Hybrid Transfer Learning Using VGG Networks for Robust Ear Biometric Recognition

Mahamadabrar D. Maneri[1]*, Mrunal S. Bewoor[2] and Ashvini A. Todkar[3]

[1] College of Engineering, Bharati Vidyapeeth (Deemed to be University), Pune, India
*abrarmaneri@gmail.com
[2] College of Engineering, Bharati Vidyapeeth (Deemed to be University), Pune, India
msbewoor@bvucoep.edu.in
[3] Annasaheb Dange College of Engineering and Technology (ADCET), Ashta
ash.todkar@gmail.com

Abstract. Ear biometrics offers a reliable means of personal identification, as the shape of the ear remains largely stable across a person's lifetime and is unaffected by facial expressions. However, existing deep neural network methods for ear recognition perform poorly in unconstrained, real-world conditions. To address this, we propose a Hybrid Transfer Learning framework based on VGG Networks, which combines the strengths of multiple pre-trained models through score-level ensemble fusion. Four VGG variants - VGG11, VGG13, VGG16, and VGG19 - are individually fine-tuned and evaluated on two benchmark datasets: the controlled AMI Ear Database and the in-the-wild EarVN1.0 dataset. The top-performing pair on each dataset is selected by validation accuracy and combined to form the hybrid model. The proposed ensemble achieves 96.67% recognition accuracy on AMI and 73.96% on EarVN1.0, outperforming every individual VGG variant on both benchmarks.

Keywords: Ear Biometrics, Transfer Learning, VGG Networks, Hybrid Model, Ensemble Fusion, Robust Recognition, Deep Learning, AMI Dataset, EarVN1.0 Dataset.

1 Introduction

Modern security applications heavily depend on the latest technology in biometric recognition. Because of its unique shape, which is unlikely to change much throughout a person's lifetime, ear recognition technology is a fascinating method of biometrics. The human ear is highly resistant to distortion over time, making it a stable identifier that can be used in various fields, including access control and forensic analysis [1, 6, 12]. The fact that it can be externally visible makes it a highly practical choice in situations where surveillance or authentication is necessary without having to bother users, where methods such as fingerprint readers or eye scanners may be intrusive or impractical.

M. Kaushik et al. (eds.), *Proceedings of the International Conference on Responsible, Risk-aware, and Regulated AI (RRRAI 2026)*, Advances in Intelligent Systems Research 210,
https://doi.org/10.2991/978-94-6239-723-1_33

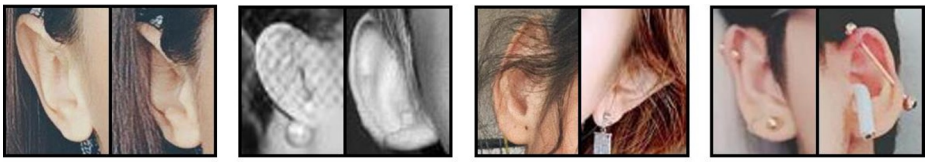

Fig. 1. Illustration of environmental and physical perturbations in the ear biometric [9].

Despite its advantages, ear-based recognition faces several practical challenges. The perceived shape of the ear varies with viewing angle and lighting conditions, while partial occlusions from hair, accessories, or other environmental factors further complicate recognition [3, 5, 8]. These conditions demand feature extraction methods that perform reliably in normal circumstances and remain robust under real-world noise and interference [4, 18].

Traditional ear recognition relied on conventional image processing, which proved insufficient for datasets with significant variation [12]. Deep learning shifted this paradigm, with CNNs learning features directly from raw images [13, 17]. However, such models risk overfitting when training data is limited, reducing generalisation [2, 11]. Transfer learning addresses this by fine-tuning models pre-trained on large datasets such as ImageNet for the ear recognition task [10, 20].

To address these limitations, this paper proposes a hybrid transfer learning approach built on the VGG network family. All four variants - VGG11, VGG13, VGG16, and VGG19 - are individually fine-tuned, and the best-performing pair on each dataset is combined via score-level ensemble fusion. Evaluated on the AMI Ear Database and EarVN1.0, the method shows improved robustness to scale, pose, illumination, and occlusion variations, yielding a meaningful gain in recognition reliability for practical security deployments.

