ROC of 0.6922.

One of the most recent studies by Luu et al. [8] in the field of ear-based kinship recognition is based directly upon the work of Dvoršak et al. The authors employed a similar Siamese architecture with pre-trained convolutional neural networks, including VGG-16, ResNet-50, ResNet-152, DenseNet-121, DenseNet-169, and vision transformers ViT-b-16 and ViT-b-32. All networks were pre-trained on the biometric dataset EarVN1.0 [9]. A new kinship-focused ear image dataset was created, with samples collected from the Internet. This dataset, temporarily named KinEarVN, was collected from open sources and social media, and is approximately three times larger in terms of family units compared to the original KinEar dataset. For training the Siamese network with the pre-trained models, both the KinEar and KinEarVN datasets were used. The best performance was achieved with DenseNet-169, reaching 98.76% and 94.19% accuracy on the KinEar and on the larger KinEarVN dataset, respectively.

The last study by Cao et al. [10] followed the work from the same university as [8]. They introduced the EarKinshipVN (previously known as KinEarVN) dataset, containing 4876 images of 156 families, The authors verified kinship via ear image pairs using feature extractors that were fed into a similarity classifier. Authors tested convolutional, transformer-based, and mixed architectures as feature extractor backbones, with the best results being 98.71% accuracy of the Mixer Attention Inception model (MAI). In this work, because of smaller images lacking sufficient detail, super-resolution and image restoration techniques based on transformer architecture were applied.

The aim of our work was to develop a deep learning method based on a Siamese architecture, capable of distinguishing between related and unrelated pairs of ear images, and test our proposed model on the combination of the publicly available datasets KinEar, Meng, and EarKinshipVN ear datasets.

3 Methodology

This section describes the methodological framework of the proposed method, which consists of image datasets description and preprocessing, description of the used deep learning model architectures, model design and training, and model evaluation.

After preparing and augmenting the data, we evaluate whether two individual inputs exhibit kinship using the Siamese architecture. We extracted features

from each input image using deep neural networks with shared weights. Several architectures were employed for feature extraction, including VGG-16, ResNet-50, ResNet-152, and ConvNeXt-T. The extracted features are subsequently passed to a similarity classifier, which determines whether a kinship relation exists between the two input images. The overall methodology is illustrated in Fig. 2.

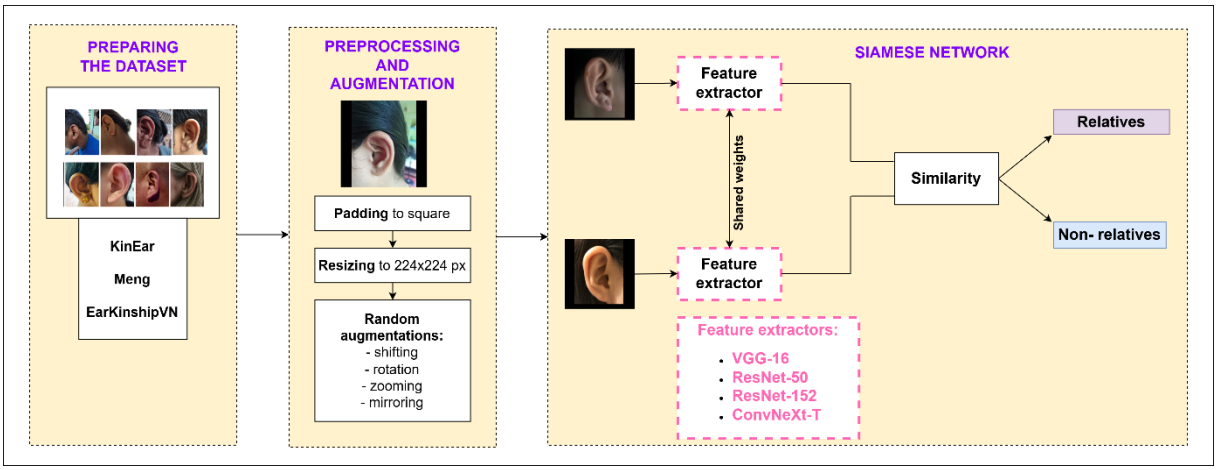

Fig. 2. Overview of the proposed methodology of kinship recognition based on ear images

3.1 Datasets

We used KinEar, Meng, and EarKinshipVN datasets in our study. Examples of images are shown in Fig. 3.

