2023 IEEE International Joint Conference on Biometrics (IJCB) | 979-8-3503-3726-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCB57857.2023.10449062

The Unconstrained Ear Recognition Challenge 2023: Maximizing Performance and Minimizing Bias*

Ž. Emeršič[1,*], T. Ohki[2], M. Akasaka[2], T. Arakawa[2], S. Maeda[2], M. Okano[2], Y. Sato[2], A. George[3], S. Marcel[3], I. I. Ganapathi[4], S. S. Ali[4], S. Javed[4], N. Werghi[4], S. G. Işık[5,6], E. Sarıtaş[5], H. K. Ekenel[5], V. Hudovernik[1], J. N. Kolf[7,8], F. Boutros[7], N. Damer[7,8], G. Sharma[9], A. Kamboj[10], A. Nigam[9], D. K. Jain[11], G. Cámara-Chávez[12], P. Peer[1], V. Štruc[1]

[1]University of Ljubljana (UL, SI), [2]Shizuoka University (SU, JP), [3]Idiap Research Institute (IDIAP, CH), [4]Electrical Engineering and Computer Science Department, Khalifa University (KU, UAE), [5]Istanbul Technical University, Department of Computer Engineering (ITU, TR), [6]Microelectronic Guidance and Electro-Optical Group, ASELSAN, [7]Fraunhofer Institute for Computer Graphics Research IGD (IGD, DE), [8]Technische Universität Darmstadt (TUD, DE), [9]Indian Institute of Technology Mandi (IIT, IN), [10]HCL Imaging and Robotics R&D Lab (HCL, IN), [11]School of Artificial Intelligence, Dalian University of Technology (DLUT, CN), [12]Federal University of Ouro Preto (UFOP, BR)

Abstract

The paper provides a summary of the 2023 Unconstrained Ear Recognition Challenge (UERC), a benchmarking effort focused on ear recognition from images acquired in uncontrolled environments. The objective of the challenge was to evaluate the effectiveness of current ear recognition techniques on a challenging ear dataset while analyzing the techniques from two distinct aspects, i.e., verification performance and bias with respect to specific demographic factors, i.e., gender and ethnicity. Seven research groups participated in the challenge and submitted a seven distinct recognition approaches that ranged from descriptor-based methods and deep-learning models to ensemble techniques that relied on multiple data representations to maximize performance and minimize bias. A comprehensive investigation into the performance of the submitted models is presented, as well as an in-depth analysis of bias and associated performance differentials due to differences in gender and ethnicity. The results of the challenge suggest that a wide variety of models (e.g., transformers, convolutional neural networks, ensemble models) is capable of achieving competitive recognition results, but also that all of the models still exhibit considerable performance differentials with respect to both gender and ethnicity. To promote further development of unbiased and effective ear recognition models, the starter kit of UERC 2023 together with the baseline model, and training and test data is made available from: http://ears.fri.uni-lj.si/.

*This research was supported in parts by the ARRS Research Programmes P2-0250(B) "Metrology and Biometric Systems", P2-0214 "Computer Vision", and TUBITAK Research Programmes 120N011, 2210 "Graduate Scholarship Program" and Turkcell Research Scholarship Program.

1. Introduction

Ear recognition is an important area of research within biometrics [4, 29, 32]. However, existing work in this field has mostly been focused on maximizing raw recognition performance, while other aspects, critical for the deployment of biometrics recognition techniques in practice, have largely been ignored. One such example is demographic bias [13]. Modern ear recognition approaches are not only expected to be highly effective when recognizing individuals, but also to be equally fair and reliable in their decisions, regardless of the demographic characteristics of the subjects with respect to, e.g., gender or ethnicity. In fact, the fairness of the automated decisions is a key aspect of ear recognition models that has important implications for their trustworthiness and real-world deployment [24].

While a considerable amount of work has been done on studying bias with different biometrics modalities [3, 8, 14, 28], on understanding the corresponding reasons and causes for the presence of bias [1, 22], and on designing mitigation measures [12, 15], research on demographic bias and performance differentials caused by demographic factors in ear recognition models is still limited in the open literature. Consequently, important research questions remain unanswered, e.g., "Are ear recognition techniques biased?", "How does gender and ethnicity impact performance of modern ear recognition models?", "Do certain model architectures lead to stronger performance and less biased results?", "Is it possible to improve verification performance and reduce bias given a fixed model design?".

To answer these and related questions, the third Unconstrained Ear Recognition Challenge was organized in the scope of the 2023 IEEE international Joint Conference on Biometrics (IJCB). The idea behind the challenge was to provide a platform for evaluating ear recognition models under a common experimental protocol, while focusing on two distinct aspects, i.e., verification performance and de-

mographic bias. Thus, the participants of the challenge were asked to design models that not only perform well across a wide range of diverse ear images, captured in challenging unconstrained conditions, but that also lead to limited performance differentials with respect to variations in gender and ethnicity. A joint evaluation criterion was, therefore, defined that penalized verification errors, but also variations in performance due to demographics.

