Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061583

ARTICLE

Enhanced Kinship Verification through Ear Images: A Comparative Study of CNNs, Attention Mechanisms, and MLP Mixer Models

Thien-Tan Cao, Huu-Thanh Duong, Viet-Tuan Le, Hau Nguyen Trung, Vinh Truong Hoang and Kiet Tran-Trung*

Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh, 722000, Vietnam

*Corresponding Author: Kiet Tran-Trung. Email: kiet.tt@ou.edu.vn

Received: 28 November 2024; Accepted: 21 March 2025; Published: 19 May 2025

ABSTRACT: Kinship verification is a key biometric recognition task that determines biological relationships based on physical features. Traditional methods predominantly use facial recognition, leveraging established techniques and extensive datasets. However, recent research has highlighted ear recognition as a promising alternative, offering advantages in robustness against variations in facial expressions, aging, and occlusions. Despite its potential, a significant challenge in ear-based kinship verification is the lack of large-scale datasets necessary for training deep learning models effectively. To address this challenge, we introduce the EarKinshipVN dataset, a novel and extensive collection of ear images designed specifically for kinship verification. This dataset consists of 4876 high-resolution color images from 157 multiracial families across different regions, forming 73,220 kinship pairs. EarKinshipVN, a diverse and large-scale dataset, advances kinship verification research using ear features. Furthermore, we propose the Mixer Attention Inception (MAI) model, an improved architecture that enhances feature extraction and classification accuracy. The MAI model fuses Inceptionv4 and MLP Mixer, integrating four attention mechanisms to enhance spatial and channel-wise feature representation. Experimental results demonstrate that MAI significantly outperforms traditional backbone architectures. It achieves an accuracy of 98.71%, surpassing Vision Transformer models while reducing computational complexity by up to 95% in parameter usage. These findings suggest that ear-based kinship verification, combined with an optimized deep learning model and a comprehensive dataset, holds significant promise for biometric applications.

KEYWORDS: Biometric analytics; ear kin; Inceptionv4; kinship verification; kin; ear images

1 Introduction

Kinship verification is an emerging field of research with various practical applications. It plays a crucial role in the management of family albums, the analysis of social networks, and the search for lost family members. With the ever-increasing amount of photo data, it is essential to use related identities to organize family photos effectively. Kinship verification can also aid in forensic investigations and help locate missing children more accurately and quickly by utilizing various biometric systems [1]. Furthermore, by analyzing and identifying kinship relations, we can gain insights into behavior patterns and propose appropriate content based on information from other family members.

Kinship recognition research has advanced significantly, shaped by numerous influential studies. Fang et al. [2] pioneered the use of handcrafted feature descriptors, including Scale-Invariant Feature Transform (SIFT) [3], Local Binary Patterns (LBP) [4], and Gabor filters [5], in kinship verification. Wu et al. [6]

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

employed SIFT to locate facial feature points, while Qin et al. [7] leveraged SIFT-based feature extraction for multi-view learning. Additionally, Van et al. [8] combined LBP with Support Vector Machines (SVM) to improve recognition accuracy. More recently, Chouchane et al. [9] introduced the Hist-Gabor method, integrating Gabor features with deep learning architectures, demonstrating significant performance gains in kinship verification.

Later advancements in metric learning methods focus on refining distance metrics by enhancing inter-class separation and minimizing intra-class variation. These methods aim to bring samples of the same class closer together while pushing samples from different classes farther apart. To achieve this, a distance metric is learned to quantify similarity. Neighborhood Repulsed Metric Learning (NRML) [10] a widely adopted and influential approach in metric learning, with recent studies like Ramazankhani et al. [11] fusing features to enhance performance. NRML generates discriminative vectors, verified by an SVM classifier for kinship recognition. Huang et al. [12] proposed an innovative cross-pair metric learning framework, leveraging a refined k-tuplet loss to effectively capture both low-order and high-order discriminative features from multiple negative pairs, enhancing the model's ability to distinguish complex inter-class and intra-class variations.

A recent survey [13] of Wang et al. highlights that deep learning architectures outperform metric learning methods and handcrafted feature descriptors. Therefore, CNNs are widely adopted for their ability to extract meaningful feature representations from images. For example, Serraoui et al. [14] integrated CNNs with metric learning to achieve strong performance on the KinFace dataset, while Chen et al. [15] designed a shared-parameter deep neural network, reducing model complexity without compromising performance. Beyond CNNs, Li et al. [16] proposed a Graph-Based Kinship Reasoning (GKR) network that applies relational reasoning on extracted features using Graph Neural Networks. Autoencoders, similar to CNNs, effectively retain essential genetic features for kinship recognition. Dehghan et al. [17] proposed gated autoencoders with a discriminative neural network layer, while Wang et al. [18] developed a deep kinship verification model incorporating metric learning.

CNN architecture [19] has been widely adopted in image processing tasks due to its ability to effectively handle spatial and sequential data. Among these, Inception v4 [20] refined the design further with minor modifications to the module and stem layers, leading to improved accuracy and efficiency. However, CNNs have an inherent limitation: they are prone to information loss, as highlighted in prior research [21]. To address this issue, hybrid architectures such as CoAtNet [22] have emerged, combining CNNs and Transformers to balance complexity with the integration of local and global receptive fields. Although transformers offer robust performance, their complexity may be unnecessary for tasks involving relatively simple data, such as ear images. Alternatively, Mixer architectures present a compelling solution by providing a global receptive field [23], effectively overcoming the limitations of CNNs while maintaining lower computational complexity compared to Transformers.

Driven by the previously discussed insights, we propose a new and improved model that combines the Inception v4 architecture [20], attention modules, and MLP Mixer [24]. Our objective is to extract highly descriptive image representations with rich semantic information for a variety of visual tasks. This study compares the performance of our model with other prominent CNN models, evaluating their strengths and weaknesses based on the methodology outlined by Dvoršak et al. [25]. Additionally, we explore our model's focus on ear images using Grad-CAM [26], addressing a previously unexplored area by identifying the specific ear regions most critical for recognition.

As mentioned earlier, there is a scarcity of datasets specifically designed for this task. In a study by Dvoršak et al. [25], the KinEar dataset was created to benchmark this problem. However, the dataset's small

size and limited generalizability present significant challenges. To address these limitations, we propose a novel, larger dataset called EarKinshipVN.

In summary, the main contributions of this paper are two-folds:

- We have constructed a novel dataset of ear images to evaluate the fairness of kinship verification. This dataset integrates the largest number of previously published datasets, ensuring comprehensive diversity in terms of race, gender, and image resolution.
- We propose a new architecture that combines Inception v4, an Attention mechanism, and an MLP Mixer, effectively mitigating the limitations of each individual component. Our model achieves state-of-the-art performance across multiple datasets while maintaining a significantly lower computational footprint compared to other leading models. This advancement fosters more efficient and equitable kinship verification systems.

This paper is organized as follows: Section 2 reviews related research, Section 3 introduces the EarKinshipVN dataset, Section 4 details the proposed model, Section 5 covers the experiments, and Section 7 presents the conclusions.

2 Related Works

Kinship Verification

Most kinship verification research has focused on facial imagery due to dataset limitations. The Siamese Neural Network (SNN) is commonly used in this domain, with architectures based on VGGFace and ResNet-101 incorporating attention mechanisms and fairness improvements [27]. Other approaches, such as SNNs with lightweight backbones like SqueezeNet, have also shown effectiveness in kinship verification tasks [28].

The primary objective of this study is to determine kinship based on pairs of input photos using advanced techniques by exploring the configurations of Siamese models.