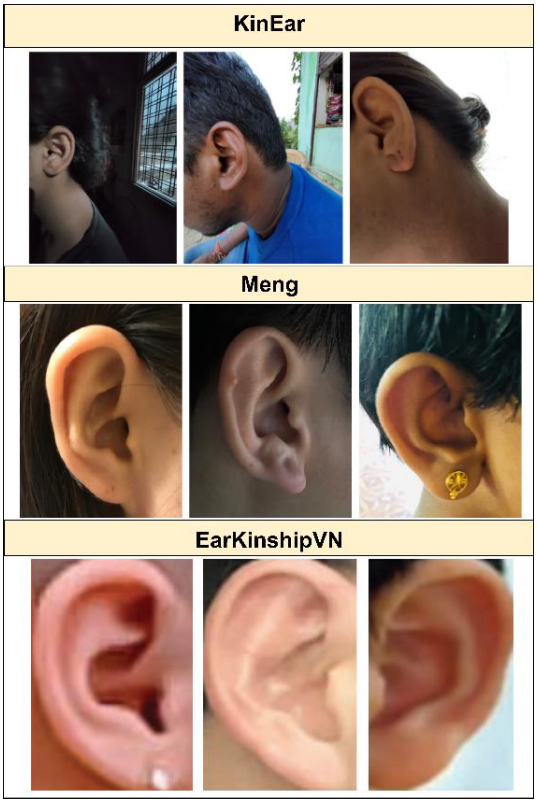

Fig. 3. Examples of the datasets used

KinEar dataset [7] contains 1477 ear images of 76 persons from 19 families. There are seven kinship possibilities between family members, mostly parents and children. The dataset is heterogeneous, with various lightning conditions, ear sizes, their positions on an image, and different capturing angles. Some images contain ears partially covered with hair or earrings. According to the dataset description, it includes the exact location of the ear on the image, described by the coordinates of the bounding box. According to this, the authors ensured that the ear region in the image was always at least 250×250 pixels in size to provide an appropriate image size for deep learning training. The dataset was reduced according to Dvoršak et al. [7] and they removed 436 images with insufficient lighting, blurred, or images where the ear is not clearly visible. After this reduction, the dataset contains 1041 images, still some images were not of excellent quality, but those are in limited quantity. From the cleaned version of the dataset, no family or subject was excluded, thus preserving the number of kinship relationships.

The Meng dataset [6] contains 142 images of 71 persons from 21 families, mostly images of parents and their children, with a limited number of sibling images. This dataset is more homogeneous – the ears are mostly captured under similar angles, centered in the image. Some of the images have different lighting conditions and include earrings. The original resolution of the images, except for one, is 300×400 pixels.

The Meng dataset is not organized by families, but by left and right ears, so we had to completely reorganize it by families and subjects to ensure an identical directory structure to the KinEar database. Since no file indicating family relationships between subjects was provided with the dataset, it was necessary to create one as a reference for generating related and unrelated pairs. We therefore generated a JSON file in which the main identifier of each person is the name of the folder containing the ear samples for that individual. Each person's record also includes a list of individual images and a list of relatives, where each relative is assigned an identifier and a numerical label ranging from 0 to 6, representing the relationship of the person to those relatives. The generated reference relationship file was then used to create positive (related) and negative (unrelated) image pairs with the corresponding relationship label.

