Machine Learning with Applications 19 (2025) 100610

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Demographic bias mitigation at test-time using uncertainty estimation and human–machine partnership☆

Anoop Krishnan Upendran Nair [a,c], Ajita Rattani [b,*]

[a] School Of Computing, Wichita State University, 1845 Fairmount St., Wichita, 67260, KS, USA
[b] Department of Computer Science and Engineering, University of North Texas Discovery Park, 3940 North Elm Street, 76207 Denton, TX, USA
[c] Edwin L. Cox School of Business, Southern Methodist University, 6214 Bishop Blvd, Dallas, 75275, TX, USA

ARTICLE INFO

Keywords:
Facial attribute classifier
Human–machine partnership
Demographic bias mitigation
Continual learning
Test-time methods

ABSTRACT

Facial attribute classification algorithms frequently manifest demographic biases by obtaining differential performance across gender and racial groups. Existing bias mitigation techniques are mostly in-processing techniques, i.e., implemented during the classifier's training stage, that often lack generalizability, require demographically annotated training sets, and exhibit a trade-off between fairness and classification accuracy. In this paper, we propose a technique to mitigate bias at the test time i.e., during the deployment stage, by harnessing prediction uncertainty and human–machine partnership. To this front, we propose to utilize those lowest percentages of test data samples identified as outliers with high prediction uncertainty. These identified uncertain samples at test-time are labeled by human analysts for decision rendering and for subsequently re-training the deep neural network in a continual learning framework. With minimal human involvement and through iterative refinement of the network with human guidance at test-time, we seek to enhance the accuracy as well as the fairness of the already deployed facial attribute classification algorithms. Extensive experiments are conducted on gender and smile attribute classification tasks using four publicly available datasets and with gender and race as the protected attributes. The obtained outcomes consistently demonstrate improved accuracy by up to 2% and 5% for the gender and smile attribute classification tasks, respectively, using our proposed approaches. Further, the demographic bias was significantly reduced, outperforming the State-of-the-Art (SOTA) bias mitigation and baseline techniques by up to 55% for both classification tasks. The demo shall be released on https://github.com/hashtaglensman/HumanintheLoop.

1. Introduction

Automated facial analysis (FA) encompasses diverse applications, ranging from face detection to attribute classification, such as gender and age prediction, and actual face recognition. These applications play a prominent role in contemporary smartphones, law enforcement, border control, and surveillance (Almadan, Krishnan, & Rattani, 2020; Kiruthika & Masilamani, 2021; Krishnan, Neas, & Rattani, 2022; Levi & Hassner, 2015; Masood, Gupta, Wajid, Gupta, & Ahmed, 2018; Nadimpalli & Rattani, 2022; Rattani, Derakhshani, & Ross, 2019; Salim, Sankaranarayanan, & Jayaraman, 2021; Siddiqui, Rattani, Ricanek, & Hill, 2022; Villa et al., 2020; Zhang, Gao et al., 2017). Commercial entities, including Amazon Rekognition (Rekognition, 2022), DeepVision AI (Vision, 2022), FaceX (FaceX, 2022), and Microsoft Azure Cognitive Services (Services, 2022), have released SDKs featuring automated FA.

Despite these advancements, recent research indicates pervasive demographic biases in facial analysis technology, particularly across demographic groups such as gender, race, and age groups (Abdurrahim, Samad, & Huddin, 2018; Albiero et al., 2020; Best-Rowden & Jain, 2018; Buolamwini & Gebru, 2018; Chouldechova, 2017; Grother, Quinn, & Phillips, 2011; Klare, Burge, Klontz, Bruegge, & Jain, 2012; Krishnan, Almadan, & Rattani, 2020a, 2020b, 2021; Muthukumar, 2019; Raji & Buolamwini, 2019; Vera-Rodríguez et al., 2019). Specifically, differential performance is obtained for women, dark-skinned people, and the elderly. Fairness is the absence of prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics. Thus, an unfair (biased) algorithm is one whose decisions are skewed towards a particular group of people. The presence of demographic bias in these systems has significant ramifications encompassing erroneous

☆ This document is the result of the research project partially funded by the National Science Foundation, United States.
* Corresponding author.
E-mail addresses: axupendrannair@shockers.wichita.edu (A.K.U. Nair), ajita.rattani@unt.edu (A. Rattani).
URLs: https://scholar.google.com/citations?user=z7GEbVwAAAAJ&hl=enn (A.K.U. Nair), https://scholar.google.com/citations?user=9esyU2EAAAAJ&hl=en (A. Rattani).

https://doi.org/10.1016/j.mlwa.2024.100610
Received 1 April 2024; Received in revised form 25 September 2024; Accepted 26 November 2024
Available online 6 December 2024

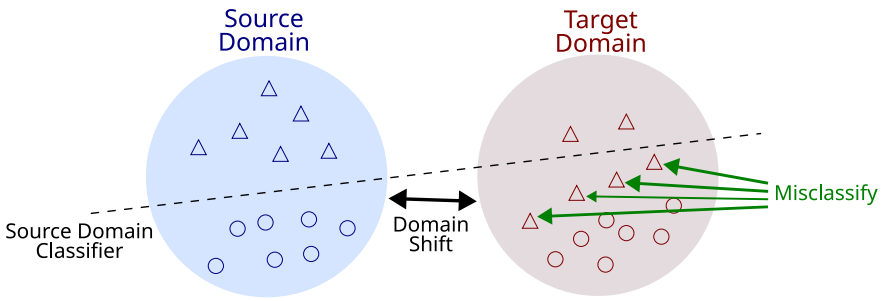

Fig. 1. Visualization of the domain shift challenge, where the model's decision boundary (represented by the dotted line) fails to accurately classify data points from the target domain (represented by red shapes) due to the divergence from the source domain (represented by blue shapes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

identification and other discriminatory outcomes such as rejection of applications from certain demographic groups in automated hiring (Huang, Zhang, Mao, & Yao, 2023). This bias can emanate from multiple origins, including training data annotation error during crowdsourcing (Klie, Webber, & Gurevych, 2023), biased (skewed) training data distribution (Krishnan et al., 2020b) and algorithmic bias (Palma, Kiani, & Lloyd, 2019) (Lingenfelter, Davis, & Hand, 2022). Bias in facial analysis technology can also perpetuate and reinforce existing societal biases and inequalities.

For addressing the challenges associated with bias in facial analysis technology such as facial attribute classification, a flurry of research includes the examination of the bias (Prabhu, Yap, Wang, & Whaley, 2019) in the dataset and algorithms and the development of bias mitigation strategies. Specifically, debiasing techniques based on the regularization strategies (Kamishima, Akaho, Asoh, & Sakuma, 2012; Krishnan & Rattani, 2023), attention mechanism (Majumdar, Singh, & Vatsa, 2021), adversarial debiasing (Chuang & Mroueh, 2021; Zhang, Lemoine, & Mitchell, 2018), over-sampling the minority class using Generative Adversarial Networks (GANs) (Ramaswamy, Kim, & Russakovsky, 2021), multi-task classification (Das, Dantcheva, & Brémond, 2018) and consistency regularization based technique (Krishnan & Rattani, 2023) have been proposed for the bias mitigation of facial attribute classifiers. Most of these aforementioned techniques are *in-processing techniques* i.e., fairness-related penalties are introduced during the training stage to obtain a fairer model. Thus, these mitigation strategies are often offline processes.

The limitations of most of these in-processing bias mitigation techniques include the need for a demographically annotated training set, poor generalizability, and high computational complexity (Ramaswamy et al., 2021; Zhang et al., 2018). Furthermore, the use of these mitigation strategies often introduces a trade-off between fairness and classification accuracy (Zhang et al., 2018). This is also called *Pare-to inefficiency* which implies that fairness is often obtained at the cost of reduced overall classification accuracy (Berk, Heidari, Jabbari, Kearns, & Roth, 2021; Chen, Zhang, Sarro, & Harman, 2022; Wick, Panda, & Tristan, 2019).

It is also crucial to emphasize the intricate challenges inherent in static deep neural network architectures. These challenges include inherent biases, overfitting, and progressive performance degradation primarily stemming from domain shift phenomena. Domain shift arises when the test data distribution diverges significantly from the training data distribution (Attenberg, Ipeirotis, & Provost, 2015; Reiter, 1977), as depicted in Fig. 1.