Ear Recognition

Ear recognition systems have evolved from early geometric and structural methods [29] to modern deep-learning approaches [30], significantly improving performance in uncontrolled environments. Shailaja et al. [31] introduced a simple, rotation-invariant geometric approach, while Sinha et al. [32] combined SVMs, HOG, and CNNs for ear localization and recognition. Alshazly et al. [33] compared CNN architectures, with ResNeXt101 achieving state-of-the-art performance, while Xu et al. proposed an efficient lightweight model based on MobileNetV2.

Kinship Recognition Using Ear Images

Research on kinship recognition from ear images remains limited [34]. Meng et al. proposed a model-based approach using handcrafted features, incorporating geometric features and HOG with distance metrics for verification. Their model achieved 95.6% accuracy on the USTB dataset using the Manhattan distance metric, with preprocessing techniques such as the Hough transform and affine transformation improving performance.

More recently, Grega Dvoršak et al. introduced deep learning models for kinship verification using VGG16, ResNet152, and other architectures. Their study, based on 37,282 image pairs, demonstrated that ear images can serve as a reliable biometric for kinship recognition, with models achieving over 60% in ROC-AUC. Notably, VGG16 performed best with 64.01% accuracy and 69.2% ROC-AUC, suggesting that networks with fewer layers provide more relevant features for this task.

Inspired by previous research, our article proposes a novel dataset named KinEarVN and experiments with new robust architectures that are improvements upon Inceptionv4 [20].

3 Proposed Method

The dataset contains images of varying sizes, with smaller images lacking sufficient detail for effective training. Drawing inspiration from previous research, we apply super-resolution and image restoration techniques to enhance the quality of these images. Studies by Umirzakova et al. [35] and Gunturk et al. [36] have also demonstrated that applying these image preprocessing techniques can improve the accuracy of results.

3.1 Data Preprocessing

To prepare the data for modeling, we employ three methods: Super Resolution, Image Restoration, and Image Normalization. Images are preprocessed using the Hybrid Attention Transformer (HAT) model for super-resolution and restoration, with initial loading handled by the Pillow library. The parameters are kept at their default values as specified in the studies by Chen et al. [37,38]. Afterward, we resize all images to 320×320 before applying normalization. We selected this size as it strikes a balance between being small enough to avoid loss of detail, which could lead to errors, and large enough to prevent overfitting. Once resized, the images are normalized so that each pixel value for each color channel ranges from 0 to 1. The process is illustrated in detail in Fig. 1.

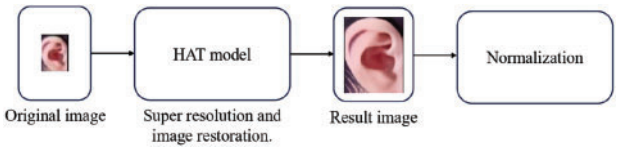

Figure 1: Data preprocessing process

3.2 Attention-Based Inception Block

Inceptionv4 is an improved version of the earlier Inception models, known for its exceptional performance in image classification tasks. It achieved a top-5 error rate of just 3.08. However, in deep neural networks, the absence of residual connections can lead to a decline in performance, particularly as the network depth increases.

Gradient vanishing According to the research by Hochreiter [39], the vanishing problem is a serious issue in the machine learning world. It has significant effects on deep networks, causing the gradients used to update the network to become extremely small or "disappear" as they are back-propagated from the output layer to previous layers. This phenomenon is known as vanishing gradient, which leads to slow convergence, the network getting stuck at low minima, and degraded learning ability.

Deviation in function class The advent of deep neural networks has fostered a common belief that adding more layers improves data extraction. However, this also leads to a different set of functions. As the network deepens, its layers become more powerful and complex, but they also become increasingly distinct from one another. Fig. 2 illustrates how different classes of functions can diverge, with the letter f representing the function class.

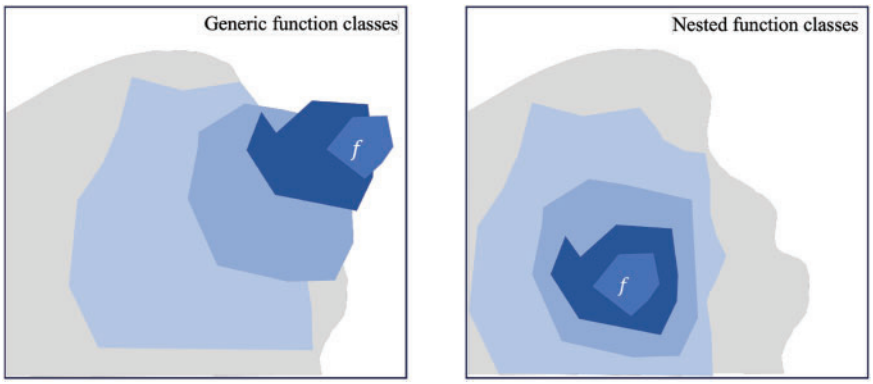

Figure 2: For non-nested function classes, a larger class does not always bring us closer to the "truth" function. In contrast, nested classes consistently refine previous layers, keeping the model within more accurate function classes

It is not always guaranteed that adding new layers to neural networks will increase their expressive capabilities unless the new function classes are already contained within the larger ones. In deep neural networks, if we can train the newly added layer to perform an identity function $f(\mathbf{x}) = \mathbf{x}$, the new model will operate as effectively as the original. Additionally, the new layer may help reduce training errors by finding better ways to fit the training dataset.

To solve the δ_j problems, we propose to use residual mapping published by He et al. [40]. The identity function, the simplest function, is obtained when the input is directly added to the output in a vanishing gradient situation, where all parameters are zero or negligible. Regarding deviation in function classes, increasing the number of parameters can cause the model to drift away from the initially identified function. When an additional layer is added to the model, it introduces an inductive bias but still performs the recognition function while preserving the output of the previous layers. The key difference is that the output can now be further processed for the next input, allowing the model to refine its predictions.

We apply Residual Learning to selected stacked layers. A building block of this approach is illustrated in the figure below. Where $x, \mathcal{F}(x,\ W_i)$ represent the input and output of the matrix under consideration, respectively. Specifically in this article, the function $\mathcal{F}(x,\ W_i)$ is calculated using the formula:

$$\mathcal{F}(x, W_i) = \text{AttentionBlock}(\text{InceptionBlock}(x)) + x \tag{1}$$

The Formula (1) uses an Inception block, representing both Inception A and Inception B. The effectiveness of this architecture is demonstrated in Section 5. Additionally, we use the Reparameterization technique from Ding et al. [41] to enhance computational efficiency during the inference stage. We experiment with four different types of attention mechanisms to focus on important parts of the image while eliminating less relevant areas. Detailed information about the backbones is provided below.

- **Squeeze-and-Excitation Attention** Proposed by Hu et al. [42], this mechanism enhances CNNs by modeling channel interdependencies. It uses global average pooling to capture channel statistics, applies fully connected layers to learn adaptive weights, and scales the original features, improving network performance.
- **Large Kernel Attention (LKA)** Introduced by Guo et al. [43], LKA improves CNNs by incorporating large convolutional kernels into the attention mechanism. This approach captures long-range dependencies and enhances feature focus, leading to better accuracy and robustness in image classification and segmentation.

- **Shuffle Attention (SA)** Zhang et al. [44] proposed SA to improve attention mechanisms by dividing channel dimensions into sub-features and processing them in parallel. A shuffling operation facilitates information exchange, effectively capturing spatial and channel dependencies.
- **Triplet Attention** Misra et al. [45] introduced this lightweight mechanism, which captures multi-dimensional interactions using a three-branch structure. By leveraging rotation and residual transformations, Triplet Attention encodes both spatial and inter-channel information with minimal computational cost.

3.3 The Backbone Model

The main objective of the proposed method is to develop a lightweight and high-performance model for kinship recognition. In this section, we first provide a detailed explanation of the model and how it enhances accuracy. Secondly, we present an in-depth overview of the commonly used backbones that we experimented with.