EarKinshipVN dataset [10] (previously named KinEarVN) contains 4876 images of 498 persons from 156 families. It was created in the Faculty of Information Technology, Ho Chi Minh City Open University in Vietnam. The images are mostly smaller than 224×224, although the authors applied the Hybrid Attention Transformer (HAT) to super-resolution and image restoration in their original article when presenting this dataset. This dataset emphasizes a diverse range of racial backgrounds and ages.

Description of all three datasets is provided in Tab. 1.

Table 1. Description of all three datasets used

Dataset	Nº of images	Nº of families	Nº of persons	Image size (px)
KinEar [7]	1477	19	76	250×250*
KinEar- reduced [7]	1041	19	76	250×250*
Meng [6]	142	21	71	300×400
EarKinshipVN [10]	4876	156	498	< 224×224
Together	6059	196	645	-

*minimal ear resolution, according to the authors

By merging KinEar-reduced, Meng, and EarKinshipVN databases, the number of families increased to 196, from which we randomly selected 20 families for the validation set and 20 for the test set. We then generated all possible related pairs within each subset, along with an equal number of unrelated pairs. Of all generated pairs, the training set contains 78.69% (144,240) pairs, the validation set 6.42% (11,768) pairs, and the test set 14.89% (27,298) pairs.

The pairs were divided into training, validation, and test sets based on the families to which the subjects belonged, ensuring that each subset contained a certain number of related and unrelated subjects. If the data were split solely according to the number of related and unrelated pairs in each subset, it would be easy to create a situation where samples from the same individuals appear in both the training and validation or test data.

3.2 Image preprocessing and augmentation

Image preprocessing consisted of resizing the samples to 224×224 pixels, while preserving their original aspect ratio, as simple resizing can distort the shape of the ear. If empty space remained after resizing, it was filled with black color.

Data augmentation was performed using augmentation layers from the Keras library and included:

- Random shifting of the image in all directions with a maximum shift of 15% of the sample's height and width
- Random rotation in both directions within the range of 0 to 67.5 degrees
- Random zooming in or out in both height and width, with a maximum shift of 15% of the sample's height and width
- Random horizontal flipping with a 50% probability

After applying augmentations, any remaining empty space in the sample image was filled with black color, consistent with the preprocessed dataset.

Data normalization differs for the ConvNeXt network, where the input data must remain in the original 0-255 range after loading and should not be normalized using a function, as the required pre-processing steps are already integrated into the network's architecture.

3.3 Architectures of the used deep learning models

In the following subsections, we present the architectures of the deep neural networks employed in the proposed method.

Siamese architecture

The Siamese architecture [11] derives its name from Siamese twins, who are physically connected from birth. Similarly, in a Siamese network, two identical neural networks are interconnected, sharing their weights and thus functioning as twins. This architecture was first applied to kinship recognition based on facial photographs [12], and later extended to ear images [7]. In these cases, a Siamese convolutional neural network was employed, where the subnetworks – i.e., the twin networks – are convolutional structures suitable for image processing. The convolutional subnetworks process these images and produce their low-dimensional representations, or feature vectors. These feature vectors are then compared within the comparison module of the Siamese network, where, for example, a similarity metric based on Euclidean distance is computed according to the chosen contrastive loss function ($L_{\text{contrastive}}$). This function minimizes the distance between the feature vectors of samples belonging to the same class, while increasing their distance when the samples belong to different classes [13].

$$L_{\text{contrastive}} = Y\,D_{euclid}^2 + (1 - Y)\,[max\,(0,\, m - D)]^2 \quad (1)$$

where D_{euclid} is a similarity metric

$$D_{\text{euklid}} = \|f_1\,(X_1) - f_2\,(X_2)\|_2 \quad (2)$$

Different deep learning models have been used as convolutional twin networks:

VGG-16

VGG [14] is a family of convolutional neural networks that were the first to demonstrate that network depth is crucial for achieving higher performance. The VGG architecture builds upon the AlexNet network, which was the first convolutional neural network to achieve first place in the ImageNet classification challenge, thereby redirecting the focus of artificial intelligence research toward deep neural networks. VGG improves upon AlexNet by more than doubling the number of convolutional layers and reducing the filter size from the original 7×7 to 3×3, effectively decreasing the number of parameters. The network alternates between convolutional and max-pooling layers. In the fully connected layers, the regularization method Dropout is applied with a probability of 0.5. The convolutional layers use filters of size 3×3 with a stride of 1 and zero padding to preserve spatial dimensions. Max-pooling layers employ filters of size 2×2 with a stride of 2, which progressively reduces the spatial resolution by half, while the number of channels doubles after each convolutional block. All convolutional layers are followed by ReLU activation functions [14].

ResNet variants (50 and 152 layers)

ResNet [15] is a class of convolutional neural networks that introduced residual blocks, enabling the construction of substantially deeper architectures with hundreds of convolutional layers. The number following the model name indicates the total number of convolutional layers (for example, ResNet-50 contains 50 convolutional layers). Earlier networks, such as VGG and AlexNet, relied on sequential data processing, where information flows linearly through the network, with each layer connected in a single, continuous chain. Although the use of He initialization and batch normalization mitigates the problems of vanishing and exploding gradients, performance metrics begin to degrade once a certain network depth is reached. This degradation affects both the training and testing datasets, ruling out overfitting as a cause. The phenomenon is referred to as fragmented gradients. The impact of fragmented, vanishing, and exploding gradients is alleviated in ResNet by the introduction of residual or shortcut connections within the residual blocks [15].

ConvNeXt

The authors of the ConvNeXt [16] network integrated selected components from Vision Transformers into convolutional architectures, specifically ResNet-50 and ResNet-200. They adopted the training strategies used in the Swin and DeiT Vision Transformers and modified the network structure by adjusting the number of residual blocks. In Vision Transformers, the stem layer performs a stronger spatial downsampling of the input image compared to residual networks. The ConvNeXt design is also inspired by the ResNeXt architecture and its modified residual blocks, where a grouped convolution layer divides the input channels into groups that are processed separately and then concatenated back into the original dimensionality. However, the authors employed a specialized variant known as depthwise convolution, in which the number of input channels is also divided, but the number of groups is equal to the number of input channels. This layer is placed at the center of an inverted bottleneck residual block, where

the first convolutional layer increases the number of channels fourfold, and the final convolutional layer restores the original number of channels. Furthermore, all ReLU activation functions were replaced with a single GELU (Gaussian Error Linear Unit) activation, which is based on the cumulative distribution function of the normal distribution and provides a smoother, more continuous version of ReLU. Additionally, all Batch Normalization layers within the blocks were replaced with a single Layer Normalization layer. Unlike Batch Normalization, which normalizes activations using the mean and variance across the entire mini-batch, Layer Normalization performs normalization based on the mean and variance of each individual sample. ConvNeXt-T blocks are developed according to this principle [16].

3.4 Model training and evaluation

ResNet-152 and ConvNeXt-T were pretrained on the ImageNet dataset, ResNet-50 was pretrained on the VGGFace2 database [17], and the VGG-16 network was pretrained on the VGGFace database [18]. All layers of the convolutional sub-networks were fully trainable. The multi-dimensional feature tensors produced by the convolutional twin networks are transformed by the Siamese network into vectors using a Global Max Pooling layer and subsequently concatenated into a single comprehensive feature vector. This vector is processed by a fully connected layer with 192 neurons and a ReLU activation function, followed by a batch normalization layer, which we positioned between the ReLU activation and the fully connected layer, and a Dropout layer with a dropout rate of 0.4. The output of the network is a fully connected layer with a single neuron using a Sigmoid activation function.