Additionally, Attenberg et al. (2015) argued that offline models might exhibit systematic misclassification errors for data representations not encompassed by the training set, termed "unknowns", as the trained model fails to generalize effectively to unseen data representations. Consequently, the model assigns high confidence scores to these erroneous predictions. Such data representations are termed "unknown unknowns" (UUs) since the classification model remains oblivious of such errors (Han, Dong, & Demartini, 2021). These "unknown unknowns" reside in the critical region, defined as the region of the data distribution where test instances are misclassified with high confidence by the model, constituting the predominant source of predictive errors and classification fallibilities. *We conjecture that the phenomenon of domain shift, coupled with the errors engendered by data representations extraneous to the training set, termed UUs exert a profound impact on the systematic bias exhibited by the model.*

Thus, effectively managing the aforementioned contributing factors and mitigating resulting demographic bias demands meticulous scrutiny and the development of robust strategies that accommodate the dynamic nature of the environment at the test time. Moreover, in practical scenarios, it may not be feasible to retrain or fine-tune the deployed model (Lohia et al., 2019; Wang et al., 2022). However, limited research has been dedicated to bias mitigation for already deployed models at the test time or during the deployment stage (Kong, Yuan, Hao, & Henao, 2023; Marcinkevics, Ozkan, & Vogt, 2022).

This paper **aims** to propose strategies for demographic bias mitigation of facial attribute classifiers at test-time using uncertainty estimation and human–machine partnership, using labeling and continual learning framework as illustrated in Fig. 2. Notably, among facial image-based visual attributes such as gender, ethnicity, and age, gender stands out as an important demographic attribute. The automated gender classification (Albiero, Zhang, King, & Bowyer, 2022; Krishnan et al., 2020a; Tapia, Perez, & Bowyer, 2016) holds significant relevance in various applications, including image retrieval, surveillance, and human–computer interaction. Furthermore, smile attribute classification (Becker, Kenrick, Neuberg, Blackwell, & Smith, 2007; Steephen, Mehta, & Surampudi, 2017) is another dimension gaining importance in facial analysis applications, contributing to emotion recognition and enhancing user interaction (Bostan & Klinger, 2018; Demszky et al., 2020). In this context, our study **addresses** the bias of facial image-based gender and smile attribute classification tasks, with gender and race as the protected attributes, at the test time as a case study.

1.1. Our contribution

In summary, the main *contributions* of this work are as follows:

- We explored five label-agnostic methods for uncertainty estimation (quantification) of the samples at test-time for two different facial attribute classification tasks i.e., facial image-based gender- and smile classification.
- We evaluated the efficacy of decision rendering and continual learning by annotating the test samples classified with high uncertainty via a human–machine partnership in mitigating bias of the face attribute classifier at the test time.
- Extensive evaluation of the proposed approaches in bias mitigation at test-time has been explored in the intra- and cross-data evaluation for gender classification & smile attribute classification. We utilized the FairFace (Kärkkäinen & Joo, 2021) for

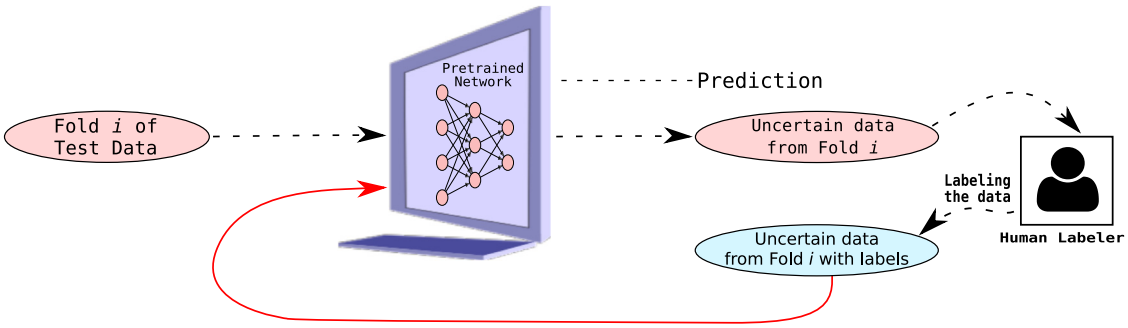

Fig. 2. Illustration of our proposed approach that uses human–machine partnership to mitigate bias of facial attribute inference/analysis at the test time. At each time stamp, any test data samples classified with uncertainty as estimated by an uncertainty quantification technique will be referred to a human analyst for the label assignment, and decision rendering, and for iteratively fine-tuning the classifier by incorporating periodic labeling of outliers by the human analyst.

training the gender classifier and the gender-balanced version of CelebA (Liu, Luo, Wang, & Tang, 2015) for smile attribute classifier.

Later, the gender classifier was evaluated across different folds of FairFace, UTKFace (Zhang, Zhifei, Song, Yang and Qi, 2017), and DiveFace (Morales, Fiérrez, Vera-Rodríguez, & Tolosana, 2021) datasets. Similarly, the smile attribute classifier was evaluated on the LFW (Huang, Ramesh, Berg, & Learned-Miller, 2007) dataset. Evaluations were conducted across different gender-racial demographic groups such as Asian Males, Asian Females, White Males, White Females, Indian Males and Indian Females, African Males, and African Females.

- Cross-comparison with the state-of-the-art (SOTA) bias mitigation techniques on facial attribute classifier was performed.
- Lastly, we performed a cause-and-effect analysis by using GradCAM and combining GradCAM with Guided Backpropagation (Gildenblat & contributors, 2021; Selvaraju et al., 2017). This allowed us to gain a comprehensive understanding of the enhanced details learned by the adaptive model through the proposed human–machine partnership in the continual learning framework.

This paper is organized as follows: relevant related work on examining, and mitigating bias on face-based gender and smile attribute classifiers, and prior work on human–machine partnership is discussed in Section 2. Section 3 discusses the methods for uncertain data estimation. Section 4 discusses the datasets used, implementation details, and the evaluation metrics. Furthermore, the results and key findings are discussed in Sections 5 and 6, respectively. Finally, the concluding remarks and future work are discussed in Section 7.

2. Related work

In this section, we will discuss the related work on investigating and mitigating bias in facial attribute (gender & smile attribute) classification algorithms, and the related work on human–machine partnership.

On Examining & Mitigating Bias in Gender Classification: The following foundational work has identified the systematic failings of gender classification algorithms across gender and race (Barlas, Kyriakou, Guest, Kleanthous, & Otterbacher, 2020; Buolamwini & Gebru, 2018; Joo & Kärkkäinen, 2020; Krishnan et al., 2020b; Li & Xu, 2021; Muthukumar, 2019).

Specifically, Buolamwini and Gebru (2018) evaluated the fairness of five COTS gender classifiers and suggested unequal accuracy for dark-skinned people and women on the Pilot Parliaments Benchmark (PPB) dataset. Muthukumar (2019) suggested that age, hair length, and facial hair likely cause the performance differential for women and dark-skinned people when evaluated on the PPB dataset. Krishnan et al. (2020b) evaluated the efficacy of different CNN architectures (ResNet-50, Inception-V4, VGG-16/19, and VGGFace) in gender classification across gender-racial groups when evaluated on the UTKFace and FairFace datasets, respectively. The authors suggested that architectural differences impact unequal accuracy rates. The authors in Joo and Kärkkäinen (2020) proposed an encoder–decoder network to synthesize facial images with varying gender and race attributes to measure counterfactual fairness in commercial computer vision classifiers. They also reported skewed gender representations in online search services, which may explain the biases in the models. Barlas et al. (2020) discussed the issue of differential performance of computer vision algorithms across gender and race. They found that the training data often has too many images of people and situations that exhibited social stereotypes contributing to biased performance. Li and Xu (2021) proposed a new framework for discovering unknown biased attributes of an image classifier without human conjecture. They introduced a novel total-variation loss and orthogonalization penalty within this framework to optimize a hyperplane in a generative model's latent space, representing the biased attribute. This approach aimed to assist in the automatic discovery of biases that may not be apparent to humans, reducing the need for extensive human effort in annotating test images for bias analysis.

Multiple approaches have been proposed to mitigate the bias of gender classification algorithms (Chiu, Chung, Chen, Shi, & Ho, 2023; Das et al., 2018; Georgopoulos, Oldfield, Nicolaou, Panagakis, & Pantic, 2021; Krishnan & Rattani, 2023; Majumdar et al., 2021; Park, Hwang, Kim, & Byun, 2021; Ramachandran & Rattani, 2023; Zhang et al., 2018). Ramachandran and Rattani (2023) introduced an approach leveraging generative views, structured learning, and evidential learning to improve fairness as well as classification accuracy of gender classifiers. Das et al. (2018) proposed a Multi-Task Convolution Neural Network (MTCNN) to jointly classify gender, age, and ethnicity, as well as to minimize the impact of protected attributes. The proposed model was evaluated on UTKFace and BEFA datasets. Georgopoulos et al. (2021) presented a style-based neural data augmentation framework that enhances demographic diversity using a novel style transfer method. To obtain the joint demographic attribute style, the authors introduced a tensor-based mixing structure that captures multiplicative interactions between attributes in a multilinear fashion to mitigate demographic bias and improve fairness metrics. Park et al. (2021) introduced a Fairness-aware Disentangling Variational Auto-Encoder (FD-VAE) to combat bias by disentangling the influence of protected attributes while preserving features pertinent to the main classification task. Majumdar et al. (2021) proposed an Attention Aware Debiasing (AAD) method utilizing an attention mechanism to learn unbiased feature representations pertinent to the main classification task. Finally, Krishnan and Rattani (2023) proposed a bias mitigation technique based on consistency-based regularization utilizing image-level and feature-level augmentation to alleviate bias of the gender classifier. Zhang et al. (2018) proposed a method to reduce biases in machine learning models by using adversarial learning techniques. They introduced an adversarial debiasing framework that involves training a

predictor model and an adversary model together. The goal is to ensure fairness in the predictor model's predictions. They demonstrated the effectiveness of this approach on the UCI dataset.