Another task addressed as part of UERC 2023 was model improvement. Here, participants were given an existing ear recognition model and were tasked with the design of training strategies, model components and learning objectives that yielded improved recognition results and/or reduced bias. A total of seven research groups participated in UERC 2023 and submitted seven distinct models that ranged from ensemble methods, convolutional neural networks (CNNs), transformers and hybrids between learned and handcrafted models. All models were evaluated on a sequestered test dataset and analyzed for performance and bias behavior. The research effort of the participants led to the following main contributions that are presented in this paper:

- A comprehensive evaluation and comparative analysis of seven state-of-the-art ear recognition models with a focus on performance as well as demographic bias. The evaluation includes benchmarking of the models on a challenging (new) ear dataset.
- An in-depth analysis of the impact of gender and ethnicity on the performance of the evaluated ear recognition models.

2. Related work

UERC 2023 is the third in the series of Unconstrained Ear Recognition Challenges, initially started in 2017. The first UERC was organized in the scope of the 2017 edition of the International Joint Conference on Biometrics (IJCB) [31] and introduced one of the first large-scale ear recognition benchmarks with appearance-rich ear images, collected from the internet in so-called in-the-wild conditions. UERC 2017 explored the identification performance of the submitted ear recognition models in the presence of a large number of distractor identities. The second UERC was held as part of the 2019 IAPR International Conference on Biometrics (ICB) [30]. This edition of UERC again focused on identification experiments, but also investigated the sensitivity of the submitted models to image resolution, head rotations and presence of occlusions. UERC 2023 builds on the previous editions of the challenge, but investigates ear-recognition models in a verification setting, while also taking demographic bias into account. The goal of the competition, is, hence, to promote the development of more accurate, fair and unbiased recognition techniques that are less likely to produce errors or false positives for certain groups of people and are, therefore, well suited to be deployed in practice.

3. Methodology

In this section, we present the methodology adopted for UERC 2023. We first discuss the organization of the challenge and the experimental datasets used, and then describe the experimental protocol, performance metrics, and starter kit distributed to the challenge participants.

3.1. UERC 2023 Organization

UERC 2023 was organized as a two-track competition, where each track focused on one specific goal. Participants were free to enter only a single track or compete in both. A detailed description of the two tracks is given below.

Track 1: Fair Ear Recognition. The idea of the first UERC 2023 evaluation track was to collect ear recognition models and analyze their behavior on ear images captured in unconstrained environments. Different from previous challenges, we were not interested solely in the recognition performance, but also in the fairness of the submitted models. Thus, the participants were asked to develop ear-recognition solutions that perform well in verification experiments but also exhibit limited performance differentials (i.e., low demographic bias) with respect to different demographic groups. To rank the submissions, a performance measure was designed that takes verification errors as well as demographic bias into account. The participants were free to develop any type of model that (in their opinion) maximized performance, while minimizing bias. The final submissions for this track included the computed feature vectors over a sequestered dataset and a working solution (i.e., source code or a compiled binary), which the organizers ran to score the submissions.

Track 2: Model Improvement. The second UERC 2023 track addressed model improvement strategies. Here, a baseline ResNet-18 model was made available to the participants and the goal was to implement strategies that maximize the performance of the model, while minimizing the demographic bias (due to gender and ethnicity). Thus, the model architecture in the second track was fixed, and the participant were asked to improve the provided model towards better verification performance, reduced bias or, ideally, both. To score performance, differential performance indicators were designed that compared the submissions to the UERC baseline. Similarly to the first track, participants had to submit feature vectors and a working solution that the organizers evaluated on the sequestered test dataset.

3.2. Training and Testing Data

The training and testing data for UERC 2023 consisted of images captured in-the-wild. Such images exhibit a con-

Figure 1. **Example images from two subjects (in rows) from the sequestered test dataset of UERC 2023.** The test data was sequestered from the training data and not made available to the participants to ensure fair evaluation results. The data was captured in-the-wild and exhibits considerable appearance variability.

siderable degree of appearance variability and are, therefore, particularly challenging for existing ear recognition models. A few example images are presented in Figure 1.

Training Data. For training and model development, the following data was provided to the UERC participants:

- A newly collected dataset of a total of $14,004$ images of 650 distinct subjects. Images were harvested from the web, with $2,304$ of them taken from the training images from UERC 2017 and UERC 2019.
- Over $234,651$ images of 660 subjects from the VGGFace-Ear dataset [26]. The data for this part was generated by cropping the ear region from face images from the VGGFace dataset and then normalizing the cropped regions to a fixed size.

The training dataset was made available to allow the participants to train their models and select suitable hyperparameters. Additionally, a validation split was proposed that enabled progress tracking and performance evaluation during model training. However, the participants were free to split the training set as they wished and also use additional external training data for the development.

The training data provided by the organizers was annotated with binary gender, i.e., female (f), male (m), and 7 ethnicity labels, i.e., Caucasian (1), Asian (2), South Asian (3), Black (4), Middle Eastern (5), Hispanic (6), and Other (7). Identity labels were also made available.