Mixer Attention Inception In our architecture, we aim to combine attention layers, CNNs, and multi-layer perceptrons to fully leverage the model's capabilities. We selected the Inception architecture for its diverse combination of convolutional filters, which allows for feature extraction at different scales. As shown in the Fig. 3a, we use the "spatial multiplier" in Inception-v4 to reduce the computational cost of convolution operations and employ a "stem module" to enhance information flow through the network. Additionally, the grid size reduction modules help decrease the spatial size of feature maps while maintaining their depth, thereby improving network efficiency. The reduction block architecture (Figures remains identical to that of Inceptionv4). Additionally, the incorporation of residual connections and attention mechanisms, as illustrated in Figs. 3 and 4, respectively, enhances the model's representational capacity and efficiency. The reduction block architecture (Fig. 4b) was applied to reduce spatial dimensions and aggregate Information. Additionally, the incorporation of residual connections and attention mechanisms, as illustrated in Fig. 4c and d, respectively, enhances the model's representational capacity and efficiency. Subsequently, the output feature map of size $H \times W \times C$ is divided into patches of size $P \times P$, resulting in dimensions $H' \times W'$, where H' and W' denote the spatial dimensions after reduction. The reshaped output has a form of $M \times D$, where $M = \frac{H'W'}{P^2}$ represents the number of patches, and $D = P^2 \cdot C$ is the dimensionality of each patch. This representation is then fed into the MLP Mixer block (Fig. 4e). The application of the MLP Mixer enables the model to capture global context, addressing the limitations of the local receptive field of CNN.

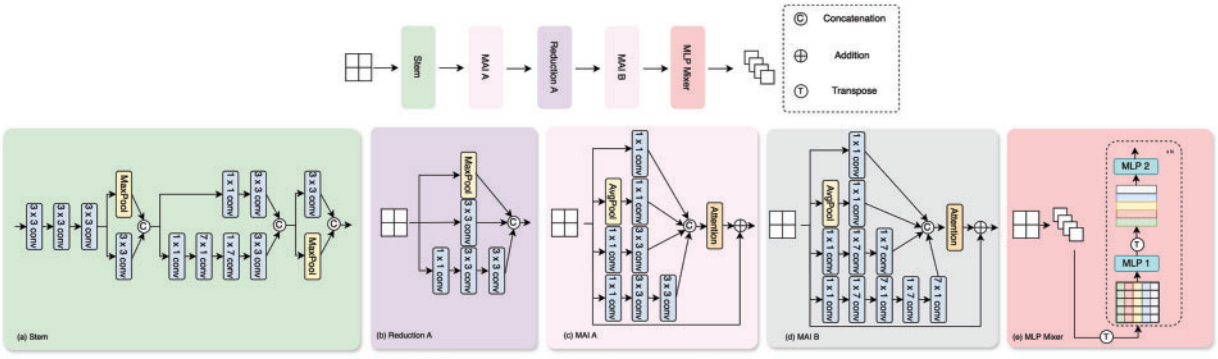

Figure 3: Mixer Attention Inception architecture, as well as the (a) Stem block, (b) Reduction A, (c) Mixer Attention Inception block A (MAI A), (d) Mixer Attention Inception block B (MAI B), (e) MLP Mixer block

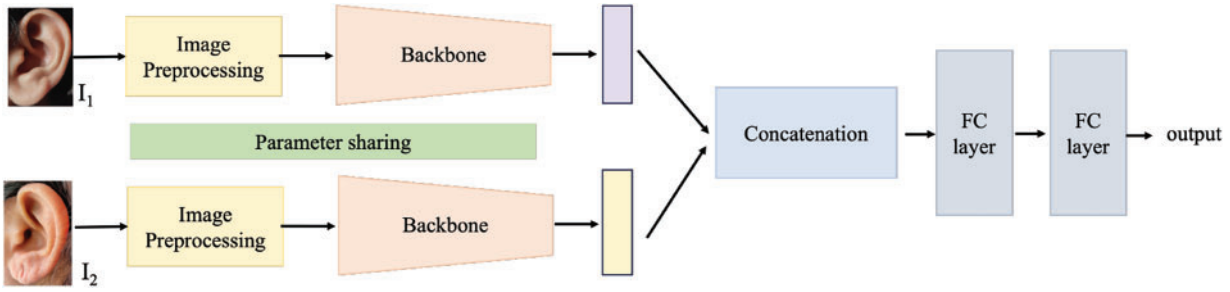

Figure 4: Synopsis of the used framework. We utilize a Siamese model configuration, which can be easily built utilizing different backbones, as the basis for the study, drawing on current literature in relevant problems

Next, we will explain why combining CNNs with the MLP Mixer model leads to significant efficiency gains. Modern deep vision architectures consist of layers that combine features at a given spatial location, between different spatial locations, or both simultaneously. In CNNs, 1×1 convolutions operate at a given spatial location, while larger kernels handle interactions both at the same location and between different spatial locations. In the MLP Mixer model, interactions at the same spatial location are managed through channel-mixing, while interactions across different spatial locations are handled through token-mixing. Specifically, convolution uses a fixed kernel to gather data from the nearby receptive field. In more detail, the convolution operation involves applying a filter (kernel) to an input image, resulting in the generation of a feature map. Mathematically, the convolution of an input X with a filter F at a position (i, j) is given by:

$$(X * F)(i, j) = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} X(i+m, j+n) \cdot F(m, n) \tag{2}$$

where $X(i+m, j+n)$ represents the value of the input image at position $(i+m, j+n)$, $F(m, n)$ is the value of the filter at position (m, n), and h_w and k_w are the height and width of the filter, respectively. The property of Translation Equivariance means that when the input X is shifted, the resulting feature map Y produced by the convolution will also be shifted by the same amount.

Let X be the original input and $X_{shifted}$ be the input shifted by (δ_i, δ_j):

$$X_{\text{shifted}}(i, j) = X(i - \delta_i, j - \delta_j) \tag{3}$$

The output feature map for the shifted input is:

$$Y_{\text{shifted}}(i, j) = (X_{\text{shifted}} * F)(i, j) \tag{4}$$

Substituting $X_{shifted}$:

$$\begin{aligned} Y_{\text{shifted}}(i, j) &= \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} X_{\text{shifted}}(i+m, j+n) \cdot F(m, n) \\ Y_{\text{shifted}}(i, j) &= \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} X(i+m-\delta_i, j+n-\delta_j) \cdot F(m, n) \end{aligned} \tag{5}$$

Notice that this is equivalent to the output feature map of the original (δ_i, δ_j):

$$Y_{\text{shifted}}(i, j) = Y(i - \delta_i, j - \delta_j) \tag{6}$$

Therefore, the output feature map for the shifted input is simply the shifted version of the original output feature map, demonstrating translation equivariance. In contrast, MLP Mixer allows the receptive field to be the entire spatial location and the interaction between the same channel and channels is followed by the formula:

$$\begin{aligned} \mathbf{X}^2_{*,i} &= \mathbf{X}^1_{*,i} + \mathbf{W}_2\boldsymbol{\sigma}\left(\mathbf{W}_1\mathrm{LN}(\mathbf{X}^1)_{*,i}\right), \quad \text{for } i = 1\ldots C \\ \mathbf{Y}_{j,*} &= \mathbf{X}^2_{j,*} + \mathbf{W}_4\boldsymbol{\sigma}\left(\mathbf{W}_3\mathrm{LN}(\mathbf{X}^2)_{j,*}\right), \quad \text{for } j = 1\ldots S \end{aligned} \tag{7}$$

A series of S non-overlapping picture patches, each projected to a chosen hidden dimension C, are input into the mixer. LN represents Layer Normalization and σ is an element-wise nonlinearity (GELU [46]). Through the analysis δ_j, we identify key characteristics of each model. Translation equivariance enhances generalization on small datasets, but MLP-Mixers lack this property due to their use of absolute positional embeddings, explaining why CNNs often outperform them in such cases. While MLP-Mixers benefit from a global receptive field that improves contextual learning, this comes at the cost of increased computational complexity, requiring trade-offs in efficiency. To enhance feature extraction, we integrate Attention mechanisms with CNNs; however, to avoid missing less prominent details, we incorporate MLP-Mixers for a more comprehensive evaluation, ultimately improving classification accuracy.