On Examining & Mitigating Bias in Smile Attribute Classification: The literature review in this domain has extensively explored the inherent biases present in smile attribute classification algorithms. A noteworthy observation, as reported by studies (Becker et al., 2007; Steephen et al., 2017), indicates that existing algorithms exhibit systematic failings, particularly in classifying smiles across genders. Studies show that automated systems tend to judge women as happier than men. Such systems are also better at detecting angry expressions on men's faces and happy expressions on women's faces (Becker et al., 2007; Steephen et al., 2017).

In a related study, Denton, Hutchinson, Mitchell, and Gebru (2019) investigated the impact of altering facial features for smile classification. Their findings reveal that a smiling classifier trained on CelebA tends to predict "smiling" faces more frequently when certain alterations, such as beard removal or application of makeup and lipstick, are introduced while keeping other aspects unchanged. This underscores the presence of psychological biases, leading us to hypothesize the existence of systematic annotation bias in large, in-the-wild expression datasets. We posit that this, coupled with data representation bias, significantly contributes to the observed gender bias in trained models.

Furthermore, Wang et al. (2020) conducted an extensive study on demographic bias mitigation for the smile attribute classifier. Their study encompassed data balancing, fairness through blindness, and fairness through awareness. Their results demonstrated that fairness through awareness obtained superior outcomes in mitigating bias, especially in the context of smile attribute classification using the CelebA dataset.

Human–Machine Partnership and its Benefit: This review delves into the realm of human–machine collaboration across various domains (Correia & Lécué, 2019; Han et al., 2021; Russakovsky, Li, & Fei-Fei, 2015; Yao, Gall, Leistner, & Gool, 2012). Correia and Lécué (2019) introduced a reinforcement learning-based human-in-the-loop framework, enhancing machine-learning classification tasks by selecting pertinent features based on expert feedback. Russakovsky et al. (2015) proposed a collaborative framework for object annotation, incorporating human–machine collaboration. This system aims to efficiently and accurately localize objects in images by considering annotation constraints such as precision, utility, and human cost. Yao et al. (2012) presented an incremental learning approach for refining models for object detection. Additionally, Han et al. (2021) proposed an iterative strategy leveraging human intelligence to identify and retrain models on unknown unknowns, augmenting prediction accuracy and facilitating effective classification confidence evaluation. These advancements underscore the significance of human–machine partnerships in optimizing model performance and addressing evolving challenges.

The ongoing discourse on the human–machine partnership paradigm underscores the benefits inherent in integrating this approach within a deep learning framework. Noteworthy advantages encompass enhanced model accuracy, bolstered robustness, elevated user trust, and the facilitation of iterative learning. The infusion of human presence into the loop contributes to the transparency of the system, rendering it comprehensible and interpretable for human operators. This symbiotic collaboration between human agents and artificial intelligence (AI) establishes a shared responsibility, thereby elucidating the decision-making process. Additionally, the human–machine partnership paradigm adeptly assimilates human judgment, aligning AI systems with human preferences and expertise. This method reduces the burden of perfecting algorithms, prioritizing collaboration and continuous improvement. The iterative process, guided by human intelligence, enables adaptive responses to dynamic environments and outlier samples, ultimately leading to enhanced system performance (Amershi, Cakmak, Knox, & Kulesza, 2014; Wang, 2019).

3. Methods for uncertainty quantification

Uncertainty quantification pertains to the ability of a pre-trained machine learning model to furnish probabilistic estimates regarding its predictive confidence or uncertainty (Sensoy, Kaplan, & Kandemir, 2018). In contrast to exclusively generating deterministic predictions, the model additionally provides a metric indicating the level of certainty associated with each prediction. The incorporation of uncertainty quantification holds critical significance in the development of AI systems with robust assurance, ensuring a comprehensive understanding of the model's predictive reliability in various scenarios. This nuanced approach to prediction enhances the model's applicability in engineering contexts, where informed decision-making relies on a nuanced comprehension of the model's confidence in its predictions (Amini, Schwarting, Soleimany, & Rus, 2020; Malmström, Skog, Axehill, & Gustafsson, 2022).

Next, we will discuss some of the uncertainty estimation techniques employed in this study for the detection of test samples classified with high prediction uncertainty.

- **Boundary Proximity Confidence-based Outlier Detection (BPCOD):** Boundary Proximity Confidence-based Outlier Detection (BPCOD) (Monarch & Manning, 2021) is based on the underlying assumption that the samples located in proximity to the decision boundary are more likely to be misclassified. To quantify this proximity, we measure the standard deviation (dev) of the distance unit between the feature embeddings of the test samples and the averaged feature embeddings of the training data of each class given as follows:

$$dev = \sigma\Big(d\left(f(x), \Sigma f\left(c_1\right)\right), d\left(f(x), \Sigma f\left(c_2\right)\right), \ldots\Big) \tag{1}$$

where $f(x)$ is the feature embedding of the input test sample, $\Sigma f\left(c_n\right)$ is the averaged feature embedding of the training data belonging to the class c_n, and d is the distance between the two feature embeddings, and σ is the standard deviation of all the distances.

Additionally, we compute the ratio of confidence, i.e. the ratio between the two most confident predictions as shown in Eq. (2), which reflects the level of certainty in the predictions. Mathematically, for an input x with c possible classes, the confidence ratio (CR) is obtained as

$$CR = \frac{p\tilde{c}}{p\tilde{c}2} \tag{2}$$

where, pc is the model prediction probability for class c, $\tilde{c}$ is the class with maximum prediction probability ($\tilde{c}$ = argmaxc pc), $\tilde{c}2$ is the class with second-highest prediction probability.

By combining these proximity and confidence measures as shown in Eq. (3), the BPCOD method effectively identifies outliers that are likely to be misclassified, thereby enhancing the robustness of the classification process. Thus, the test data sample with a standard deviation close to zero and a confidence ratio above a threshold will be identified as an outlier. Therefore, we define sample x as uncertain if:

$$uncertain(x) = \left\{ \begin{matrix} True,\ if\ \ CR \geq \tau_1\ \&\ dev \leq \tau_2 \\ False,\ otherwise \end{matrix} \right\} \tag{3}$$

where τ_1 and τ_2 are two predetermined thresholds.

- **Ensemble-based Outlier Detection (EBOD):** EBOD (Ouyang, Song, Li, Sant, & Bauchy, 2021) strategically harnesses the knowledge and perspectives of multiple expert models generated using pruning and quantization techniques (Kuzmin, Nagel, van Baalen, Behboodi, & Blankevoort, 2023), which also enhance their execution efficiency on resource-constrained devices. The fundamental principle behind incorporating both the pruned & quantized and the original model into an ensemble lies in capitalizing on the diverse knowledge and perspectives

offered by multiple models while concurrently mitigating computational complexity and memory requirements. EBOD employs a comparative approach, pitting the predictions of the primary expert classifier against the fused output vector. This comparison, as depicted in Eq. (4), enables the identification of uncertain data points or outliers.

For an input x belonging to a set of c possible classes, let f represent the primary expert classifier, f_i denote the ith expert model, and $(f_i)''$ be its pruned and quantized counterpart, respectively. We define the uncertain input condition for x as follows:

$$uncertain\,(x) = \left\{ \begin{matrix} True,\ if\ argmaxc\big(f\,(x)\big)\ \neq argmaxc\left(\sum_{i=1}^{n}(f_i)''\,(x)\right) \\ False,\ otherwise \end{matrix} \right\} \quad (4)$$

where $argmaxc\ f\,(x)$ represents the class with the maximum predicted probability according to the primary expert classifier f evaluated on input x. The uncertain input condition is satisfied if the class yielding the maximum prediction probability from the primary expert classifier f differs from the class corresponding to the maximum of the summed prediction probabilities across the pruned and quantized versions $(f_i)''$ of the ensemble of expert models f_i, evaluated on the input x.

Thus, this synergistic collaboration culminates in an enhanced outlier detection performance in a resource-constraint environment, demonstrating the effectiveness of EBOD in addressing the challenges caused by limited computing resources.