Testing Data. The testing data was sequestered and made available to the participants without identity labels. The sequestered test data consisted of *six distinct groups* of ear images with different ethnicity-gender combinations from the following base categories: (i) **Ethnicities**: Asian, Black and White, and (ii) **Gender categories**: Female and Male. Each ethnicity-gender group in the sequestered test dataset consisted of 10 subjects and around 250 images, resulting in a total of $1,670$ images that were available for ranking of the submitted approaches and the analysis of their behavior.

3.3. Experimental Setup

Experimental Protocol. To score the developed models in both of the UERC 2023 competition tracks, participants were requested to submit feature representations of the testing data to the organizers. The organizers then performed a matching procedure using the cosine similarity to compute the relevant performance indicators and analyze the developed models. For the matching procedure, within-group experiments were considered, where all mated images from a selected ethnicity-gender group were compared against each other to produce mated comparison scores, and all non-mated image pairs from the same group were computed to generate non-mated comparison scores for the evaluation. The presented procedure was performed for each demographic group and each submitted model separately.

Performance Indicators for Track 1. The main goal of the first UERC 2023 track was to evaluate both verification performance as well as demographic bias of the developed ear recognition models. A joint evaluation criterion R was, therefore, defined to rank the submitted approaches. The criterion is based on a weighted average between the Gini index computed over the different demographic groups and the Equal Error Rate computed over all available test data irrespective of the demographics and is defined by the following equation:

$$R = \lambda G + (1-\lambda)EER, \tag{1}$$

where smaller values indicate better performance, λ stands for the balancing weight between the two evaluation criteria, and G is the Gini coefficient, defined as:

$$G = \frac{\sum_i \sum_j |EER_i - EER_j|}{2n \sum_j EER_j} = \frac{\sum_i \sum_j |EER_i - EER_j|}{2n^2 \overline{EER_j}},$$

where, n is the number of demographic groups and EER_i is the Equal Error Rate (EER) of the i-th demographic group. Note that the indices i and j run over the same six ethnicity-gender test sets and a balancing factor of $\lambda = 0.2$ is selected to give a somewhat higher preference to verification performance and smaller to the demographics-induced performance differentials. We note that the EER is a commonly used performance indicator in biometric verification systems. The Gini coefficient, on the other hand, is a measure of bias and defined by the ratio of the sum of absolute differences between all pairs of values in a set, to the total number of possible pairs used to quantify the performance. Both measures are between 0 and 1, and were combined into the joint performance measure R that was utilized to rank the submitted approaches.

In addition to the joint evaluation criterion from Eq. (1), further performance measures were also considered to facilitate an in-depth analysis of the submitted models, including: the Area Under the (ROC) Curve (AUC), and the False Non-Match Rate (FNMT) at a 1% False Match Rate (FMR).

Performance Indicators for Track 2. For the second track, performance indicators that quantify verification errors and demographic bias were used. However, because the second UERC track is interested in the relative changes of the performance scores in comparison to the provided baseline model, we define differential measures that calculate the relative performance loss/increase as well as the relative demographic bias decrease/increase, i.e.,

$$S_p = \frac{EER' - EER}{EER};\ S_g = \frac{G' - G}{G}, \tag{2}$$

where EER' and G' denote the Equal Error Rate and Gini index after the model improvement process. To rank the submissions, a joint criterion was again defined, i.e., $S_p + 0.2Sg$, where higher preference was again given to verification performance improvements, compared to the changes in demographic bias. Lower values of this criterion correspond to better performance.

The UERC 2023 Starter Kit and Baseline. To allow for a quick start into UERC 2023, a starter kit was provided to the participants. The starter kit included the UERC baseline model, i.e., the fairly lightweight ResNet18. The model was initially pre-trained on ImageNet and then fine-tuned on the training part of the UERC dataset. Additionally, dropout regularization was used after each Conv2d layer with a probability of 0.2, which helped to reduce overfitting by randomly setting a fraction of input units to 0 at each update of the training process. The data representation from the pen-ultimate model layer was used as the feature vector for the given input image.

The baseline was trained with the Adam optimizer with a learning rate of 10^{-4}, weight decay of 10^{-3} and beta values of 0.8 and 0.999, using the Cross-Entropy loss. A batch size of 512 was adopted over 100 epochs. Model performance was evaluated at the end of each epoch on a separate validation dataset, and the model with the highest validation accuracy was used as the final baseline model. Both the training code and the model with weights was provided to the participants as part of the starter kit to have a decent starting point for development of the competition entries.

4. Summary of Participating Approaches

UERC 2023 received a total of 7 submissions from research groups belonging to 7 distinct institutions. A high-level summary of the submitted solutions is provided in Table 1 and a brief description is given below.