Considering the comparison provided in Table 1, the ideal model should be able to combine the three desirable attributes.

Table 1: Desirable properties found in attention-base convolution or MLP Mixer

Properties	Attention-based convolution	MLP Mixer
Translation equivariance	✓	
Highlight key features	✓	
Global receptive field		✓

In a study by Dvoršak et al. [25], it was demonstrated that the design of neural network backbones with varying depths can significantly impact the output results. Through experiments involving five different backbones, the researchers concluded that a moderately shallow backbone is the most suitable choice for this specific problem. Based on these findings, we developed two types of backbones that are not excessively deep, yet are designed to effectively address the optimization challenges inherent in this task.

The details of the layers are provided in Table 2. The stem block with Reduction remains consistent with Inceptionv4. Block A and Block B are based on the Inception A and Inception B designs, respectively. The MAI architecture prioritizes compactness, while MAIm focuses on enhancing accuracy. However, both architectures are intentionally kept relatively shallow to align with our initial design goals.

Table 2: The table details the mixer attention inception architecture including MAIs (size s) and MAIm (size m)

Layer	MAIs	MAIm
Stem	Stem × 1	
Attention-based Inception Block (1)	Block-A × 1	Block-A × 2
Reduction	Reduction-A × 1	

(Continued)

Table 2 (continued)

Layer	MAIs	MAIm
Attention-based Inception Block (2)	Block-B × 2	Block-B × 2
MLP-Mixer	Mixer-Block × 2	Mixer-Block × 3

We consider the five different backbone models, denoted as δ_j, within the framework to extract image representations for kinship verification based on ear images. These backbones (**Vgg** [47], **Inception-v4** [20], **Resnet** [40], **DenseNet** [48], **EfficientNet** [49], **Efficientnetv2** [50], **ViT** [51]) are all publicly available to ensure reproducibility and have been selected for their state-of-the-art performance in various vision tasks.

3.4 Network Architectures

The proposed framework is illustrated in Fig. 4. For input image with $I_1 \in \mathbb{R}^{H\times W\times 3}$ and $I_2 \in \mathbb{R}^{H\times W\times 3}$ representing two RGB (Red, Green, Blue) ear images, we have 2 images $I_1' \in \mathbb{R}^{320\times 320\times 3}$, $I_2' \in \mathbb{R}^{320\times 320\times 3}$ corresponding to I_1, I_2 after going through the preprocessing process. We have a model named θ which has been trained to generate a kinship score based on the inputs II_1' and I_2'. Specifically, after passing through the backbone to extract features, the image pair I_1 and I_2 will produce two feature vectors. These two vectors will then be combined before going through two fully connected layers. The purpose of having two fully connected layers is to maximize the use of information extracted from the backbone. The kinship verification task assigns the input pair (I_1', I_2') to either the class of images with kin relationships (1) or the class of images without kin relationships (0), or formally:

$$(I_1', I_2') = \begin{cases} 1, & if \quad \theta(I_1', I_2') > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Inspired by face-related kin recognition models, we experiment on Siamese network architecture that takes two ear images, I_2' and I_2', of different people as input. The synthesis from two input photos is extracted by Backbone to create a feature map, which is then flattened into a vector and concatenated. Following concatenation, the vector be sent to the completely connected layer, where it be used to compute the kinship score, which establishes the relationship between the subjects in the input photos.

Architecture. The framework uses a Siamese model architecture and takes two separate ear pictures of different people, x_1 and x_2, as input. The Siamese architecture consists of two branches implemented using a chosen backbone model with shared common parameters. These backbones generate two image representations, or image embeddings, referred to as y_1 and y_2. These embeddings are then combined and passed through several fully connected layers that analyze the relationships between the two embeddings. Ultimately, this process produces the kinship score $\theta(x_1, x_2)$, which determines the relationship between the individuals in the input images.

Training. We use binary cross-entropy as the learning target and apply binary supervision for picture pairings that include or exclude kin relations when training the Siamese model. Even though the EarKinshipVN dataset, which was added later, provides more data than earlier datasets in this field, data augmentation is still necessary to prevent overfitting in the models. Therefore, we take advantage of the capabilities found in the Torchvision libraries throughout the training process including: normalization and resize image. With KinEar dataset, during the training procedure, we applied: color jitter with brightness and

hue set to 0.2, 0.5, respectively, Gaussian blur with kernel size 5 × 9, random sharpness adjustment with the sharpness factor set to 2 and horizonal flip. All augmentations for KinEar dataset are given probability 0.5.

4 Dataset Preparation

The KinEar dataset, developed by the University of Ljubljana [25], comprises 1477 photos from 19 families but is limited in size and racial diversity. To address these shortcomings, we created the EarKinshipVN database, which contains four times more images and over twice the number of relationship pairs, making it suitable for both kinship and ear recognition research. Table 3 provides a comparative analysis of related datasets.

Table 3: Detailed comparison table of datasets for kinship identification

Characteristic	Value	
	KinEar (Grega Dvorsak et al.)	EarKinshipVN (Our)
Family	19	157
Members/Subjects	76	498
Subject-2-Subject Kin Relations	96	508
Images	1477	4876
Kin image pairs	37,282	73,220

Our dataset includes images from a diverse range of racial backgrounds and ages (12 to 75 years) to minimize bias. After collecting photos, we crop ear regions, apply filtering techniques to remove occluded or low-resolution images, and ensure that families with multiple relationships are represented.

The gender distribution details are depicted in Fig. 5a. In this visualization, individuals categorized as mother and daughter are considered female, while those categorized as father and son are considered male. According to the accompanying data table, 3378 individuals are identified as male, and 1498 are identified as female. The data clearly shows that the number of males is significantly higher than the number of females. This imbalance presents a challenge that we aim to address in the field of gender recognition by ear.

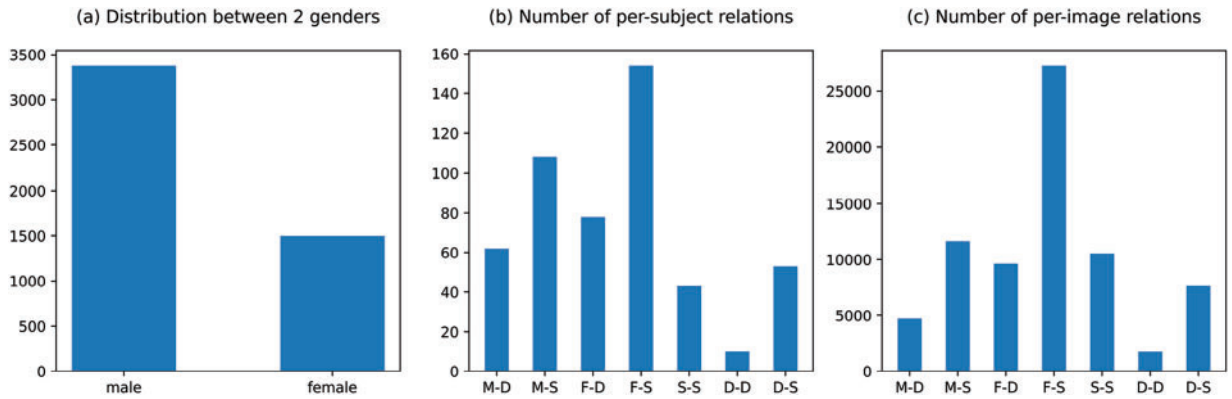

Figure 5: Visualization of distribution of relationships in EarKinshipVN

Fig. 5b and c shows the distribution of 7 pairs and Table 4 provides detailed numbers for each pair, including mother-daughter (M-D), mother-son (M-S), father-daughter (F-D), father-son (F-S), son-son (S-S), daughter-daughter (D-D), and daughter-son (D-S). This classification expands the kinship identification

problem by specifically identifying these seven kinship relations. However, it is evident that the father-son relationship has a significantly higher number of pairs compared to others, while the daughter-daughter pair has the fewest. This data imbalance introduces a new challenge in handling skewed datasets. The left panel shows the distribution rate of each pair, while the right panel displays the distribution rate of images within each pair.