- **Dirichlet Uncertainty Estimation (DUE):** This Dirichlet Uncertainty Estimation (DUE) technique, as presented in Sensoy et al. (2018), employs the Dirichlet distribution to quantify prediction uncertainty within deep neural networks. This method treats model predictions as subjective opinions by learning a function that combines evidence from data through a deterministic neural network. The resulting predictor for multi-class classification takes the form of another Dirichlet distribution, where the continuous output of the neural network determines the parameters. The Dirichlet distribution is a multivariate generalization of the beta distribution and is commonly used as a prior distribution in Bayesian statistics for categorical data. It is a continuous probability distribution defined on the simplex of K-dimensional vectors whose components sum to 1. The Dirichlet distribution is parameterized by a vector of positive real numbers $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_K$, where each $\alpha_i > 0$ is called a concentration parameter. The probability density function (PDF) of the Dirichlet distribution with parameters α is given by:

$$f(x_1, x_2, \ldots, x_K \mid \alpha_1, \alpha_2, \ldots, \alpha_K) = \Gamma\left(\sum \alpha_i\right)/(\Pi\ \Gamma(\alpha_i)) \times \Pi\ x_i^{\alpha_i - 1}\} \quad (5)$$

where: $x_i \geq 0$ and $\sum x_i = 1$, $\Gamma(x)$ is the gamma function, The mean and variance of the i_{th} component of the Dirichlet distribution are:

$$E[X_i] = \alpha_i / \sum \alpha_j \quad (6)$$

$$Var[X_i] = (\alpha_i(\sum \alpha_j - \alpha_i))/((\sum \alpha_j)^2 \times (1 + \sum \alpha_j)) \quad (7)$$

In contrast to standard softmax neural networks, Dirichlet-based uncertainty model predict the parameters of a Dirichlet distribution, the natural prior for categorical distributions, based on input $x(i)$ (i.e., $q(i) = Dir(\alpha_i)$, where $f(x_i|\theta) = \alpha_i \in R_C^+$). Consequently, the epistemic distribution $q(i)$ encapsulates uncertainty on $x(i)$, signifying uncertainty on the categorical distribution prediction $p(i)$. During training, the model learns the Dirichlet distribution parameters (α), enabling the modeling of prediction distributions for each class. In the evaluation phase, when confronted with new data samples, the model calculates predicted probabilities for each class.

Mathematically, for input sample i with c possible classes, let $f\left(x_i|\theta\right)$ is the output of the network of sample i, where θ is the network parameters, and

$$\alpha_i = f(x_i|\theta) \quad (8)$$

for $i = 1, \ldots, K$, where K is the number of classes, and the Dirichlet strength S is given as,

$$S = \sum_{i=1}^{K} \alpha_i \quad (9)$$

and, the uncertainty u is given as,

$$u = \frac{K}{S} \quad (10)$$

Therefore, we define x as uncertain if:

$$uncertain\,(x) = \left\{ \begin{matrix} True,\ if\ u \geq \tau \\ False,\ otherwise \end{matrix} \right\} \quad (11)$$

where τ is the predetermined threshold.

- **Entropy-based Uncertainty Estimation (EUE):** The Entropy-based Uncertainty Estimation (EUE) (Shannon, 1948) has utilized Shannon's entropy to measure the uncertainty of data samples within a deep neural-network framework. By examining the output probability distribution of the model for each sample, the entropy of the distribution can be calculated. This involves employing the probabilities assigned to each class by the model as an input to Shannon's entropy formula. The entropy is determined by taking the negative sum of the probability of each class multiplied by the logarithm of that probability. By computing the entropy for each sample, a threshold can be established to identify uncertain samples. Those samples whose entropy surpasses the threshold are considered uncertain. Thus, the EUE method leverages Shannon's entropy as a measure of uncertainty to identify and handle uncertain data samples in the deep neural network framework. Mathematically, for an input x with c possible classes and prediction probability pc for each class, the entropy is:

$$H\,(x) = -\sum c \times pc \times \log pc \quad (12)$$

where pc is the predicted probability for class c and log is the logarithm (typically base 2 or e). Higher entropy indicates higher uncertainty in the predictions. Thus, we define x as uncertain if:

$$uncertain\,(x) = \left\{ \begin{matrix} True,\ if\ H\,(x) \geq \tau \\ False,\ otherwise \end{matrix} \right\} \quad (13)$$

where τ is a predetermined entropy threshold.

- **Multiview Disagreement (MD):** Multiview Disagreement (MD) (Monarch & Manning, 2021) constitutes an alternative technique employed to estimate uncertainty in deep neural networks (DNN). This approach for outlier identification entails an examination of the prediction of multiple views of the same test sample. These views are generated by applying diverse transformations, such as rotations or translations, to the input data.

Consequently, each view obtains a distinct prediction for a given input sample. If there exists a disagreement between the predictions across different views, then the prediction is deemed uncertain. Thus, the MD method leverages the concept of multi-view disagreement to provide a measure of uncertainty within DNN models, thereby facilitating a more comprehensive understanding of the model's confidence in its prediction.

Mathematically, for an input x with c possible classes, with x_1 and x_2 as the two views of the same input sample, then we define the prediction on x as uncertain if:

$$uncertain\,(x) = \left\{ \begin{matrix} True,\ if\ f\left(x_1\right) \neq f\left(x_2\right) \\ False,\ otherwise \end{matrix} \right\} \quad (14)$$

where $f\left(x_1\right)$ and $f\left(x_2\right)$ are the predictions on x_1 and x_2, respectively, obtained by the model f.

Table 1
Overview of the classification tasks studied and the datasets used.

Task	Training dataset	Test dataset
Gender classification	FairFace (Kärkkäinen & Joo, 2021)	FairFace, UTKFace (Zhang, Zhifei et al., 2017) DiveFace (Morales et al., 2021)
Smile attribute classification	CelebA (Liu et al., 2015)	LFW (Huang et al., 2007)

Fig. 3. Sample face images from FairFace training set.

4. Experimental setup

In this section, we will discuss the datasets used and the details of model training.

4.1. Dataset

As illustrated in Table 1, we employed the gender- and race-balanced FairFace facial attribute dataset for training the gender classifier, and a gender-balanced subset of the large-scale face attribute dataset, CelebFaces Attributes Dataset (CelebA), for training the smiling attribute classifier. The trained gender classifier was subsequently evaluated on the holdout subsets of the FairFace, UTKFace, and DiveFace datasets. The trained smiling attribute classifier was evaluated on the holdout subset of the Labeled Faces in the Wild (LFW) dataset. We did not evaluate the smiling attribute classifier on the CelebA test set across gender and race attributes, as these annotations are not available for CelebA. The images in all the used datasets exhibit variations in age, gender, pose, illumination conditions, and facial expressions. A detailed discussion of these datasets is provided as follows:

FairFace: The Fairface dataset (Kärkkäinen & Joo, 2021) consists of 108,501 images, with an emphasis on balanced race composition in the dataset. The dataset is labeled with the seven-race groups, namely White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino Hispanic across male and female and age groups ranging from 0–9, 10–19, 20–29, 30–39, 40–49 and 50+. The training portion of the FairFace dataset consists of 47% females and 53% males. Table 2 showed the training distribution and Fig. 3 showed few training samples used in the work. For the gender classification task in this study, we used the training partition of the dataset for training the models, and the test partition for evaluating their performance.

UTKFace: The UTKFace dataset (Zhang, Zhifei et al., 2017) is a facial image dataset with a long age span (ranging from 0 to 116 years old). It contains over 20,000 face images annotated with age, gender, and ethnicity, namely White, Black, Asian, Indian, and Others (which include Hispanic, Latino, and Middle Eastern) with significant variations across pose, expression, illumination, occlusion, and resolution. Due to the vagueness of the "Other" category, we excluded it from this study. We used 25% of the entire dataset with an equal number of female and male images across different races as our test set.

DiveFace: The DiveFace dataset (Morales et al., 2021) is a facial image dataset and contains a total of 139,677 images. It contains gender and race annotations equally distributed to three ethnic groups (namely East Asian, Sub-Saharan and South Indian, and Caucasian). We used 25% of the entire dataset with an equal number of female and male images across races as our test set.

Fig. 4. Sample face images from CelebA training set.

CelebA: The CelebFaces Attributes Dataset (CelebA) (Liu et al., 2015) is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter. CelebA has a large demographic diversity, large sample size, and rich annotation. We used a gender-balanced subset of the CelebA dataset to train the smile attribute classifier. Table 3 showed the training set distribution and Fig. 4 showed a few training samples used in the work for smile classification.