2 Related Work

2.1 Ear Recognition Methods

The study of the shape and size of the human ear for the purposes of identification has developed from basic geometric methods into modern methods reliant on complex data analysis. Research in the field has shown that ear recognition is a valid biometric [1, 6, 12]. In controlled environments, these approaches were successful, but they failed when applied in real-world conditions. Researchers consequently developed and explored texture-based features such as scale-invariant feature transforms and local binary patterns [4, 18].

The field has recently been transformed by deep learning technology. CNNs allow for end-to-end learning from an ear image. For instance, Dodge et al. [5] pioneered unconstrained ear detection through the use of Residual Learning architectures; notable performance improvements were demonstrated when compared with traditional ear detection methodologies. Similarly, Priyadharshini et al. [13] employed a deep neural network employing a convolutional architecture for facial identity tasks. Nonetheless, their

model performed poorly in tests involving images with objects covering faces. Both Mollona et al.[11] and Sharkas et al.[15] have employed the application of multi-scale convolutions combined with attention mechanisms to enhance the invariance properties of the system.

2.2 Transfer Learning in Biometrics

Transfer learning has proven instrumental in biometrics by adapting ImageNet-pre-trained weights to domain-specific tasks. In ear biometrics specifically, Alshazly et al. [2] demonstrated the effectiveness of fine-tuning VGG and ResNet backbones, producing ensembles that surpassed baseline classifiers on mixed datasets. Their follow-up work [3] extended this to unconstrained scenarios, emphasising domain adaptation to bridge the gap between synthetic and real imagery. Hybrid fusions have further amplified these benefits; Hansley et al. [8] combined learned embeddings with handcrafted edge features for occlusion-robust matching [10]. Despite these advances, transfer learning applications remain fragmented across studies [19, 20], limiting generalisation across varying illumination and pose conditions.

2.3 VGG-Based Hybrid Architectures in Biometrics

VGG networks have served as a foundation for hybrid CNN applications in biometrics. Tian and Mu [16, 17] adapted VGG variants for ear classification and achieved strong results on small controlled databases. Sarangi et al. [20] proposed combining predictions from multiple machine learning models to improve classification, while Kumar and Singh [19] demonstrated noise robustness through a deep ensemble of VGG-style feature extractors. Shallower VGG variants preserve fine spatial detail, whereas deeper variants capture broader contextual representations [2, 15]. However, prior work rarely applies a systematic, validation-guided approach to select the optimal model pair for fusion [14], and most evaluations are restricted to constrained conditions. The present work addresses both gaps: VGG11 through VGG19 are comprehensively assessed under the same training protocol, and the best-performing pair is selected on a per-dataset basis, yielding a 4-7% accuracy improvement over the strongest individual model.

3 Datasets

In order to properly test the robustness of the proposed hybrid learning approach, we evaluate it on two standard databases that cover a range of acquisition conditions. The controlled and uncontrolled ear databases comprise the AMI ear database [7] and the EarVN 1.0 database [9]. This framework for the evaluation of biometric systems allows performance in ideal and realistic conditions to be compared.

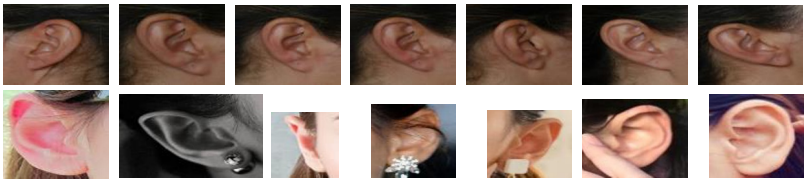

Fig. 2. Representative samples from the evaluation datasets. (Top row) AMI Ear Database. (Bottom row) EarVN1.0 Dataset.

Table 1. Dataset Characteristics.