DHCF. The Deep HOG-CNN Fusion (DHCF) approach consists of a hybrid neural network model that incorporates both neural network components and handcrafted features in an end-to-end manner. The model comprises a spatial transformer network (STN) [20], a pre-trained CNN-based edge detector (LDC) [27], and a ResNet-18 module. The STN aims to learn an aligned representation that optimises classification performance. Subsequently, the CNN edge detector component is employed to emphasize ear shape over texture. The output from the edge detector passes through the ResNet-18 module, generating a 512-dimensional output. Concurrently, the image post-STN is processed by a Histogram of Oriented Gradients (HOG) [10] feature extractor, which yields a representation that is further passed through a two-layer fully connected layer with ReLU activation, resulting in another 512-dimensional representation. The representations from the HOG feature-extraction process and the CNN are merged using a fully connected layer in an end-to-end fashion, followed by an additional fully connected layer for classification. The model was trained on a GPU for 25 epochs with a learning rate of 1e-3 and the CrossEntropy loss. To enhance robustness, data augmentation techniques such as scale changes, rotations, horizontal flips, and variations in brightness, contrast, saturation, and hue were applied within specified ranges. All images were converted to grayscale to prevent overfitting related to skin colour.

KU-EAR. The KU-EAR solution utilizes the ResNet18 architecture as the backbone model for recognition. The model is trained using the supervised contrastive loss, which encourages ear images from the same class to have smaller distances in the embedding space while pushing different class samples apart. By leveraging the contrastive loss, the KU-EAR approach aims to enhance the discriminative power of the model for accurate ear recognition. The training was performed on the data provided by the UERC 2023 organizers, but data augmentation played a crucial role in increasing the diversity of the training set. Here, random rotations (± 10 degrees), color jitter and vertical flipping are adopted as the main augmentation techniques. Rotations help the model generalize to different orientations, color jitter increases robustness to varying lighting conditions, whereas vertical flipping introduces further appearance variations into the training dataset.

PreWAdaEar. The PreWAdaEar approach consists of a fine-tuned AdaFace model [21], a powerful recognition approach designed specifically for diverse and low-quality images. During training, PreWAdaEar is initialized with the AdaFace weights, initially learned for the face recognition task in [21]. Next, the AdamW optimizer is used to fine-tune the model for ear recognition. Here, different learning rates are adopted for different parts of the mode. Specif-

Table 1. **High-level summary of the approaches submitted to UERC 2023.** Seven groups participated in the challenge and provided seven solutions for the first track and four for the second track of the challenge. The submitted solution span a range of deep learning models, including CNNs and transformers, but also ensemble techniques and combinations of hand-crafted features and CNNs.

Organization	Model	Brief Summary	Track 2	External Data	Model Footprint
Idiap Research Institute, Switzerland	DHCF	Hybrid approach with learned CNN and handcrafted HOG features	Yes	No	$15.4M$
Khalifa University, UAE	KU-EAR	ResNet-18 trained with a supervised contrastive loss	Yes	No	$11.9M$
Istanbul Technical University, Turkey	PreWAdaEar	Quality-aware AdaFace model fine-tuned for ear recognition	No	Yes	$65.2M$
Shizuoka University, Japan	MEM-Ear	Ensembe technique with 3 CNNs and weighted feature aggregation	No	No	$97.0M$
Fraunhofer IGD, Germany	UERC-IGD	ResNet-18 trained with CosFace objective in a multi-task setting	Yes	No	$24M$
IIT Mandi, India	RecogEAR	Two-Stream Inflated 3D ConvNet trained with ArcFace loss	No	No	$12.8M$
University of Ljubljana, Slovenia	ViTEar	DINOv2 Vision Transformer trained with margin-penalty softmax losses	No	Yes	$304.9M$
Organizers	UERC Baseline	ResNet-18 with dropout layers	No	No	11.9M

ically, learning rates of 1, 1/2, and 1/10 are employed for the output layer, the main network layers, and the input layer, respectively. To address the scarcity of Asian subjects in the UERC 2023 training dataset, the external EarVN1.0 dataset [18] is also included in the fine-tuning process. This dataset consists of ear images from 98 males and 66 females of Asian origin. The main part of the backbone network is utilized similar to AdaFace, while a linear classifier is employed for classification. The model was trained for 30 epochs with the presented parameters and data.

MEM-Ear. The multi-algorithm ensemble approach to ear recognition (MEM-Ear) combines diverse data representations extracted with the ConvNext-tiny, iResNet100, and EfficientNet-B3 networks for recognition purposes. MEM-Ear starts with the ear normalization procedure, proposed by Hansley *et al.* in [16] to minimize appearance variations due to pose variations and size. Next, three feature extractors are trained using the normalized UERC 2023 dataset, that is, ConvNext-tiny [23], EfficientNet-B3, and iResNet100 [2]. To account for bias, a weighted loss functions is adopted for each model that incorporates precomputed demographic proportions into the learning objective. For the evaluation procedure, the feature vectors extracted with each of the models are aggregated into the final ear representation using a weighted summation, where the weights are determined in accordance with the recognition performance on the training data.