LFW: The Labeled Faces in the Wild (LFW) (Huang et al., 2007) dataset is a widely used benchmark dataset for face analysis tasks in computer vision research. The original LFW dataset contains over 13,000 labeled images of faces collected from the internet, representing a diverse range of individuals, poses, and lighting conditions. The images are mostly unconstrained, captured in real-world settings, and include variations in facial expressions, illumination, and pose. We used the deep funneled version of the LFW dataset (Huang, Mattar, Lee, & Learned-Miller, 2012) which is a preprocessed version of the original LFW dataset as the test set. The deep funneled version applied a series of geometric and photometric corrections to the original images, including aligning the faces based on facial landmarks and normalizing the illumination conditions. This preprocessing step aims to reduce variations in pose, expression, and lighting. We used a holdout subset with an equal proportion of smiling and non-smiling faces with almost equal proportions across genders and races.

4.2. Implementation details

We implemented deep-learning models for two facial-attribute classification tasks i.e., gender classification, and smile attribute classification discussed as follows.

4.2.1. Model training

ResNet18 (He, Zhang, Ren, & Sun, 2016) architecture and the Vision Transformer, i.e.,ViT-B/32 (Dosovitskiy et al., 2021) architecture, using an input size of 224 × 224, a patch size of 32 × 32, and pre-trained on ImageNet, were employed for experimentation and validation of the proposed mitigation techniques on two classification tasks. To this **Baseline** architectures based on the ResNet18/ViT backbone, a penultimate layer with 1000 nodes is added for capturing image sample feature embeddings, followed by the output binary classification layer as shown in Fig. 5.

This model was trained on FairFace for gender classification and CelebA for smile attribute classification, utilizing binary cross-entropy loss as defined in Eq. (15). The presented findings primarily focused on utilizing the ResNet18 architecture for gender and smile attribute classification tasks. Similar trends and behaviors were observed for the ViT-B/32 architecture, and for the sake of brevity, these results have been included in the supplementary materials (refer Tables 27–44). These models were also trained to optimize the classification loss which is a cross-entropy loss. To mitigate the challenge of catastrophic forgetting (French, 1999; McCloskey & Cohen, 1989; Ratcliff, 1990) during continuous model retraining, we have incorporated a prior preservation loss into the process. The prior preservation loss constitutes the classification loss derived from a random subset of the original training data, as illustrated in Fig. 6 as l_{prior}.

$$l_{BCE} = -(\, y\log(p) + (1-y)\log(1-p)\,) \tag{15}$$

Table 2
Gender attribute classifier: FairFace training dataset distribution used in our study.

	White	Black	East Asian	Indian	Middle Eastern	Latino Hispanic	Southeast Asian	Total
Female	7826 (9%)	6137 (7%)	6141 (7%)	5909 (7%)	2847 (3%)	6715 (8%)	5183 (6%)	40 758 (47%)
Male	8701 (10%)	6096 (7%)	6146 (7%)	6410 (7%)	6369 (8%)	6652 (8%)	5612 (7%)	45 986 (53%)
Total	16 527 (19%)	12 233 (14%)	12 287 (14%)	12 319 (14%)	9216 (11%)	13 367 (16%)	10 795 (13%)	86 744 (100%)

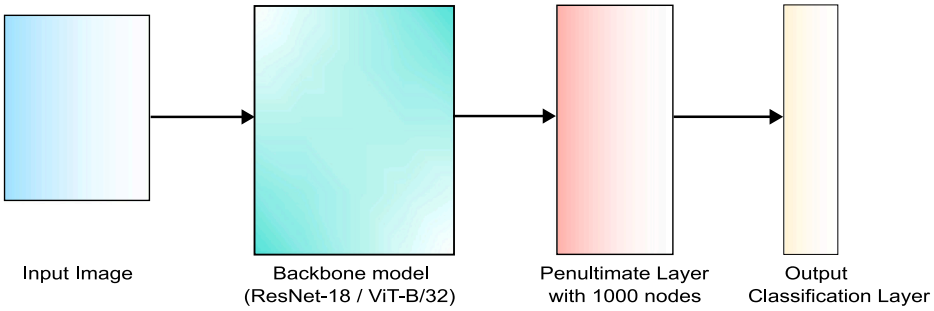

Fig. 5. Baseline architecture used in this work.

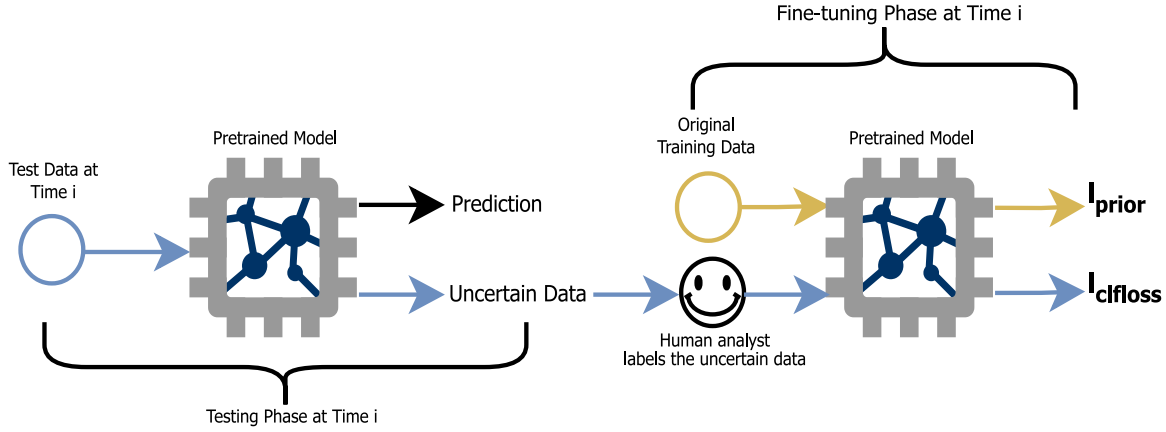

Fig. 6. Continual learning paradigm with human–machine partnership.

Table 3
Smile attribute classifier: CelebA training dataset distribution used in our study.

	Female	Male	Total
No smile	27 259 (25%)	27 259 (25%)	54 518 (50%)
Smile	27 259 (25%)	27 259 (25%)	54 518 (50%)
Total	54 518 (50%)	54 518 (50%)	109 036 (100%)

This classification loss from the few original data samples aids in preserving the initial patterns learned by the model. Integrating a prior preservation loss during continual learning has multiple advantages for deep neural networks, including safeguarding pre-acquired knowledge from pre-training or reference datasets, ensuring the retention of crucial features and patterns, and mitigating catastrophic forgetting. Additional benefits noted in previous research (Pan & Yang, 2010; Yosinski, Clune, Bengio, & Lipson, 2014) include the facilitation of knowledge transfer, regularization effects, improved robustness to variations and noise, and efficient fine-tuning convergence by leveraging the preserved knowledge as a strong initialization point, thereby reducing the reliance on an extensive dataset for fine-tuning.

Consequently, the continual learning paradigm optimizes the objective function delineated in Eq. (16), wherein the total loss is quantified as a composite of the classification loss on uncertain data instances and the loss associated with preserving the prior knowledge:

$$Total\,Loss = l_{prior} + l_{clfloss} \tag{16}$$

Here, l_{prior} denotes the loss related to preserving prior information, and $l_{clfloss}$ represents the classification loss incurred from the outliers labeled by the human annotators. Fig. 6 outlines the schema, wherein the test data at timestamp i undergoes evaluation using a pre-trained model. For instance, at i equals 1, the pre-trained model depicted in Fig. 6 serves as the baseline. Subsequently, human analysts label the outliers, which are then assimilated into the fine-tuning process along with the subset of original training data for 10 epochs. This integration results in an updated parameter configuration for the model. The resultant updated model would be the new predictive model for the subsequent iteration as shown in Tables 8–13, 16, 17.

All the models were trained with an empirically chosen batch size of 128 across 2 NVIDIA RTX 8000 GPUs, and label smoothing of 0.1. The training was performed using an RMSprop optimizer with cosine annealing and a warm restart with an initial learning rate of 3×10^{-4} and weight decay of 1×10^{-5}. We also utilized stochastic weight averaging along with mixed precision and an early stopping mechanism. All the experiments were done using the PyTorch-lightning framework.

4.2.2. Outlier detection

We comprehensively evaluated outlier detection methods on data folds available at different time stamps, focusing on gender and smile attribute classification tasks using FairFace, UTKFace, DiveFace, and LFW test datasets. For gender classification, Jenson Shannon's Divergence distance metric was empirically identified as effective for BPCOD, while cosine distance proved effective for the smile attribute classifier.

We determined specific threshold values, setting τ_1 to 0.11 and τ_2 to 2.33 in Eq. (3) based on empirical evidence.