Dataset	Subjects	Images	Environment	Key Challenges
AMI Ear [7]	100	700	Controlled (indoor, uniform light)	Pose tilts, focal shifts
EarVN1.0 [9]	164	28,412	Unconstrained (in-the-wild)	Pose, scale, illumination, occlusion

3.1 AMI Ear Database

The University of Las Palmas de Gran Canaria's AMI Ear database is a controlled [7], benchmark resource. This dataset comprises 700 images captured by 100 people. The photographs were taken in an indoor environment with consistent lighting. Six of these views are recorded from the subject's right ear, using six different angles - frontal, 20 degrees up from the front, 20 degrees down from the front, from the side, from the front again but looking at the back, and finally a frontal view taken at a 200 mm focal length with the subject 135 mm away. The main challenges include geometric transformations caused by tilting and zooming, resulting in blurring. In a controlled test environment, noise is reduced by using a standard image setup.

3.2 EarVN1.0 Dataset

In contrast, the EarVN1.0 dataset [9] is an unconstrained collection of 28,412 images gathered from 164 Vietnamese volunteers using handheld mobile cameras in everyday settings. Images vary in resolution and are subject to inconsistent lighting, viewpoint differences, and frequent occlusions from hair, glasses, or scarves. This high intra-class variability necessitates substantial data augmentation during training to ensure the model generalises rather than overfits to a narrow distribution of appearances.

4 Proposed Methodology

In the proposed method, several transfer learning approaches utilising the VGG network are combined to obtain robust ear identification results. This approach uses a number of different, pre-trained models and combines the best results from them in order to overcome variations in the pose of figures, changes in lighting, differences in size and partial obscuration of the images. [2, 15, 20].

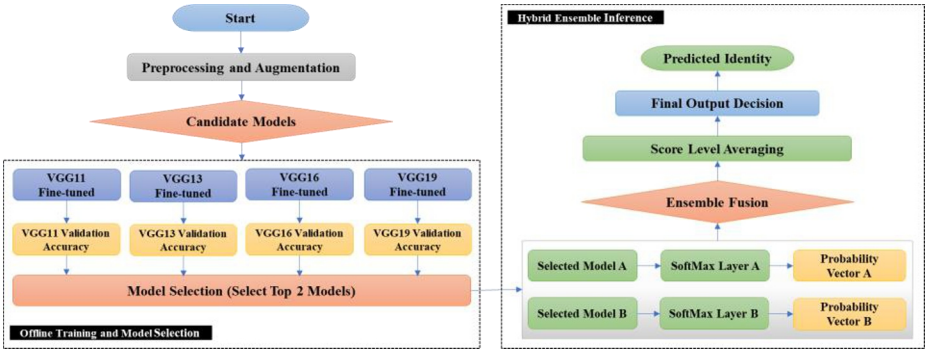

Fig. 3. Overview of the proposed Ensemble of VGG Networks architecture.

4.1 Transfer Learning with VGG Architectures

The network utilises four VGG variants with batch normalisation, which have been pre-trained on the ImageNet database [16]. The architecture of these models features a uniform progressive deepening of their convolutional stacks [2].

For each variant, the classification layer is replaced with a customised head tailored to the number of speakers in the target database [13, 17]. Each of the network's levels is next trained using supervised learning to fine-tune the results obtained so far. The approach enables the networks to maintain crucial early-layer representations, which are required for the differentiation of textures and contours. This also allows the adaptation of generic knowledge from the image database into ear-specific features [13, 17].

4.2 Model Selection and Hybrid Ensemble Fusion

Each of the models fine-tuned has its own respective set of characteristics due to variations in depth and field value that it can perceive. To exploit the benefits of this complementary relationship, we combine the two best models in each set for predictions.

The ensemble model works at the image-level, where the model's SoftMax probabilities are averaged:

$$P_{\text{hybrid}} = \frac{1}{2}(P_{\text{VGG}_a} + P_{\text{VGG}_b})$$

The final classification is determined by taking the maximum likelihood from the hybrid model's probabilities. The framework presented offers a trade-off between inference speed and the quality of the results produced, which proves to be effective in a wide range of applications.

5 Experimental Setup

In order to ensure the reproducibility of the results obtained and allow a fair comparison between the models, experiments were carried out under controlled conditions.