UERC-IGD. The UERC-IGD solution consists of a ResNet-18 [17] model trained on the UERC 2023 dataset. The data is split into 90% for training and 10% for validation. The CosFace [33] learning objective is utilized to learn discriminative identity representations from the UERC 2023 dataset. The margin of CosFace is set to 0.2 and the scale factor to 8. Two additional classification layers are added to the base ResNet-18 model to encourage it to learn more descriptive (and non-biased) features in a multi-task setting. The first is optimized for gender classification and the second is optimized for learning ethnicity classification. The model is trained for 6 epochs with the SGD optimizer, a learning rate of 0.1 and the weight decay of 5e-4. During the training phase, training samples are augmented with RandAug [9], following the settings in [6].

RecogEAR. The RecogEAR approach uses a dedicated deep convolutional neural network, called, two-Stream Inflated 3D ConvNet (I3D) [7], where filters and pooling kernels of standard 2D ConvNets are expanded into 3D. Here, two of the dimensions correspond to spatial data, while the third dimension is temporal. To make the I3D architecture applicable for ear recognition, 3DRecogEAR first converts the given (single) input image into a video by creating at least 30 sequential patches of size 50×50. The patches are created by densely sampling from each given input image. The network itself consists of 3D convolutional layers, 3D batch normalization, 3D max-pooling and 3D average pooling operation. The first 3D convolutional layer takes a 4-dimensional image tensor as input (color channels $\times$ number of frames/patches $\times$ width $\times$ height). and extracts spatial and temporal features from the input image. Next, a 3D max-pooling layer is utilized to reduce the dimensionality of the generated feature representations and focus the processing on the most informative image features. This procedure is repeated across multiple layers of the model, where at each layer an inception structure with multi-scale processing capabilities is used. Finally, a global average pooling is used at the top of the model to aggregate the computed (spatial and temporal) features and reduce the generated data representation into a form that can be fed into a classification layer. To learn the model, an ArcFace loss [11] is minimized over the UERC 2023 training data.

ViTEar. The ViTEar approach utilizes the DINOv2 Vision Transformer [25] to extract discriminative data representations from the input images. The approach starts with a pretrained DINOv2 model and then fine-tunes the model on aligned data from the UERC 2023 and EarVN datasets [18] to improve its performance for the ear recognition task. Here, margin penalty softmax losses, including CosFace [33], ElasticCosFace, and ElasticCosFace+ [5], are used for the fine-tuning. To minimize the impact of pose variations, the two-stack hourglass network from [19] is employed for the image normalization before the learning procedure. To improve the model's generalization capabilities, various augmentation techniques, such as resiz-

ing, random horizontal flipping, affine transformations, histogram equalization, perspective transformations, color jittering, Gaussian blur, grayscale conversions, and normalization procedures are applied at run-time to the training images. To further boost performance, the embeddings of the different models (i.e., fine-tuned with different losses) are concatenated into the final data representation. This ensemble achieves an impressive rank 1 accuracy of 96.27% on the UERC 2019 test dataset. Overall, ViTEar effectively combines unsupervised pretraining, margin penalty softmax losses, alignment techniques, and ensemble modeling to achieve strong recognition performance.

5. Experiments and Results

In this section, we present the results of UERC 2023. We perform an in-depth analysis of the submitted approaches and study different aspects of ear recognition technology, with a focus on performance and demographic bias due to gender and ethnicity.

5.1. Performance vs. Bias (Track 1)

Overall Comparison. We first analyze the behavior of the submitted models with respect to recognition performance and demographic bias. To this end, all models are scored on the UERC 2023 test data and the summary results are reported in Table 2. Here, the Equal Error Rates (EER), the Area Under the ROC Curve (AUC), and the False Non-Match Rate at a 1% False Match Rate (F1F) are provided to quantify verification performance and the Gini index is reported to quantify bias due to gender and ethnicity. To gain further insight into the submitted models, the trade-off between verification performance, demographic bias and model size (in terms of #parameters) is illustrated in Figure 2. As can be seen, the MEM-Ear approach yields the strongest verification results both in terms of EER and AUC scores, followed closely by the ViTEar and DHCF solutions. ViTEar especially, leads to highly competitive verification performance at the 1% False Match Rate with a F1F score of 0.278. The next group of techniques, i.e., IGD, KU-Ear and PreWAdaEAR, yield slightly lower but still competitive results and convincingly outperform the UERC baseline. The weakest performance is exhibited by the RecogEAR approach, with an AUC score of around 0.5. It is interesting to note that the three top performers are conceptually quite distinct, with MEM-Ear focusing on CNN ensembles, ViTEar on transformers and DHCF on a combination of learned and handcrafted features.

If we look at the behavior of the models with respect to demographic bias, we can see that RecogEAR achieves the lowest Gini index of 0.019 followed in order by the DHCF, KU-Ear, PreWAadaEar and MEM-Ear approaches with scores around 0.1 and the IGD and ViTEar techniques

Table 2. **Performance comparison across different performance measures.** The table reports Equal Error Rates (EER), the Area Under the ROC Curve (AUC), the False Non-Match Rate at 1% False Match Rate – FNMR @ 1% FMR (F1F) and the Gini index computed over the EER (G). The results are sorted according to the EER. The symbol ↓ suggests that lower is better and with ↑ higher is better.