Table 4: Summary table of the number of relationships by kinship

Relation	Number of subject relation	Number of pair-image relation
Mother-daughter	62	4705
Mother-son	108	11,673
Father-daughter	78	9628
Father-son	154	27,297
Son-son	43	10,555
Daughter-daughter	10	1756
Daughter-son	53	7636

The EarKinshipVN dataset includes images of six pairs, as shown in Fig. 6. Each image represents a family relationship described in the caption. The selected photos highlight the ear accessories in each image, which are notable features. Some images in the dataset, like pairs 6a, 6b, 6c, and 6d, include earrings. Additionally, the dataset includes attributes such as skin tone, hair texture, and style, offering a comprehensive representation of diverse visual features. These attributes enhance the dataset's utility for studying the verification and recognition of kinship relations based on ear-related features and overall appearance.

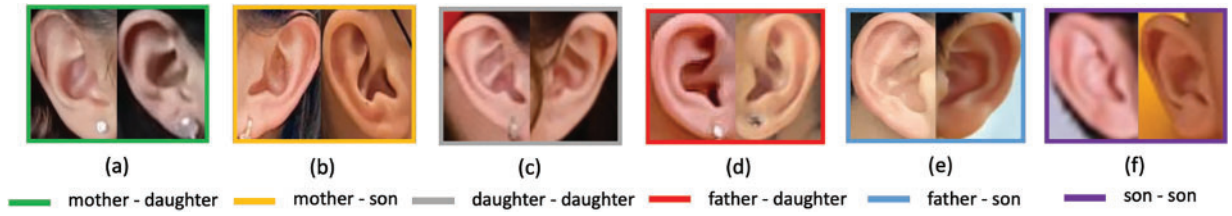

Figure 6: Example of pair images from six different relation

5 Experimental Result

Dataset In this study, we used the EarKinshipVN dataset and the KinEar dataset to evaluate our proposed methodology. We divided the datasets into sets based on the same relationship categories and then generated corresponding unrelated pairs. The kinship relations we focused on include Father-Son (F-S), Father-Daughter (F-D), Mother-Son (M-S), Mother-Daughter (M-D), Daughter-Daughter (D-D), Son-Son (S-S), and Daughter-Son (D-S). For the EarKinshipVN dataset, we shuffled all pairs before dividing them into three sets, with the training, validation, and testing sets distributed at 50%, 15%, and 35%, respectively. For the KinEar dataset, we structured it similarly to previous experiments, with 14 families in the training set, 2 in the validation set, and 3 in the testing set. The KinEar testing set contains a total of 12,960 possible image pairs, with 9692 negative pairs and 3268 positive pairs.

Implementation Details. All of our models are developed and trained using the PyTorch framework, utilizing hardware that includes two 2080Ti GPUs, 256 GB of RAM, and an Intel Xeon E5 CPU. The models are trained using the Adam optimizer with a weight decay of 5e-5, $\beta_1 = 0.9$, $\beta_2 = 0.99$. We train the networks

for 10 epochs with a learning rate of 1e-5 for the EarkinshipVN dataset and for 15 epochs with a learning rate of 1.5e-5 for the KinEar dataset to ensure effective convergence. To evaluate the backbones within our framework, we use various performance indicators, including classification accuracy, precision, recall, and F1-score.

We compare our backbones with other well-performing backbones on the same task across two datasets: EarKinshipVN and KinEar.

EarKinshipVn The backbones we selected for experimentation and comparison with our model include VGG, ResNet, DenseNet, EfficientNet, EfficientNetv2, and Inception-v4. The comparison details are provided in Table 5.

Table 5: Benchmark on EarKinshipVn Dataset. The Mixer Attention Block (MAIs) model is combined with attention including Squeeze-and-Excitation Attention (SE), Large Kernel Attention (LKA), Shuffle Attention (SA), Triplet Attention (TA). Top 2 results are markedin bold and red

Backbone		#Params	Accuracy	F1-score	Precision	Recall
VGG [47]	vgg16	17.9M	92.89	92.92	92.69	93.15
	vgg19	23.2M	95.54	95.64	93.63	97.74
Resnet [40]	resnet101	59.3M	74.78	77.36	70.20	86.14
	resnet152	74.9M	70.16	74.42	65.13	86.81
Densenet [48]	densenet121	20.1M	98.59	98.60	98.13	99.08
	densenet169	33.9M	98.44	98.44	**98.40**	98.49
Inception [20]	inceptionv4	41.3M	89.46	90.10	84.97	95.88
Efficientnet [49]	efficientb4	17.8M	82.82	85.00	75.45	97.31
	efficientb5	28.6M	79.76	82.52	72.62	95.54
	efficientb6	41.0M	84.14	85.68	78.09	94.90
	efficientb7	64.1M	80.44	81.86	76.32	88.26
Efficientnet_v2 [50]	efficient_v2s	20.3M	79.94	83.07	71.87	98.40
	efficient_v2m	53.0M	85.10	86.75	78.13	97.51
	efficient_v2l	117.3M	83.30	85.01	77.12	94.70
ViT [51]	vit_b16	86.8M	97.51	97.52	97.34	97.69
	vit_b32	88.3M	97.99	98.00	97.49	98.52
	vit_l32	306.7M	96.75	96.79	95.60	98.01
MAIs (Our)	MAIs + TA	**15.5M**	96.33	96.42	94.36	98.56
	MAIs + SE	15.8M	98.29	98.31	97.48	99.14
	MAIs + SA	15.6M	**98.71**	**98.72**	98.16	99.28
	MAIs + LKA	17.9M	98.44	98.46	97.63	**99.31**

This table presents different backbone architectures with distinct features and performance metrics, highlighting the strengths and weaknesses of each model. VGG models showed quite impressive performance in a previous paper by Dvoršak et al. This model, especially VGG16 and VGG19, shows clear improvements with increased parameters, with VGG19 achieving an accuracy of 95.54% and an F1 score of 95.64%. DenseNet models, especially DenseNet169, stand out for their outstanding performance, having an impressive accuracy of 98.44% and a leading accuracy of 98.40%, illustrating their effectiveness with

little information. more number. The EfficiencyNet models, along with the v2 variants, show moderate performance, where EfficiencyNet-v2m achieves 85.10% accuracy, suggesting room for improvement. Vision Transformer (ViT) models, such as vit-b32, give remarkable results with 97.99% accuracy, highlighting the potential of transformer-based approaches in achieving High precision and balanced data. The proposed our MAI methods significantly outperform other models, with MAIs + SA achieving an outstanding accuracy of 98.71% and the highest F1 score of 98.72%.