5.1 Hardware and Software Environment

This was implemented using the PyTorch library, with Python 3.10 as the programming language. The experiments were conducted on an NVIDIA GeForce RTX 3080 GPU, running under CUDA 11.8 for computation. Automatic mixed precision training was used in order to accelerate the training process and decrease the memory needed.

5.2 Image Preprocessing and Augmentation

The images used were preprocessed. Each of the ear images is resized to 224 by 224 pixels. The images were standardised with the mean of 0.485, 0.456, and 0.406 and the standard deviation of 0.2247, 0.2247, and 0.2255 from ImageNet [16]. The AMI dataset [7] grayscale images were replicated across the colour channels.

Extensive data augmentation was used in training to increase robustness, with particular emphasis on the variations present in the EarVN1.0 database [9]. These variations include random cropping at a scale of 0.6 to 1.0, random flipping of images horizontally at a 50% chance, rotation by as much as 20 degrees to either side, random colour modification to alter brightness, contrast, and saturation, and the occasional blurring of an image using a Gaussian blur. In the validation and testing phases, deterministic transformations were utilised; in this case, images were resized to a pixel dimension of 256 and then centred and cropped to 224x224 pixels.

5.3 Training Protocol

An 80-20 train-validation split was applied to both data collections. After model validation, the remaining 20% of the data was used for final testing. A common methodology was utilised across all experiments; AdamW was used as the optimiser with a learning rate of 0.001 and a weight decay of 1×10^{-3}. This experiment used cross-entropy as the loss function, with 32 samples per dataset fed into it at one time. The training continued for a maximum of 100 iterations. A cosine learning rate schedule with period for the learning rate and with period for the second moment of the gradients was applied. This was with a warm restart every iteration.

5.4 Model Selection and Ensemble Construction

All of the variants of the VGG network (VGG11, VGG13, VGG16, and VGG19 using batch normalisation) were fine-tuned individually. Each model was ranked by its validation accuracy on the individual data sets after training was complete. The top-performing pair for the ensemble method selected on AMI data was VGG 11 and VGG 16. On the other hand, the top pair selected for the hybrid ensemble on EarVN1.0 data were VGG 11 and VGG 19. During inference, the hybrid model was created by averaging the SoftMax probabilities of the selected pair as explained in section 4.2. In Section 6, the performance metrics presented are the top one accuracy rates on the unseen test groups.

6 Results and Discussion

6.1 Quantitative Evaluation

Table 2 summarises the top-1 recognition accuracy and parameter count for each VGG variant and the proposed hybrid ensemble across both datasets, confirming that the ensemble consistently outperforms any single model.

Table 2. Comparative recognition accuracies and parameters across models and datasets.

Model	AMI (%)	EarVN1.0 (%)	Parameters (Millions)
VGG11	96.00	70.28	9.4
VGG13	95.00	68.85	9.9
VGG16	95.33	69.21	14.7
VGG19	90.67	72.12	20.0
Hybrid (Avg.)	96.67	73.96	24.1

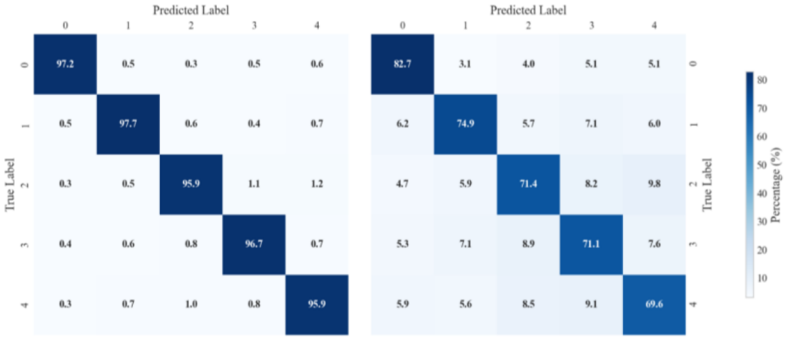

Fig. 4. Example confusion matrices illustrating classification performance on subsets of the AMI (left) and EarVN1.0 (right) datasets for the hybrid model.