Submitted Model	EER ↓	AUC ↑	F1F ↓	G ↓
MEM-EAR	**0.146**	**0.915**	0.313	0.116
ViTEar	0.177	0.908	**0.278**	0.224
DHCF	0.185	0.895	0.355	0.092
IGD	0.190	0.868	0.483	0.195
KU-EAR	0.198	0.880	0.414	0.099
PreWAdaEAR	0.204	0.887	0.378	0.101
RecogEAR	0.493	0.494	0.999	**0.019**
UERC Baseline	0.360	0.699	0.908	0.053

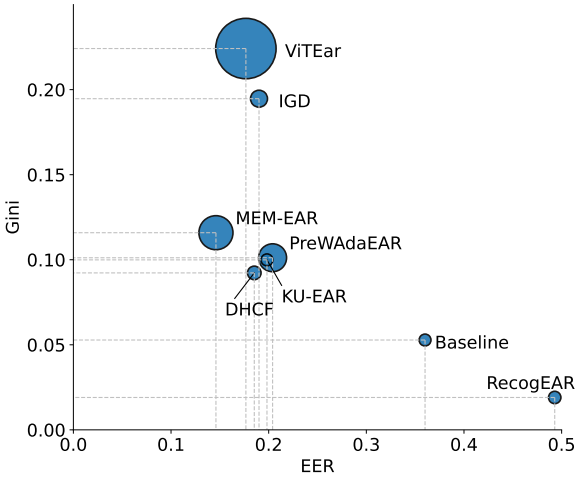
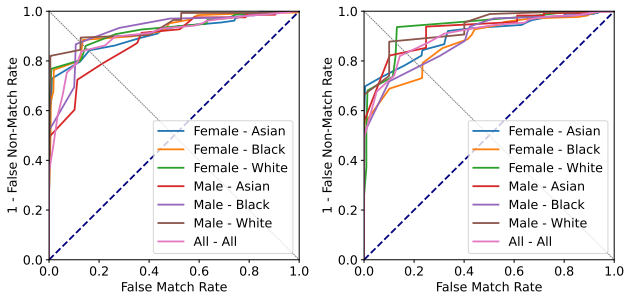


Figure 2. **Comparison of the verification performance (EER) vs. bias behavior (Gini).** The surface area of the circles represents the number of trainable parameters. For reference, the smallest model (Baseline) has around $11.7M$ parameters and the largest (ViTEar) has $304.9M$. An ideal method would be located at the origin of the coordinate systems.

that exhibit somewhat larger differential performance when it comes to gender and ethnicity groups, with Gini indices of 0.195 and 0.224, respectively. Interestingly, all but the RecogEAR technique, fared worse in terms of bias than the UERC baseline, suggesting that improved performance consistently leads to worse behavior in terms of bias. It is also worth noting that more heavily parameterized models (see ViTEar), despite resulting in stronger verification performance, do not necessarily help with bias.

Bias Analysis. Next, we focus explicitly on demographic bias and compare the submitted models across different groups of test images in terms of gender and ethnicity. Specifically, in Figure 3 we report the complete ROC curves for each demographic sub-group, whereas in Figure 4 we provide a comparison of the models at the Equal Error Rate. From the ROC curves, we can see that the models



(a) MEM-EAR (b) ViTEar

(c) DHCF (d) IGD

(e) KU-EAR (f) PreWAdaEAR

(g) UERC Baseline (h) RecogEAR

Figure 3. **Comparison of the submitted models across demographic sub-groups.** A separate ROC curve is provided for each demographic sub-group for all tested models, plotted in logarithmic scale. Observe how the differential performance changes at different operating points.

in general exhibit a considerable spread in performance for the different demographic subgroups. While these performance differentials were captured by the Gini index in Table 2 for the EER operating point, the ROC curves provide a more complete picture and suggest that among the better performing models MEM-Ear, DHCF, PreWAdaEar and KU-Ear provide the smallest performance differentials for a range of operating points, whereas, the IGD and ViTEar techniques, on the other hand, lead to consistent results across the demographic sub-groups, for some ROC operating points, but not for others. When looking at the EER results in Figure 4 it is interesting to note that the models behave differently with the different demographic subgroups. While, for instance, overall, the *black-male* sub-group leads to the strongest verification performance at the EER operating point on average, this is not universally true for all submitted recognition models. ViTEar, for example, performs the worst with this subgroup, which suggests that the images are encoded quite differently by the submitted models.

Table 3. **UERC 2023 Track 1 ranking.** The submitted models are ranked based on a weighted criterion that considers the EER score (verification performance) and Gini index (demographic bias).

Submitted Model	Track 1 Ranking (↓)
MEM-Ear	**0.140**
DHCF	0.167
KU-EAR	0.179
PreWAdaEar	0.183
ViTEAR	0.186
IGD	0.189
RecogEAR	0.398
UERC Baseline	0.299

Track 1 Ranking. In Table 3, we provide the overall UERC 2023 ranking for the first track of the challenge. The table reports a weighted score that jointly considers verification performance (i.e., EER) and demographic bias (i.e., Gini index) in accordance with Eq. (1) and where lower values correspond to a better ranking. As can be seen, MEM-Ear is the top performer of the first Track of UERC 2023 with a combined score of 0.140. The runner-up, DHCF, resulted in a join score of 0.167, while the rest of the models performed slightly worse, but again better than the UERC baseline. The only exception here is the RecogEAR technique that performed weaker than the baseline.