During the training process, we employed the binary cross-entropy loss function with a learning rate of 7×10^{-6} using the Adam optimization algorithm. The network was trained for a total of 120 epochs, with each epoch consisting of 512 training steps and 256 validation steps. The batch size for both training and validation was set to 32, with alternating positive and negative pairs, and a generator shuffled the order of data batches after each epoch. The learning rate was gradually reduced using the *ReduceLROnPlateau* function, which decreased the learning rate by one order of magnitude when the validation loss failed to improve for 5 consecutive epochs. We also incorporated an EarlyStopping mechanism configured to stop training if the validation loss did not reach a new minimum within 10 epochs.

During network training, batches were generated by two types of generators. For the training and validation sets, the first generator produced batches through random selection of various related and subsequently unrelated pairs. The second generator alternated related and unrelated samples according to a predefined order of the test data.

4 Results and analysis

This section presents the results of our proposed method described in the Methodology section.

According to ROC-AUC, accuracy, and the balance between sensitivity and specificity, ResNet-152 and ConvNeXt-T were the most successful convolutional sub-networks used in our proposed method (Tab. 2). VGG-16 was the weakest sub-network, achieving the lowest sensitivity among all experiments, as well as the lowest accuracy and ROC-AUC values. This can be explained by the architectural limitations of VGG-16, which make it more prone to vanishing gradients and suboptimal feature learning.

AUC-ROC curves and confusion matrices of the proposed method, according to the convolutional sub-network used, are depicted in Figs. 4 and 5.

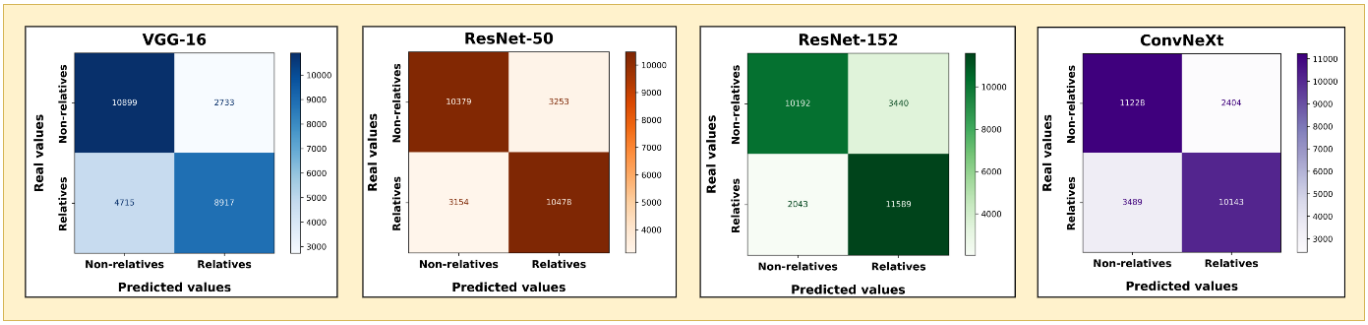

Fig. 4. Confusion matrices of the proposed method according to convolutional sub-networks

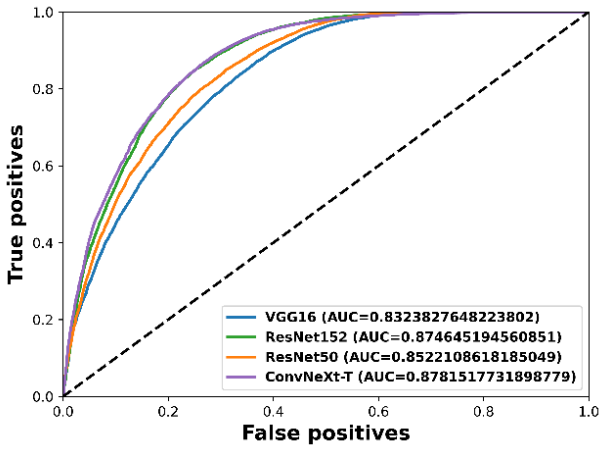

Fig. 5. AUC-ROC curves of the proposed method according to convolutional sub-networks

The comparison between our method and the results of Dvoršak et al. [7] is listed in Tab. 2. Our proposed method outperformed Dvoršak et al. Adding new databases altogether, along with changing the training parameters, significantly improved the results. We must also note that our method did not use any detection bounding boxes, which we did not have available at the time we downloaded the KinEar dataset.