For the implementation of EBOD, gender, and smile attribute classifiers were trained utilizing DenseNet121, Vision Transformer, and EfficientNet-v2 architectures. Global structured filter pruning (Frankle & Carbin, 2019) was employed to optimize model efficiency, involving the removal of the lowest-ranked weights or neurons based on the lowest L1 norm of the filter weights, followed by fine-tuning the remaining connections. Furthermore, local pruning was applied to further optimize the model, where the L1 norm of the weights was calculated to rank individual weights and identify those with the least norm, as their removal would have the least impact on model performance. Subsequently, weights of the pruned model were quantized from their original 32-bit precision to 8 bits, reducing memory requirements and computational cost. To mitigate the potential loss of precision from quantization, quantized models were calibrated by fine-tuning the model parameters with a subset of training data to ensure optimal performance.

For MD, we considered two input sample views: one without augmentation and the second with color jitter augmentation. Empirically determined thresholds for DUE and EUE in Eqs. (11) and (13) were set to 0.9 and 0.1, respectively. This systematic evaluation and parameter selection was done to enhance the robustness and effectiveness of outlier detection across different datasets and classification tasks, contributing to the broader understanding and applicability of these methods.

4.2.3. Evaluation procedure

In our evaluation, the test datasets were randomly partitioned into four equitably sized folds, wherein each fold corresponded to a specific timestamp. We conducted a comprehensive analysis of performance spanning across these folds stratified across demographic attributes. The outliers were diligently identified within each data fold available at a specific timestamp using the methods outlined in Sections 5.1.1 and 5.2.1. Subsequently, we executed two sets of evaluations:

- **Expert Labeling**: The initial set involved evaluating the performance within a collaborative human–machine partnership. In this experiment, the system first identifies the outliers, and then the human analyst assigns labels to those outliers. The outlier detection methods are selected in such a way that it has a minimum trade-off between human involvement and the performance of the system. Finally, the decision is made through collaboration between the machine and the human. To obtain the classification machine, the model was trained only with classification loss. The results of this experiment are tabulated in Tables 5, 6, 7, and 15.
- **Continual Learning**: The second set of experiments involved leveraging these identified outliers to fine-tune the model iteratively. Considering j as a timestamp, as presented in Tables 8–13, and 19–26, the outliers from data Fold j were employed along with a subset of the original training data for fine-tuning the model. The fine-tuned model at timestamp j is denoted as Finetuned Model j in the respective tables.
  The fine-tuning procedure followed the single-cycle approach illustrated in Fig. 6. The unfrozen model was fine-tuned for 10 epochs, utilizing the set of hyperparameters and optimizers as discussed in Section 4.2.1. The fine-tuned model at timestamp j was subsequently evaluated on the subsequent fold $j+1$. Although our evaluation concluded after the third timestamp, the underlying framework implies the perpetuation of this iterative process, facilitating adaptability and progressive enhancement of the model over time.

Essentially, our evaluation involves assessing and comparing the performance of classification and bias mitigation across a static expert labeling framework (via expert labeling) and a dynamic continual learning framework.

4.3. Metrics

To conduct a comprehensive assessment of all models' performance and quantify bias following prior research (Lin, Kim, & Joo, 2022; Singh, Majumdar, Mittal, & Vatsa, 2022), the following standard evaluation criteria were employed (Krishnan & Rattani, 2023): *overall classification accuracy, Degree of Bias (DoB) represented by the standard deviation of accuracy across specified demographics, and the ratio of maximum and minimum accuracy values explained as follows:*

1. DoB (standard deviation of accuracy across demographics) assesses variability of model performance among sub-groups. Low DoB indicates consistent classification accuracy across demographics, hence, indicating reduced bias.
2. Ratio of maximum and minimum accuracy values indicates disparities across demographic subgroups. The ratio of unity signifies consistent performance across diverse demographic subgroups, indicating fairness and unbiased treatment.

In this study, particular emphasis was placed on DoB and the max–min accuracy ratio, as they are crucial measures for assessing fairness and demographic parity.

Evaluating overall classification accuracy in the expert learning paradigm is computed as follows:

$$Overall = \frac{(x \times Acc_{human}) + (y \times Acc_{machine})}{x + y} \tag{17}$$

Here, x denotes the number of test samples evaluated by a human annotator, and y denotes the number of test samples assessed by the machine classifier. Furthermore, Acc_{human} represents the true classification accuracy obtained by the human annotator for the demographic subgroup, and $Acc_{machine}$ denotes the true classification accuracy obtained by the automated classifier.

5. Results

In this section, we will discuss the results of the experiments conducted on gender and smile attribute classification tasks in terms of the best method(s) for outlier detection, the impact of labeling by the human analyst, and the deployment of a continual learning framework that updates the classifier on the test samples classified with uncertainty. Throughout these evaluations, we consider the influence of gender and race as protected attributes for performance assessment. This comprehensive evaluation enables us to analyze the classifiers' efficacy and robustness across these socio-demographic factors, facilitating a more nuanced understanding of their performance characteristics.

5.1. Gender classification

5.1.1. Outlier detection

Table 4 tabulates the percentage of outliers identified by various outlier detection methods discussed in Section 3, including BPCOD, EBOD, MD, EUE, and DUE on gender classification tasks. The percentages in Table 4 reflect the proportions of test samples identified as outliers across the entire test dataset (FairFace, UTKFace, & DiveFace). BPCOD obtains the lowest outlier percentage at 4%, closely followed by EBOD at 5.7%. These two methods obtained the least human intervention among other methods, as compared with the remaining methods, namely MD, EUE, and DUE, obtaining higher outlier percentages ranging from 6% to 15.32%.

BPCOD's effectiveness also stems from its ability to identify outliers by jointly considering two critical factors: the proximity of a test sample to the decision boundary and the confidence level of the model's prediction. On the other hand, EBOD leverages an ensemble of multiple expert models, encompassing pruned and quantized versions resulting in fewer outliers detection.

Table 4
Overall outlier detection performance on gender classification test datasets (FairFace, UTKFace, & DiveFace).

Method	BPCOD	EBOD	MD	EUE	DUE
% outliers	4%	5.7%	6%	15.28%	15.32%

In line with our goal of establishing a human–machine partnership with a minimal trade-off between human involvement and performance, our further investigation focused on the two techniques BPCOD and EBOD, as they obtained the least number of outlier cases across the datasets.

5.1.2. Expert labeling

In this section, we presented the performance analysis of the baseline gender classifier and the human–machine partnership via expert labeling. The system involved human analysts annotating the labels of outlier data, thus leveraging the collective performance of both humans and machines. Our analysis assumed that human analysts are trained experts and can correctly label the outlier instances.

Tables 5, 6, and 7 tabulates the performance analysis of the baseline gender classifier and the collaborative human–machine partnership where human analyst labels the outlier samples detected by BPCOD and EBOD for decision rendering. The classifiers were trained on the FairFace dataset and evaluated on the FairFace, UTKFace, and DiveFace test sets. By combining the efforts of humans and machines, it was observed that the overall classification accuracy improved for all folds or timestamps, and the ratio of maximum–minimum accuracy and the Degree of Bias (DoB) across the sub-groups were reduced.

FairFace: Upon analyzing Table 5, the Baseline classifier was examined on the FairFace test set across four folds, with notable results. For Fold 1, the Degree of Bias (DoB) was measured at 3.108, while the ratio of maximum–minimum accuracy stood at 1.134. Fold 2 exhibited a DoB of 3.018 and a ratio of 1.14, followed by Fold 3 with a DoB of 3.86 and a ratio of 1.94. Lastly, Fold 4 obtained a DoB of 2.67 and a ratio of 1.12. The overall classification accuracies for these data folds were respectively determined as 91.54%, 93.446%, 92.58%, and 92.58%. Additionally, we observed there is a notable gap between the maximum accuracy (97.92% for Middle Eastern Males on Fold 3) and the minimum accuracy (80.2% for Black Females on Fold 3). This large gap of around 17.7% indicates a high degree of bias in the baseline model's performance across different demographic groups.

Upon employing the human–machine partnership context with expert labeling with BPCOD as the outlier detection technique, significant performance enhancement was observed. The overall classification accuracy for each fold improved to 93.473%, 94.92%, 94.66%, and 94.53%, respectively, representing an increment of up to 2%. Furthermore, the ratio of maximum to minimum accuracy decreased to 1.126, 1.106, 1.1011, and 1.078 across the subsequent folds, while the DoB metric reduced to 2.821, 2.481, 2.53, and 1.695, indicating a reduction of up to 35%. Furthermore, the gap between the maximum and minimum has reduced to around 9% as the maximum accuracy is 98.936% for Indian females on Fold 3, and the minimum is 89.071% for Black females on Fold 3.