The Fig. 7 presents a comprehensive comparison of four models-MAI + SA, DenseNet121, VGG16, and ViT_b32—evaluated across three key metrics: accuracy, parameter count, and computational complexity (FLOPs: Floating Point Operations per Second). MAI_SA demonstrates outstanding performance, achieving the highest accuracy of 98.71%, slightly surpassing DenseNet121 at 98.59%, while VGG16 falls behind at 95.54%. Notably, MAI + SA achieves this superior accuracy with parameter efficiency, utilizing only 15.6M parameters compared to DenseNet121's 20.1M, VGG16's 17.9M, and the significantly heavier ViT_b32, which requires 88.3M parameters. Furthermore, in terms of computational efficiency, MAI + SA records the lowest FLOPs at 13.52G, followed by DenseNet121 at 11.83G, whereas VGG16 and ViT_b32 demand substantially more computation, with 62.64G and 107.32G FLOPs, respectively. These results underscore the effectiveness of MAI + SA as a model that not only delivers state-of-the-art accuracy but also excels in minimizing both parameter and computational overhead, making it an ideal choice for resource-constrained environments. This balance of efficiency and performance positions MAI + SA as a significant advancement in model design for vision tasks.

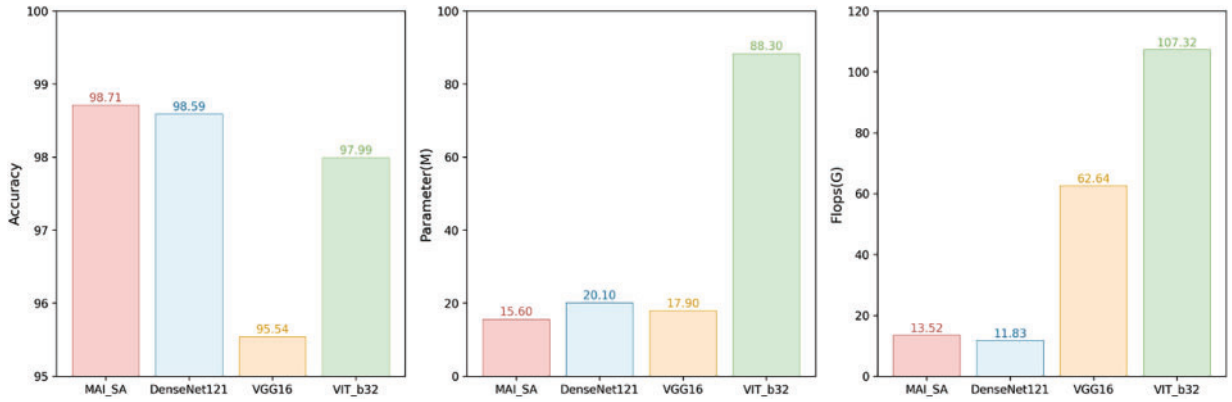

Figure 7: Comparisons between efficient models including MAI-SA (Our), DenseNet [48], VGG [47], VIT [51] in terms of accuracy, parameter, and flops. The input size of images is set to 320 × 320

The benchmark graph features ROC (Receiver Operating Characteristic) curves comparing the performance of different models in terms of their true positive rate (TPR) vs. false positive rate (FPR). This highlights the area under the curve (AUC) as an overall measure of model effectiveness. Our MAIs_SA model is particularly noteworthy with an AUC of 0.99, demonstrating near-perfect classification capabilities. Similarly, DenseNet169 and VGG19 also achieve an AUC of 0.99, indicating high reliability and precision. The ViT_b16 model follows with a strong AUC of 0.96, showcasing its robust performance. EfficientNet_v2m and Inceptionv4 exhibit moderate AUC values of 0.89 and 0.93, respectively, while EfficientNetb7 and Resnet101 show lower AUCs of 0.82 and 0.81, indicating a higher rate of false positives compared to the top performers. This comparison underscores the exceptional performance of the MAIs_SA model, highlighting its ability to maintain a high true positive rate with minimal false positives, making it a highly effective solution for classification tasks.

In the Table 6, among the models based on the Grega Dvoršak method, VGG16 achieves the highest accuracy of 64.01%, while the USTC-NELSIP model shows the lowest performance with 55.12% accuracy. ResNet152 and AFF models also show moderate accuracy levels at 57.50% and 60.00%, respectively. In contrast, our proposed MAIm models demonstrate superior performance, with MAIm + SA leading the group with an impressive accuracy of 71.10% and a notable ROC-AUC of 75%. MAIm + SE follows closely with 70.39% accuracy and the highest ROC-AUC of 78%, indicating its robust classification capabilities. MAIm + LKA and MAIm + TA also outperform traditional methods, achieving accuracies of 66.71% and 67.32%, respectively, and showing strong sensitivity and specificity metrics. These results highlight the effectiveness of the MAIm models, particularly MAIm + SA and MAIm + SE, in delivering higher accuracy and better overall performance compared to existing approaches on the KinEar dataset.

KinEar The Table 7 provides a comprehensive comparison of various backbone models used in different methods, including those proposed by Grega Dvoršak and a new approach referred to as MAIm (Our). The comparison is based on several performance metrics: the number of parameters (Params), Accuracy, F1-Score, Precision, Recall, and ROC-AUC.

Table 6: Benchmark on KinEar dataset. The Mixer Attention Block (MAIm) model is combined with attention including Squeeze-and-Excitation Attention (SE), Large Kernel Attention (LKA), Shuffle Attention (SA), Triplet Attention (TA)

Backbone		#Params	Accuracy	Sensitivity	Specificity	ROC-AUC
Grega Dvoršak method [25]	VGG16 [47]		64.01	64.01	64.01	69.22
	ResNet152 [40]		57.50	57.50	57.51	63.14
	USTC-NELSLIP [52]		55.12	55.14	55.10	57.29
	AFF [53]		60.00	60.00	60.00	64.01
	CoTNet [54]		61.85	61.84	61.86	65.88
MAIm (Our)	MAIm + LKA	20.62M	66.71	70.74	64.52	71.44
	MAIm + SA	18.02M	71.10	64.97	73.87	74.64
	MAIm + SE	18.31M	70.39	70.48	70.41	78.02
	MAIm + TA	18.02M	67.32	64.62	75.39	74.31

Table 7: Ablation study. We train all cases on EarkinshipVN for 10 epochs. Best result are bold

Architecture	Accuracy	F1-Score	Precession	Recall
Baseline	95.84	96.00	92.62	99.63
+ RAI Block only	94.55	94.81	90.45	99.63
+ Mixer only	97.23	97.29	95.64	98.90
+ RAI Block and Mixer	**98.71**	**98.72**	**98.16**	**99.31**

The Figs. 8 and 9 illustrate the trade-off between the true positive rate and the false positive rate for each model. The MAIm + SE model achieves the highest AUC at 78%, signifying superior performance compared to the others. Following closely are the MAIm + SA and MAIm + TA models with AUC values of 75% and 74%, respectively. The MAIm + KLA model has the lowest AUC at 71%. Overall, the ROC curves and AUC values confirm the findings from the table, reinforcing that MAIm + SE is the most effective model, with MAIm + SA and MAIm + TA also showing robust performance.