We also include the results of Luu et al. [8] and Cao et al. [10] in our comparison Tab. 2, which are significantly better. The main ambiguity in their methodological description concerns the procedure used to split the EarKinshipVN dataset (previously named KinEarVN) into training, validation, and test subsets. The authors state that all image pairs were shuffled before the division, but do not clarify whether the split was performed at the family level. Without this information, we cannot compare our method with their results. Cao et al.[10] also applied super-resolution and image-restoration methods using the Hybrid Attention Transformer (HAT) model, which may improve the results. However, we did not apply it when using the EarKinshipVN dataset.

The study of Meng et al. [6] is not comparable to our approach – they used the k-nearest neighbor algorithm (kNN) with different distance measures (Euclidean, Manhattan, and Mahalanobis).

We do not include the results of transformer-based architectures in the comparison with other works, as we focused on convolutional-based architectures and their potential to improve performance on multiple datasets.

Table 2. Comparison of our proposed method with other authors, with our best-performing variants marked in bold

Sub-network	Accuracy	Sensitivity	Specificity	ROC-AUC
Dvoršak et al. [7] *on the KinEar dataset*				
VGG-16	64.01%	64.1%	64.01%	0.6922
ResNet-152	57.50%	57.5%	57.51%	0.6314
USTC-NELSLIP	55.12%	55.14%	55.10%	0.5729
Our proposed method on multiple datasets				
VGG-16	72.68%	65.41%	79.95%	0.8324
ResNet-152	**79.89%**	**85.01%**	**74.77%**	**0.8746**
ResNet-50	76.50%	76.86%	76.14%	0.8522
ConvNeXt-T	**78.39%**	**74.41%**	**82.37%**	**0.8782**
Luu et al. [8] *on the KinEarVN dataset*				
VGG-16	74.39%	N/A	N/A	N/A
ResNet-50	89.38%	N/A	N/A	N/A
ResNet-152	91.36%	N/A	N/A	N/A
DenseNet-121	88.40%	N/A	N/A	N/A
DenseNet-169	94.19%	N/A	N/A	N/A
ViT-b-16	80.85%	N/A	N/A	N/A
ViT-b-32	75.79%	N/A	N/A	N/A
Cao et al. [10] *on the EarKinshipVN (previously KinEarVN) dataset*				
VGG-16	92.89%	93.15%	N/A	N/A
VGG-19	95.54%	97.74%	N/A	N/A
ResNet-101	74.78%	86.14%	N/A	N/A
ResNet-152	70.16%	86.81%	N/A	N/A
DenseNet-121	98.59%	99.08%	N/A	N/A
DenseNet-169	98.44%	98.49%	N/A	N/A
Inception-V4	89.46%	95.88%	N/A	N/A
EfficientNet-b6	84.14%	94.90%	N/A	N/A
EfficientNet-v2m	85.10%	97.51%	N/A	N/A

5 Conclusion

Significant improvements over the work of Dvoršak et al. were observed in experiments where we assume the beneficial effect of additional high-quality data from multiple datasets, as well as from increasing the number of steps per epoch and the batch size, which exposed the networks to a larger amount of data during training.

For future research, we will focus on pretraining our feature extractors on more extensive biometric ear datasets, increasing the number of both high-quality and low-quality ear samples in family datasets, and employing an ear detector for automatic ear localization and detection. Testing the transformer-based architectures on multiple datasets might also be successful.

Acknowledgement

This work was done within the grant VEGA 1/0202/23 AIDabiomeDIA.