Further improvements were realized when EBOD was employed as the outlier detection method. The overall classification accuracies on each fold increased to 95.752%, 96.617%, 96.483%, and 97.058%, surpassing the baseline accuracies by an increment of up to 5% as it detects more samples as outliers. Concurrently, the ratio of maximum to minimum accuracy decreased to 1.08, 1.071, 1.095, and 1.029 when compared to the baseline ratio. Similarly, the bias metric decreased to 1.894, 1.644, 2.223, and 0.828, resulting in a reduction of up to 69%. Furthermore, the gap between the maximum and minimum has reduced to around 9% as the maximum accuracy is 99.507% for Latino Hispanic females on Fold 2, and the minimum is 90.863% for Black females on Fold 3.

Hence, it is evident that the human–machine partnership context with expert labeling on the FairFace testset increased the overall classification accuracy by up to 5%, and reduced the bias metric by up to 69%.

UTKFace: From Table 6, we scrutinized the performance of the Baseline classifier on the UTKFace test set across four distinct folds, revealing noteworthy outcomes. For the baseline classifier, for Fold 1, the Degree of Bias (DoB) was determined to be 2.607, while the ratio of maximum–minimum accuracy stood at 1.092. Moving to Fold 2, we observed a DoB of 3.181 and a ratio of 1.122, followed by Fold 3 with a DoB of 4.482 and a ratio of 1.205. Lastly, Fold 4 displayed a DoB of 3.58 and a ratio of 1.137. The overall classification accuracies for these folds were respectively measured at 93.078%, 93%, 92.14%, and 92.79%. Additionally, we observed there is a notable gap between the maximum accuracy (96.697% for Indian Males on Fold 2) and the minimum accuracy (81.435% for Asian Females on Fold 3). This large gap of around 15% indicates a high degree of bias in the baseline model's performance across different demographic groups.

In the context of the human–machine partnership with expert labeling, with BPCOD employed as the outlier detection method, remarkable improvements were witnessed. The overall accuracy for each fold increased to 94.51%, 94.73%, 94.43%, and 94.37% respectively, with an overall increment up to 2%. Additionally, the ratio of maximum–minimum accuracy decreased to 1.571, 2.072, 2.785, and 2.573 across the subsequent folds, while the degree of bias decreased to 1.05, 1.071, 1.114, and 1.096, indicating a reduction in the degree of bias by up to 38%. Furthermore, the gap between the maximum and minimum has reduced to around 10% as the maximum accuracy is 98.656% for Black Males in Fold 4, and the minimum is 88.4% for Asian females in Fold 3.

Further, enhancements were achieved by utilizing EBOD as the outlier detection method. Concurrently, the ratio of maximum–minimum accuracy decreased to 1.06, 1.068, 1.115, and 1.097 when compared to the Baseline ratio. Similarly, the degree of bias decreased to 1.69, 1.89, 2.87, and 2.66, representing a reduction of up to 40%. Furthermore, the gap between the maximum and minimum has reduced to around 10% as the maximum accuracy is 98.795% for Black Males in Fold 4, and the minimum is 88.608% for Asian females in Fold 3.

Hence, it is evident that the human–machine partnership context with expert labeling on the UTKFace testset increased the overall classification accuracy by up to 3%, and reduced the bias metric by up to 40%.

DiveFace: Examining Table 7, we analyzed the performance of the Baseline classifier on the DiveFace test set across four distinct folds, revealing notable results. In Fold 1, the Degree of Bias (DoB) was determined to be 0.79, while the ratio of maximum–minimum accuracy stood at 1.027. Transitioning to Fold 2, we observed a DoB of 1.51 and a ratio of 1.047, followed by Fold 3 with a DoB of 1.079 and a ratio of 1.035. Lastly, Fold 4 displayed a DoB of 0.74 and a ratio of 1.018. The overall classification accuracies for these folds were respectively measured at 97.6%, 97.118%, 97.53%, and 97.6%. Additionally, we observed there is a gap between the maximum accuracy (99.46% for White Males on Fold 2) and the minimum accuracy (94.97% for Sub-Saharan & South Indian Females on Fold 2). This large gap of around 4% indicates a moderate degree of bias in the baseline model's performance across different demographic groups.

In the human–machine partnership context with expert labeling, employing BPCOD as the outlier detection method led to significant improvements. The overall accuracy for each fold increased to 98.38%, 98.07%, 97.98%, and 98.23%, respectively with an overall increment of up to 1%. Additionally, the ratio of maximum–minimum accuracy decreased to 1.017, 1.029, 1.022, and 1.022 across the subsequent folds, while the degree of bias decreased to 0.553, 0.863, 0.67, and 0.732 indicating a reduction in the degree of bias by up to 43%. Furthermore, the gap between the maximum and minimum has reduced

Table 5
Gender classification accuracy (%) on FairFace testset across different folds and gender-racial groups using expert labeling framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Baseline																	
Fold 1	85.643	85.96	91.453	91.1	94.54	93.846	89.44	92.54	97.087	94.624	91.379	91.011	93.662	89.431	1.134	91.54	3.108
Fold 2	89.64	84.7	93.01	93.264	94.5	94.38	94.86	96.55	96.208	94.9	93.407	94.631	93.214	94.98	1.14	93.45	3.018
Fold 3	90.164	80.2	92.353	94.413	95.19	94.15	93.9	94.089	95.098	95.495	90.206	95.745	93.725	91.428	1.194	92.58	3.86
Fold 4	93.229	87.437	94.118	93.33	93.989	88.442	95.161	92.453	97.92	90.426	90.27	91.52	94.719	93.133	1.12	92.58	2.67
Expert labeling-BPCOD																	
Fold 1	86.63	89.89	91.88	94.241	95.15	95.9	91.11	94.81	97.573	96.774	92.529	94.382	95.07	92.683	1.126	93.47	2.82
Fold 2	90.54	89.071	93.548	95.855	96	96.629	95.327	98.522	97.156	95.918	93.407	95.973	94.286	96.653	1.106	94.92	2.48
Fold 3	91.803	89.848	92.353	95.531	96.635	98.936	94.836	96.059	96.078	97.297	91.753	97.34	93.725	93.061	1.101	94.66	2.53
Fold 4	94.271	92.462	94.652	94.286	95.082	92.462	96.237	95.283	97.917	95.74	90.81	95.15	95.05	94	1.078	94.53	1.695
Expert labeling-EBOD																	
Fold 1	91.584	94.944	94.017	95.811	96.97	97.95	93.33	96.226	97.573	98.925	94.828	96.067	97.183	95.122	1.08	95.75	1.894
Fold 2	93.243	92.9	97.311	97.41	97	96.067	97.196	99.507	97.63	96.94	96.154	96.644	97.143	97.49	1.071	96.62	1.644
Fold 3	95.082	90.863	95.882	97.765	98.558	99.468	95.305	97.044	97.059	99.1	93.814	98.404	95.686	96.735	1.0947	96.48	2.223
Fold 4	96.875	98.492	96.79	95.714	96.72	98	96.774	97.64	98.437	96.81	95.676	96.97	97.36	96.57	1.029	97.06	0.828

to around 2% as the maximum accuracy is 99.637% for White Males in Fold 2, and the minimum is 97.283% for East Asian Males in Fold 4.

Further enhancements were achieved by utilizing EBOD as the outlier detection method. Concurrently, the ratio of maximum–minimum accuracy decreased to 1.01, 1.01, 1.02, and 1.02 when compared to the Baseline ratio. Similarly, the degree of bias decreased to 0.404, 0.41, 0.557, and 0.625, obtaining a reduction of up to 73%. Furthermore, the gap between the maximum and minimum has reduced to around 2% as the maximum accuracy is 100% for East Asian Females in multiple folds, and the minimum is 98% for Sub-Saharan & South Indian Males in Fold 4.

Hence, it is evident that the human–machine partnership context with expert labeling on the DiveFace test set slightly increased the overall classification accuracy by up to 1% as the test set was already performing well on the baseline, and reduced the bias metric by up to 40%.

In summary, these findings highlight the substantial impact of the human–machine partnership via expert labeling of the outliers as well as the effectiveness of different outlier detection methods with a minimum trade-off between human involvement and performance, in augmenting overall accuracy up to 5%, diminishing the ratio of maximum–minimum accuracy, and mitigating the degree of bias up to 8% and 70% respectively, for gender classification task across different folds of the FairFace, UTKFace, and DiveFace test sets.

5.1.3. Continual learning

In our continual learning evaluation of the gender classifier, we first used outlier detection methods to identify outliers, which were then labeled by a human analyst for classifier retraining (details in Section 4.2). We selected BPCOD and EBOD as our outlier detection methods for their optimal balance of human involvement and performance. The importance of prior preservation loss was also highlighted, ensuring initial features are maintained during learning.

Experimentally, we assumed each data fold as test samples available at different timestamps (see Section 4.2). We identified outliers from the latest timestamp and fine-tuned the model using these outliers and a portion of the original training set to maintain feature consistency (refer to Section 4.2). The fine-tuned model was then used to evaluate the next data fold.