5.2. Model Improvement (Track 2)

The goal of the second track was to improve on the UERC baseline ResNet-18 model by maximizing performance, while minimizing demographic bias. Three teams entered this track of the competition, i.e., IGD, KU-EAR and DHCF, all of which modified the initial ResNet-18 model with various mechanisms. IGD added a multi-task loss, KU-EAR fine-tuned the model with a supervised contrastive loss, and DHCF combined the ResNet-18 features with handcrafted ones. These strategies led to differences in verification performance as well as bias compared to the baseline, as summarized in Table 4 and Figure 5. As can be

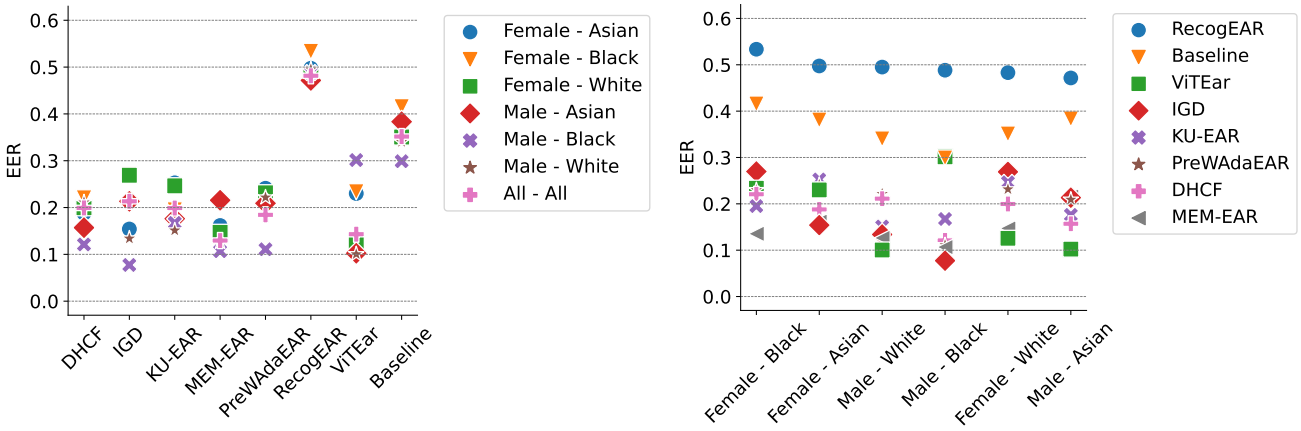


Figure 4. **Comparison of differential performance due to demographics at the EER operating point.** The left graph shows the performance differentials of the tested models w.r.t. different demographic sub-groups and the right graph shows the performance spread for each sub-group across all models. The figure is best viewed in color.

Table 4. **UERC 2023 Track 2 ranking.** The submitted models are ranked based on a weighted criterion that consider changes in verification performance and bias behavior compared to the UERC baseline. S_p denotes the relative performance loss (+) or increase (-), and S_g relative demographic bias decrease (-) or increase (+).

Submitted Model	S_p	S_g	**Track 2 Ranking** (↓)
KU-Ear	−0.45	0.89	−0.182
DHCF	−0.50	1.17	−0.166
IGD	−0.47	2.69	0.162

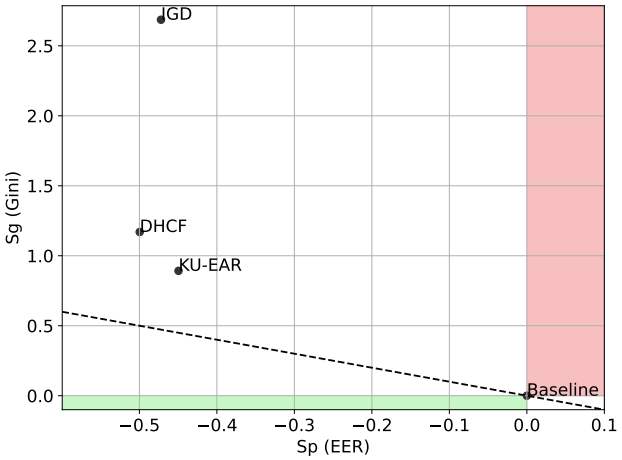


Figure 5. **Comparison of changes in verification performance Sp (EER) and demographic bias Sg compared to the baseline.** Ideally, the tested models would fall into the bottom left quadrant (marked with green), where both verification errors and bias are decreased. The worst scenario is within the top right quadrant (marked with red), where both the error and bias increase. All track 2 approaches are somewhere in between, i.e., they did not reduce bias, but they did improve verification performance.

seen, all submissions managed to improve on the verification performance (see negative S_p scores), with the hybrid DHCF approach improving the most. On the other hand, all of the submissions unfortunately also increased the demographic bias, as suggested by the positive S_g scores. Here, KU-Ear led to the lowest bias increase with a S_g score of 0.89, whereas the IGD approach, despite its explicit gender and ethnicity oriented multi-task learning objective, worsened the bias behavior the most with a S_g score of 2.69. The final ranking of the second UERC Track is based on a weighted criterion (i.e., $S_p + 0.2S_g$), where KU-Ear comes out as the top performer, followed by DHCF and IGD.