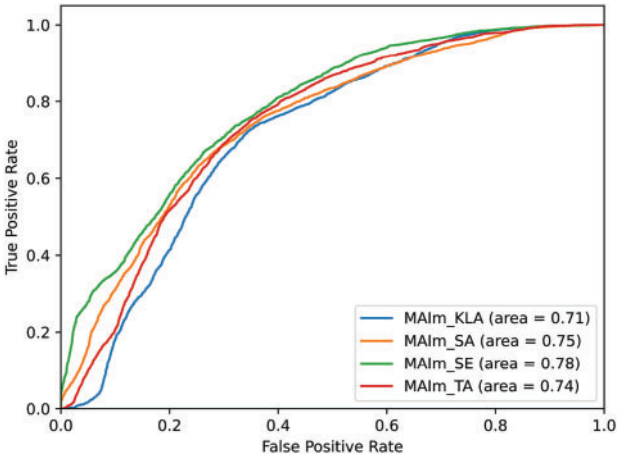

Figure 8: ROC_AUC index of our backbones (KinEar)

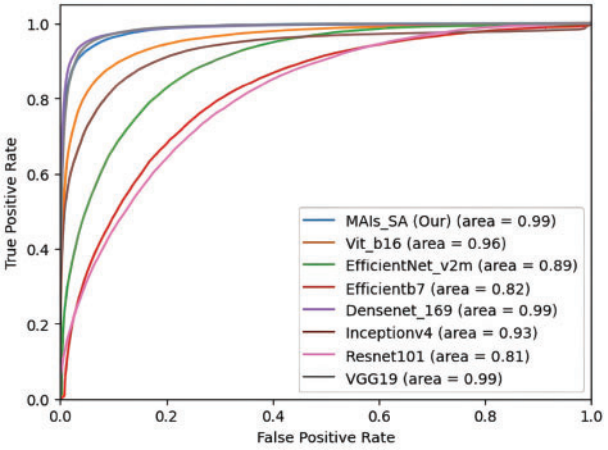

Figure 9: ROC_AUC index of some prominent backbones (EarKinshipVN)

Ablation study In Table 7, we experiment to figure out the effectiveness of each components in MAI architecture through EarKinshipVN dataset. We experimented based on 4 cases including baseline case, RAI block only, Mixer only replacement and MAI block. The results are presented according to the best results of the experiments.

Although the results slightly degrade when adding the RAI block, this happens because the convergence is slower. We believe that the convergence may be slower but the amount of information transmitted is larger. Our MAI architecture shows the efficiency of combining the Mixer and the RAI block.

6 Visualization and Analysis on Ear Kinship Relations

Most previous deep learning models that process human ear images use the entire image for learning. However, experts in human ear biometrics suggest that the helix contains the most crucial information for discrimination. In this section, we will analyze the key regions emphasized by the deep learning model using heat maps.

Ear structure The human ear begins to form very early during pregnancy and is fully developed by the time a child is born. The ear has a unique anatomical structure common to all humans, as it functions as the organ of hearing. As shown in Fig. 10a, the shapes of the tragus, antihelix, helix, lobe, and other key structural components define the appearance of the external ear. These anatomical cartilage structures vary in shape, appearance, and relative position from person to person and are often hereditary among individuals sharing the same bloodline.Some previous studies suggest that the helix plays a crucial role in identifying and performing ear-related tasks. To explore this further, the next section will present a heatmap analysis to highlight the important areas that the model focuses on.

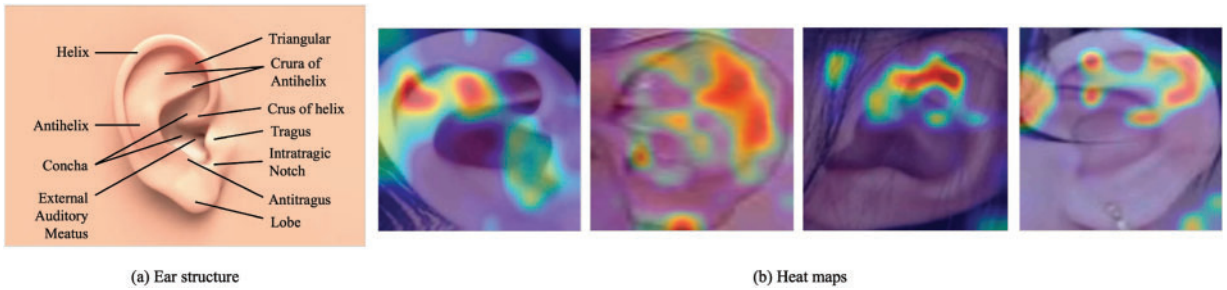

Figure 10: Ear structure

Heat map To understand the important regions that a deep learning architecture focuses on, we generated heat maps using the GradCAM algorithm [26]. These heat maps highlight varying levels of attention, with warmer colors indicating higher levels of focus across the five cognitive models. The ear images used in this analysis were randomly selected from the EarKinshipVN dataset.

The heat map indicates that the helix and the upper half of the ear are the most attention-grabbing areas (Fig. 10b). In particular, the helix, antihelix, triangular fossa, and crura of the antihelix are the primary areas of focus when evaluating kinship relations. Our research group is at the forefront of this field, aiming to identify key features that will enhance focus and reduce noise, especially for methods like SIFT or HOG.

7 Conclusions

In this paper, we presented a competitive model for kinship recognition through ear images, leveraging the combined strengths of CNNs, Attention mechanisms, and MLP Mixer models. Our approach demonstrated significant improvements in performance by effectively capturing rich semantic information and focusing on the most relevant features. The integration of these advanced techniques allowed our model to outperform other state-of-the-art models on both the EarKinshipVN and KinEar datasets, showcasing its robustness and adaptability across different data sources. Additionally, the release of the EarKinshipVN dataset contributes to the field by providing a larger and more diverse resource for future research, addressing the limitations of smaller, less generalizable datasets like KinEar.

This work opens new directions in biometric recognition. Expanding the dataset to include diverse populations and conditions is crucial for improving model generalizability and robustness. Furthermore, integrating multimodal approaches by combining ear recognition with complementary biometric modalities, such as facial recognition, gait analysis, or voice recognition, offers a compelling avenue for enhancing accuracy and reliability. Optimizing the model for real-time deployment, with a focus on computational efficiency, will be essential for practical applications in security, surveillance, and forensic investigations. Additionally, improving the model's explainability and interpretability will bolster user trust and adoption in sensitive contexts. To address current limitations, we plan to further enrich the dataset to mitigate existing imbalances and ensure a more equitable representation. These efforts aim to advance the state of the art in biometrics and foster broader adoption in real-world scenarios.

Ethical Consideration

Privacy and Consent: Biometric data, including ear images, are highly sensitive and unique to individuals. Collecting and using such data require explicit informed consent from participants, ensuring they are fully aware of how their data will be stored, processed, and used. In our study, all participants provided written consent, and data collection adhered to ethical research guidelines.

Data Security and Anonymization: To prevent unauthorized access and potential misuse, biometric datasets must be securely stored using encryption and anonymization techniques. In our research, we employ stringent data security measures, including access restrictions and pseudonymization, to protect participant information.

Fairness and Bias: Machine learning models may exhibit biases due to imbalanced datasets, leading to potential disparities in recognition accuracy across different demographic groups. To mitigate this, we ensured racial and gender diversity in our EarKinshipVN dataset and performed fairness evaluations to assess model performance across various subgroups.