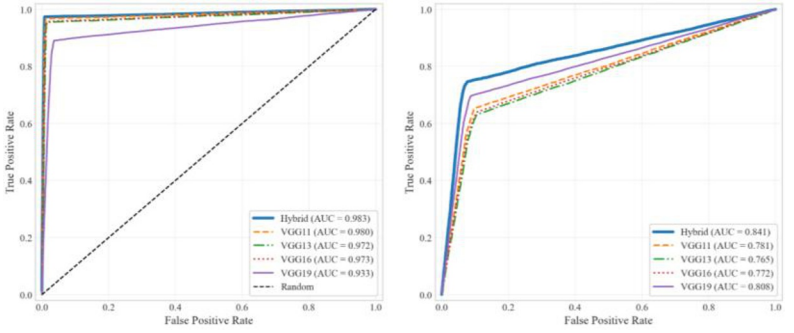

Fig. 5. ROC curves comparing individual VGG models against the proposed hybrid ensemble on the AMI (left) and EarVN1.0 (right) datasets.

6.2 Ablation Study

The proposed hybrid approach (-VGG 11 + VGG 19) on the Ear-VN1.0 database reached an accuracy rate of 73.96%. However, an ensemble of two VGG13 networks possessed a slightly lower accuracy of 71.45% when the two models had the same depth. Pairing the deepest variants (VGG19 and VGG16) led to a 72.38 % classification accuracy, a lower result than the best accuracy of the combination of a shallow and a deep network.

This occurs when either random or similar error depths are paired together, failing to capitalise on this independence. As a result, this leads to duplicated error patterns and a reduced overall performance from the ensemble. [2, 15, 20]. This experiment demonstrates the value of choosing models based on the data as opposed to choosing an arbitrary model or an ensemble of a fixed depth. This reinforces the strategy proposed of selecting the model based on the validation results.

6.3 State-of-the-Art Comparison

In order to put our results into perspective, a comparison is made in Table 3 of the proposed model with other systems that have been reported in the literature on the specified tasks.

Table 3. Performance comparison with existing ear recognition methods.

Study	Method	Dataset	Accuracy (%)
Alshazly et al. [2]	VGG/ResNet Ensemble	AMI	~97.5
Priyadharshini et al. [13]	Custom CNN	AMI	~97.5
Alshazly et al. [3]	Deep CNNs	EarVN1.0	~70-72
Mohamed et al. [11]	Advanced Deep Learning	AMI/ EarVN1.0	Up to 99.35 / ~98 (reported on variants)
This Work	Hybrid VGG Ensemble	AMI	96.67
This Work	Hybrid VGG Ensemble	EarVN1.0	73.96

6.4 Discussion

This marginal gap likely reflects the over-parameterisation of deeper VGG variants relative to the relatively small, low-diversity AMI dataset, which raises the risk of overfitting [13, 16].

EarVN1.0 presents substantial variation in pose, illumination, scale, and occlusion. Among individual models, VGG19 achieved the highest accuracy of 72.12%, benefiting from its deeper architecture's capacity to learn complex feature representations. The hybrid ensemble (VGG11 + VGG19) reached 73.96%, a 1.84% improvement, by combining the global context captured by the deeper network with the fine-grained spatial detail preserved by the shallower one [2, 15, 20].

Although the hybrid model carries a higher parameter count of 24.1 million compared to individual variants, this remains within acceptable bounds for deployment in real-world security systems where accuracy and robustness are prioritised over model compactness [10]. In practice, the system can be deployed by: (1) pre-processing input images to 224×224 pixels with ImageNet normalisation; (2) running inference through the two selected VGG models in parallel; and (3) averaging their SoftMax output vectors to produce the final identity prediction. This pipeline is straightforward to integrate into existing access control or forensic analysis infrastructure.

7 Conclusion and Future Work

This paper presented a hybrid transfer learning framework for ear biometric recognition using an ensemble of VGG network variants. The proposed method achieved 96.67% accuracy on the controlled AMI database and 73.96% on the unconstrained EarVN1.0 dataset [1, 6]. By selecting the best-performing model pair through validation and combining their SoftMax outputs, the approach effectively addresses real-world challenges, including appearance variation, inconsistent lighting, scale differences, and partial occlusion [3, 5, 8].