References

[1] X. Wu *et al.*, "Facial Kinship Verification: A Comprehensive Review and Outlook," *Int J Comput Vis*, vol. 130, no. 6, pp. 1494–1525, June 2022, doi: 10.1007/s11263-022-01605-9.

[2] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM Comput. Surv.*, vol. 45, no. 2, p. 22:1-22:35, Mar. 2013, doi: 10.1145/2431211.2431221.

[3] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, Sept. 2017, doi: 10.1016/j.neucom.2016.08.139.

[4] A. Benzaoui, Y. Khaldi, R. Bouaouina, N. Amrouni, H. Alshazly, and A. Ouahabi, "A Comprehensive survey on ear recognition: Databases, approaches, comparative analysis, and open challenges," *Neurocomputing*, vol. 537, pp. 236–270, June 2023, doi: 10.1016/j.neucom.2023.03.040.

[5] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc, "Evaluation and analysis of ear recognition models: performance, complexity and resource requirements," *Neural Comput & Applic*, vol. 32, no. 20, pp. 15785–15800, Oct. 2020, doi: 10.1007/s00521-018-3530-1.

[6] D. Meng, M. Nixon, and S. Mahmoodi, "Gender and Kinship by Model-Based Ear Biometrics," *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, July 2019, Accessed: Nov. 17, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Gender-and-Kinship-by-Model-Based-Ear-Biometrics-Meng-Nixon/8968e30050d9562ee429dfd60a8e512b22a1902f

[7] G. Dvoršak, A. Dwivedi, V. Štruc, P. Peer, and Ž. Emeršič, "Kinship Verification from Ear Images: An Explorative Study with Deep Learning Models," in *2022 International Workshop on Biometrics and Forensics (IWBF)*, Apr. 2022, pp. 1–6. doi: 10.1109/IWBF55382.2022.9794555.

[8] P. Q. Luu, B. V. Nguyen, and H. Q. Nguyen, "Kinship verification via ear images: A comparative study," *Ho Chi Minh City Open University Journal of Science – Engineering and Technology* pp. 47–57, Jan. 2025, doi: 10.46223/HCMCOUJS.tech.en.15.1.3683.2025.

[9] V. T. Hoang, "EarVN1.0: A new large-scale ear images dataset in the wild," *Data in Brief*, vol. 27, p. 104630, Dec. 2019, doi: 10.1016/j.dib.2019.104630.

[10] T.-T. Cao, H.-T. Duong, V.-T. Le, H. Trung, V. Hoang, and K. Tran-Trung, "Enhanced Kinship Verification through Ear Images: A Comparative Study of CNNs, Attention Mechanisms, and MLP Mixer Models," *CMC*, vol. 83, no. 3, pp. 4373–4391, 2025, doi: 10.32604/cmc.2025.061583.

[11] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proceedings of the 7th International Conference on Neural Information Processing Systems*, in NIPS'93. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Nov. 1993, pp. 737–744.

[12] J. P. Robinson, M. Shao, and Y. Fu, "Survey on the Analysis and Modeling of Visual Kinship: A Decade in the Making." Accessed: Nov. 17, 2025. [Online]. Available: https://www.computer.org/csdl/journal/tp/2022/08/09367013/1rDQQTlFWAU

[13] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, E. A. Chavez-Urbiola, and J. A. Romero-Gonzalez, "Loss Functions and Metrics in Deep Learning," *Artif Intell Rev*, vol. 58, no. 7, p. 195, Apr. 2025, doi: 10.1007/s10462-025-11198-7.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[16] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," Mar. 02, 2022, *arXiv*: arXiv:2201.03545. doi: 10.48550/arXiv.2201.03545.

[17] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," In: 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), https://www.robots.ox.ac.uk/~vgg/publications/2018/Cao18/

[18] O. M. Parkhi, A. Vedaldi, A. Zisseman, "Deep face recognition", In: BMVC, Proceedings of the British Machine Vision Conference 2015, British Machine Vision Association, https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf

Received 12 September 2025