6. Conclusion

The aim of the third Unconstrained Ear Recognition Challenge (UERC 2023) was to evaluate the current state of technology in the field of ear recognition with respect to verification performance and bias caused by demographic factors, such as gender and ethnicity. The results of the challenge suggest that modern ear recognition techniques achieve encouraging verification results with images captured in unconstrained settings, with the top-performer yielding an Equal Error Rate of 0.146 on the considered test data. However, compared to other modalities, there is still a gap in the overall performance. Furthermore, the results have shown that both gender and ethnicity impact results to a considerable extent, but also that the bias toward better performance for certain groups differs from model to model even if the same training data is used for the learning process. Finally, we observed that improving performance, while also minimizing bias with a predefined model is challenging, as all submissions tackling the *model improvement* task improved the verification results, but also made them more biased.

References

[1] V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2021.

[2] X. An, J. Deng, J. Guo, Z. Feng, X. Zhu, Y. Jing, and L. Tongliang. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Z. Babnik and V. Štruc. Assessing bias in face image quality assessment. In *30th European Signal Processing Conference (EUSIPCO)*, pages 1037–1041, 2022.

[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*, 2023.

[5] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1587, 2022.

[6] F. Boutros, M. Klemt, M. Fang, A. Kuijper, and N. Damer. Unsupervised face recognition using unlabeled synthetic data. In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023*, pages 1–8, 2023.

[7] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[8] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020.

[9] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[12] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chellappa. Pass: protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15087–15096, 2021.

[13] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[14] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Demographic bias in presentation attack detection of iris recognition systems. In *28th European Signal Processing Conference (EUSIPCO)*, pages 835–839, 2021.

[15] S. Gong, X. Liu, and A. K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 330–347, 2020.

[16] E. E. Hansley, M. P. Segundo, and S. Sarkar. Employing fusion of learned and handcrafted features for unconstrained ear recognition. *IET Biometrics*, 7(3):215–223, Jan. 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[18] V. T. Hoang. Earvn1.0: A new large-scale ear images dataset in the wild. *Data in Brief*, 27:104630, 2019.

[19] A. Hrovatic, P. Peer, V. Struc, and Z. Emersic. Efficient ear alignment using a two-stack hourglass network. *IET biometrics*, 12(2):77–90, 2023.

[20] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS)*, 2015.

[21] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022.

[22] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.

[24] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021.

[25] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[26] S. Ramos-Cooper, E. Gomez-Nieto, and G. Camara-Chavez. Vggface-ear: an extended dataset for unconstrained ear recognition. *Sensors*, 22(5):1752, 2022.

[27] X. Soria, G. Pomboza-Junez, and A. D. Sappa. Ldc: Lightweight dense cnn for edge detection. *IEEE Access*, 10:68281–68290, 2022.

[28] M. Vitek, A. Das, D. R. Lucio, L. A. Zanlorensi, D. Menotti, J. N. Khiarak, M. A. Shahpar, M. Asgari-Chenaghlu,

F. Jaryani, J. E. Tapia, et al. Exploring bias in sclera segmentation models: A group evaluation approach. *IEEE Transactions on Information Forensics and Security*, 18:190–205, 2022.

[29] Ž. Emeršič, D. Sušanj, B. Meden, P. Peer, and V. Štruc. Contexednet : Context-aware ear detection in unconstrained settings. *IEEE Access*, pages 1–17, 2021.

[30] Ž. Emeršič, A. K. S. V., B. S. Harish, W. Gutfeter, J. N. Khiarak, A. Pacut, E. Hansley, M. P. Segundo, S. Sarkar, H. Park, G. P. Nam, I. J. Kim, S. Sangodkar, U. Kacar, M. Kirci, L. Yuan, J. Yuan, H. Zhao, F. Lu, J. Mao, X. Zhang, D. Yaman, F. I. Eyiokur, K. B. Ozler, H. K. Ekenel, D. P. Chowdhury, S. Bakshi, P. K. Sa, B. Majhni, P. Peer, and V. Štruc. The unconstrained ear recognition challenge 2019. In *Proceedings of the International Conference on Biometrics (ICB)*, 2019.

[31] Ž. Emeršič, D. Štepec, V. Štruc, P. Peer, A. George, A. Ahmad, E. Omar, T. E. Boult, R. Safdaii, Y. Zhou, others Stefanos Zafeiriou, D. Yaman, F. I. Eyoikur, and H. K. Ekenel. The unconstrained ear recognition challenge. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 715–724, 2017.

[32] Ž. Emeršič, V. Štruc, and P. Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017.

[33] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR))*, pages 5265–5274, 2018.