Future work will explore alternative fusion strategies, such as weighted or learned score combination, and more compact architectures to reduce inference overhead. Integrating an automated ear detection and localisation module would produce a fully end-to-end biometric pipeline [14, 17]. Evaluating the approach on additional diverse datasets and investigating lightweight variants suitable for edge deployment are also planned directions.

References

1. Abaza A, Ross A, Hebert C, Harrison MAF, Nixon MS (2013) A survey on ear biometrics. ACM Comput Surv 45:1–35. doi: 10.1145/2431211.2431221
2. Alshazly H, Linse C, Barth E, Martinetz T (2019) Ensembles of Deep Learning Models and Transfer Learning for Ear Recognition. Sensors 19:4139. doi: 10.3390/s19194139
3. Alshazly H, Linse C, Barth E, Martinetz T (2020) Deep Convolutional Neural Networks for Unconstrained Ear Recognition. IEEE Access 8:170295–170310. doi: 10.1109/ACCESS.2020.3024116
4. Booysens A, Viriri S (2022) Ear Biometrics Using Deep Learning: A Survey. Applied Computational Intelligence and Soft Computing 2022:1–17. doi: 10.1155/2022/9692690
5. Dodge S, Mounsef J, Karam L (2018) Unconstrained ear recognition using deep neural networks. IET Biom 7:207–214. doi: 10.1049/iet-bmt.2017.0208
6. Emeršič Ž, Štruc V, Peer P (2017) Ear recognition: More than a survey. Neurocomputing 255:26–39. doi: 10.1016/j.neucom.2016.08.139
7. Esther Gonzalez LA and LM (2012) AMI Ear Database. In: University of Las Palmas de Gran Canaria. https://ctim.ulpgc.es/research_works/ami_ear_database/. Accessed 31 Dec 2025

8. Hansley EE, Segundo MP, Sarkar S (2017) Employing Fusion of Learned and Handcrafted Features for Unconstrained Ear Recognition. IET Biometrics. 7: 215–223. arXiv:1710.07662
9. Hoang VT (2019) EarVN1.0: A new large-scale ear images dataset in the wild. Data Brief 27:104630. doi: 10.1016/j.dib.2019.104630
10. Kamboj A, Rani R, Nigam A (2022) A comprehensive survey and deep learning-based approach for human recognition using ear biometric. Vis Comput 38:2383–2416. doi: 10.1007/s00371-021-02119-0
11. Mohamed Y, Youssef Z, Heakl A, Zaky A (2024) Advancing Ear Biometrics: Enhancing Accuracy and Robustness through Deep Learning. arXiv:2406.00135
12. Pflug A, Busch C (2012) Ear biometrics: a survey of detection, feature extraction and recognition methods. IET Biom 1:114–129. doi: 10.1049/iet-bmt.2011.0003
13. Ahila Priyadharshini R, Arivazhagan S, Arun M (2021) A deep learning approach for person identification using ear biometrics. Applied Intelligence 51:2161–2172. doi: 10.1007/s10489-020-01995-8
14. Ramos-Cooper S, Gomez-Nieto E, Camara-Chavez G (2022) VGGFace-Ear: An Extended Dataset for Unconstrained Ear Recognition. Sensors 22:1752. doi: 10.3390/s22051752
15. Sharkas M (2022) Ear recognition with ensemble classifiers; A deep learning approach. Multimed Tools Appl 81:43919–43945. doi: 10.1007/s11042-022-13252-w
16. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556
17. Tian L, Mu Z (2016) Ear recognition based on deep convolutional network. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, pp 437–441
18. Yuan L, Mu Z, Xu Z (2005) Using Ear Biometrics for Personal Recognition. pp 221–228
19. Mehta R, Singh KK (2024) An efficient ear recognition technique based on deep ensemble learning approach. Evolving Systems 15:771–787. doi: 10.1007/s12530-023-09505-0
20. Mehta R, Singh KK (2024) Ensemble of transfer learning and light-weight convolutional neural network model for an effective ear recognition system. Evolving Systems 15:115–131. doi: 10.1007/s12530-023-09561-6