Acknowledgement: This work is supported by Ho Chi Minh City Open University.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design, collection, analysis and interpretation of results, draft manuscript preparation: Thien-Tan Cao, Huu-Thanh Duong, Viet-Tuan Le. Review paper: Vinh Truong Hoang, Hau Nguyen Trung, Kiet Tran-Trung. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data support the findings of this study are available from the corresponding author, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Yan H, Song C. Multi-scale deep relational reasoning for facial kinship verification. Pattern Recognit. 2021;110:107541.
2. Fang R, Tang KD, Snavely N, Chen T. Towards computational models of kinship verification. In: 2010 IEEE International Conference on Image Processing; 2010; Hong Kong, China. p. 1577–80.
3. Lowe G. Sift-the scale invariant feature transform. Int J. 2004;2(91–110):2.
4. Pietikäinen M. Local binary patterns. Scholarpedia. 2010;5(3):9775.
5. Liu CL, Koga M, Fujisawa H. Gabor feature extraction for character recognition: comparison with gradient feature. In: 8th International Conference on Document Analysis and Recognition; 2005; Republic of Korea. p. 121–5.
6. Wu H, Chen J, Liu X, Hu J. Component-based metric learning for fully automatic kinship verification. J Vis Commun Image Represent. 2021;79:103265.
7. Qin X, Tan X, Chen S. Mixed bi-subject kinship verification via multi-view multi-task learning. Neurocomputing. 2016;214:350–7.
8. Van TN, Hoang VT. Kinship verification based on local binary pattern features coding in different color space. In: 2019 26th International Conference on Telecommunications (ICT); 2019; Hanoi, Vietnam. p. 376–80.
9. Chouchane A, Bessaoudi M, Ouamane A, Laouadi O. Face kinship verification based vgg16 and new gabor wavelet features. In: 5th International Symposium on Informatics and Its Applications; 2022; M'sila, Algeria. p. 1–6.
10. Lu J, Zhou X, Tan YP, Shang Y, Zhou J. Neighborhood repulsed metric learning for kinship verification. IEEE Trans Pattern Anal Mach Intell. 2013;36(2):331–45.
11. Ramazankhani F, Yazdian-Dehkord M, Rezaeian M. Feature fusion and NRML metric learning for facial kinship verification. J Universal Comput Sci. 2023;29(4):326. doi:10.3897/jucs.89254.
12. Huang S, Lin J, Huangfu L, Xing Y, Hu J, Zeng DD. Adaptively weighted k-tuple metric network for kinship verification. IEEE Trans Cybern. 2022;53(10):6173–86. doi:10.1109/TCYB.2022.3163707.
13. Wang W, You S, Karaoglu S, Gevers T. A survey on kinship verification. Neurocomputing. 2023;525(2):1–28. doi:10.1016/j.neucom.2022.12.031.

14. Serraoui I, Laiadi O, Ouamane A, Dornaika F, Taleb-Ahmed A. Knowledge-based tensor subspace analysis system for kinship verification. Neural Netw. 2022;151(16):222–37. doi:10.1016/j.neunet.2022.03.020.
15. Chen X, Li C, Zhu X, Zheng L, Chen Y, Zheng S, et al. Deep discriminant generation-shared feature learning for image-based kinship verification. Signal Process: Image Commun. 2022;101(2):116543. doi:10.1016/j.image.2021.116543.
16. Li W, Zhang Y, Lv K, Lu J, Feng J, Zhou J. Graph-based kinship reasoning network. In: 2020 IEEE International Conference on Multimedia and Expo (ICME); London, UK; 2020. p. 1–6.
17. Dehghan A, Ortiz EG, Villegas R, Shah M. Who do I look like? Determining parent-offspring resemblance via gated autoencoders. In: IEEE Conference on Computer Vision and Pattern Recognition; 2014; Columbus, OH, USA. p. 1757–64.
18. Wang M, Li Z, Shu X, Jingdong, Tang J. Deep kinship verification. In: 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP); 2015; Xiamen, China. p. 1–6.
19. Wiley V, Lucas T. Computer vision and image processing: a paper review. Int J Artif Intell Res. 2018;2(1):29–36. doi:10.29099/ijair.v2i1.42.
20. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2017; San Francisco, CA, USA. p. 4278–84.
21. Tomasini UM, Petrini L, Cagnetta F, Wyart M. How deep convolutional neural networks lose spatial information with training. Mach Learn: Sci Technol; 2023,4(4):045026.
22. Dai Z, Liu H, Le QV, Tan M. Coatnet: marrying convolution and attention for all data sizes. Adv Neural Inf Process Syst. 2021;34:3965–77.
23. Gao S, Li ZY, Han Q, Cheng MM, Wang L. Efficient receptive field search for convolutional neural networks. IEEE Trans Pattern Anal Mach Intell. 2022;45(3):2984–3002. doi:10.1109/TPAMI.2022.3183829.
24. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, et al. MLP-Mixer: an all-MLP architecture for vision. Adv Neural Inf Process Syst. 2021;34:24261–72.
25. Dvoršak G, Dwivedi A, Štruc V, Peer P, Emeršič Ž. Kinship verification from ear images: an explorative study with deep learning models. In: 2022 International Workshop on Biometrics and Forensics; 2022; Salzburg, Austria. p. 1–6.
26. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Why did you say that? arXiv:161107450. 2016.
27. Peng JL, Chang KW, Lai SH. KFC: kinship verification with fair contrastive loss and multi-task learning. arXiv:230910641. 2023.
28. Nandy A, Mondal SS. Kinship verification using deep siamese convolutional neural network. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019); IEEE; 2019. p. 1–5.
29. Choras M, Choras RS. Geometrical algorithms of ear contour shape representation and feature extraction. In: 6th International Conference on Intelligent Systems Design and Applications; IEEE; 2006. Vol. 6, p. 451–6.
30. Ziga Emeršič Vv, Peer P. Ear recognition: more than a survey. Neurocomputing. 2017;255(3):26–39. doi:10.1016/j.neucom.2016.08.139.
31. Shailaja D, Gupta P. A simple geometric approach for ear recognition. In: 9th International Conference on Information Technology (ICIT'06); IEEE; 2006. p. 164–7.
32. Sinha H, Manekar R, Sinha Y, Ajmera PK. Convolutional neural network-based human identification using outer ear images. In: Soft computing for problem solving. Singapore: Springer; 2019. p. 707– 19.
33. Alshazly H, Linse C, Barth E, Martinetz T. Deep convolutional neural networks for unconstrained ear recognition. IEEE Access. 2020;8:170295–310. doi:10.1109/ACCESS.2020.3024116.
34. Meng D, Nixon MS, Mahmoodi S. Gender and kinship by model-based ear biometrics. In: 2019 International Conference of the Biometrics Special Interest Group (BIOSIG); 2019; Darmstadt, Germany. p. 1–5.
35. Umirzakova S, Ahmad S, Khan LU, Whangbo T. Medical image super-resolution for smart healthcare applications: a comprehensive survey. Inform Fusion; 2024;103:102075.
36. Gunturk BK, Batur AU, Altunbasak Y, Hayes MH, Mersereau RM. Eigenface-domain super-resolution for face recognition. IEEE Trans Image Process. 2003;12(5):597–606. doi:10.1109/TIP.2003.811513.

37. Chen X, Wang X, Zhou J, Qiao Y, Dong C. Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; Nashville, TN, USA.
38. Chen X, Wang X, Zhang W, Kong X, Qiao Y, Zhou J, et al. Hat: hybrid attention transformer for image restoration. arXiv:230905239. 2023.
39. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncert, Fuzziness Knowl-Based Syst. 1998;6(2):107–16. doi:10.1142/S0218488598000094.
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, NV, USA. p. 770–8.
41. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. RepVGG: making VGG-style ConvNets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021; Nashville, TN, USA. p. 13733–42.
42. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City, UT, USA. p. 7132–41.
43. Guo MH, Lu CZ, Liu ZN, Cheng MM, Hu SM. Visual attention network. Comput Visual Media. 2023;9(4):733–52. doi:10.1007/s41095-023-0364-2.
44. Zhang QL, Yang YB. SA-Net: shuffle attention for deep convolutional neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021; Toronto, ON, Canada.
45. Misra D, Nalamada T, Arasanipalai AU, Hou Q. Rotate to attend: convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021. Waikoloa, HI, USA.
46. Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv:160608415. 2016.
47. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:14091556. 2014.
48. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, HI, USA. p. 4700–8.
49. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning; 2019; Long Beach, CA, USA. p. 6105–14.
50. Mingxing T, Quoc L. Efficientnetv2: smaller models and faster training. In: International Conference on Machine Learning; PMLR; 2021. p. 10096–106.
51. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:201011929. 2020.
52. Yu J, Li M, Hao X, Xie G. Deep fusion siamese network for automatic kinship verification. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition; 2020; Buenos Aires, Argentina. p. 892–9.
53. Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021; Waikoloa, HI, USA. p. 3560–9.
54. Li Y, Yao T, Pan Y, Mei T. Contextual transformer networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2022;45(2):1489–500. doi:10.1109/TPAMI.2022.3164083.