BPCOD. **FairFace:** Table 8 shows the performance of the baseline and fine-tuned gender classifier when trained on FairFace and evaluated on the FairFace test set with BPCOD used as a means to detect the outliers. On fine-tuning with subsequent outliers for each fold at different timestamps, it was observed that Overall classification accuracy was improved and the ratio of max–min accuracy and Degree of Bias (DoB) were reduced mostly across all the folds and across the sub-groups. The Black Female and Male subgroups performed the least for all the baseline models.

The Baseline classifier was assessed on four-folds, where Fold 1 obtained a DoB of 3.108 and a max–min accuracy ratio of 1.134. Similarly, Fold 2 exhibited a DoB of 3.018 and a ratio of 1.14, Fold 3 had a DoB of 3.86 and a ratio of 1.94, and Fold 4 obtained a DoB of 2.67 and a ratio of max–min accuracy of 1.12. The overall classification accuracy for the folds was 91.54%, 93.446%, 92.58%, and 92.58%, respectively.

Subsequently, fine-tuning using outliers from Fold 1 (Finetuned Model 1), showed a slightly reduced overall accuracy of 92.91%, and an increment in bias as denoted by the increased DoB of 4.11 from 3.018, and the ratio of max–min accuracy of 1.184 from 1.14 on Fold 2. Further, Finetuned Model 2 and Finetuned Model 3 showed improved overall accuracy from 92.58% to 92.9% in Fold 3, and from 92.58% to 93.167% in Fold 4 respectively. Moreover, the intensity of bias is reduced, as the DoB reduced from 3.86 to 3.122 on Fold 3, and from 2.67 to 2.31 on Fold 4 on Finetuned Model 2 and 3 respectively. Similarly, the ratio of max–min accuracy reduced to 1.126 from 1.194 on Fold 3, and from 1.12 to 1.083 on Fold 4 using Finetuned Model 2 and 3 models, respectively. The gap between the maximum and minimum has reduced to around 9% (initially, it was 17.7% (maximum of 97.92% for Middle Eastern Males on Fold 3 and a minimum of 80.2% for Black Females on Fold 3)), with a maximum accuracy of 96.875% for Middle Eastern Males and a minimum of 89.447% on Black Females.

In summary, the human–machine partnership with the continual learning process with BPCOD as outlier detection led to a slight improvement in overall accuracy by up to 0.5%. This slight improvement could be because fewer iterations of continual learning have been considered for our evaluation. Additionally, there was a reduction in the degree of bias as evidenced by the decreased ratios of max–min accuracy and DoB by up to 31% and 21%, respectively, on the FairFace dataset.

UTKFace: As seen from Table 9, we observed a fair significant improvement in the overall classification accuracy and a reduction in the bias on the UTKFace dataset. Evaluation of the Baseline model on four folds of UTKFace obtained the lowest overall classification accuracy of 93.078% (Fold 1), 93% (Fold 2), 92.139% (Fold 3), and 92.786% (Fold 4), and the highest ratio of max–min accuracy of 1.092 (Fold 1), 1.122 (Fold 2), 1.205 (Fold 3), and 1.137 (Fold 4) and the highest DoB of 2.607 (Fold 1), 3.181 (Fold 2), 4.482 (Fold 3), and 3.58 (Fold 4).

On fine-tuning with the outliers of different Folds, we observed an improvement in the overall accuracy. On Finetuned Model 1, the

Table 6
Gender classification accuracy (%) on UTKFace testset across different folds, and gender-racial groups using expert labeling framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and races; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across genders and races.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Baseline											
Fold 1	91.036	88.387	96.078	91.058	96.516	93.12	94.815	93.615	1.092	93.078	2.607
Fold 2	91.198	86.433	97	92.16	96.697	93.056	94.95	92.427	1.122	93	3.181
Fold 3	91.686	81.435	98.107	92.13	93.45	92.732	94.968	92.602	1.205	92.139	4.482
Fold 4	89.922	86.797	98.656	90.796	95.841	91.05	95.326	93.896	1.137	92.786	3.58
Expert labeling-BPCOD											
Fold 1	92.437	92.043	96.242	93.796	96.7	94.5	95.48	94.91	1.05	94.512	1.571
Fold 2	92.176	91.028	97.508	94.425	97.064	94.907	95.527	95.164	1.071	94.725	2.072
Fold 3	92.637	88.4	98.451	94.242	94.867	96.49	95.758	94.604	1.114	94.431	2.785
Fold 4	92.248	90.043	98.656	93.274	96.187	92.841	96.068	95.652	1.0956	94.371	2.573
Expert labeling-EBOD											
Fold 1	95.518	92.473	97.386	95.985	98.083	94.5	96.593	97.153	1.061	95.961	1.691
Fold 2	97.06	91.904	97.674	96	98.165	94.676	97.258	96.077	1.068	96.108	1.893
Fold 3	97.625	88.608	98.795	96.16	95.22	95.739	96.837	96.432	1.115	95.677	2.869
Fold 4	96.382	90.26	99.04	96.106	98.094	93.29	97.033	97.157	1.097	95.92	2.656

Table 7
Gender classification accuracy (%) on DiveFace testset across different folds and gender-racial groups using expert labeling framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race.

Race	East Asian		Sub-Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Baseline									
Fold 1	97.51	98.84	97.642	96.237	98.074	97.287	1.027	97.6	0.79
Fold 2	97.94	96.953	97.82	94.97	99.46	95.56	1.047	97.118	1.51
Fold 3	98.03	99.33	97.02	96	98.053	96.739	1.035	97.529	1.079
Fold 4	97.1	98.66	96.92	97.3	98.61	97.015	1.0179	97.6	0.74
Expert labeling-BPCOD									
Fold 1	97.865	99.5	98.428	97.85	98.249	98.373	1.017	98.378	0.553
Fold 2	98.127	98.03	98.322	97.486	99.637	96.803	1.029	98.067	0.863
Fold 3	98.029	99.33	97.207	97.64	98.053	97.645	1.022	97.983	0.666
Fold 4	97.283	99.424	97.464	98.558	98.61	98	1.022	98.225	0.732
Expert labeling-EBOD									
Fold 1	99.11	100	99.017	98.925	98.95	99.638	1.01	99.273	0.404
Fold 2	99.064	99.82	99	99.46	99.82	98.756	1.01	99.319	0.41
Fold 3	99.104	100	98.138	99.273	99.47	99.094	1.019	99.18	0.557
Fold 4	98.913	100	98	99.64	99.306	99.17	1.02	99.173	0.625

Table 8
Gender classification accuracy (%) on FairFace testset for ResNet-18 across different folds, and gender-racial groups using continual learning framework with prior preservation loss. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from data Fold *i* identified using BPCOD.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Baseline																	
Fold 1	85.643	85.96	91.453	91.1	94.54	93.846	89.44	92.54	97.087	94.624	91.379	91.011	93.662	89.431	1.134	91.54	3.108
Fold 2	89.64	84.7	93.01	93.264	94.5	94.38	94.86	96.55	96.208	94.9	93.407	94.631	93.214	94.98	1.14	93.446	3.018
Fold 3	90.164	80.2	92.353	94.413	95.19	94.15	93.9	94.089	95.098	95.495	90.206	95.745	93.725	91.428	1.194	92.58	3.86
Fold 4	93.229	87.437	94.118	93.33	93.989	88.442	95.161	92.453	97.92	90.426	90.27	91.52	94.719	93.133	1.12	92.58	2.67
Finetuned Model 1																	
Fold 2	85.586	81.967	95.161	94.819	96.5	94.382	95.327	97.044	96.68	91.837	92.307	92.617	93.57	92.887	1.184	92.91	4.11
Finetuned Model 2																	
Fold 3	89.617	85.787	92.353	93.737	95.192	94.149	95.305	94.09	96.57	96.4	89.175	94.681	95.29	89.39	1.126	92.9	3.122
Finetuned Model 3																	
Fold 4	89.583	89.447	92.513	93.809	94	90.452	94.624	95.283	96.875	94.681	89.73	94.54	94.389	94.421	1.083	93.167	2.31

Table 9
Gender classification accuracy (%) on UTKFace testset for ResNet-18 across different folds, and gender-racial groups using a continual learning framework with prior preservation loss. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model i is the fine-tuned model with the outliers from Fold i identified using BPCOD.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Baseline											
Fold 1	91.036	88.387	96.078	91.058	96.516	93.12	94.815	93.615	1.092	93.078	2.607
Fold 2	91.198	86.433	97	92.16	96.697	93.056	94.95	92.427	1.122	93	3.181
Fold 3	91.686	81.435	98.107	92.13	93.45	92.732	94.968	92.602	1.205	92.139	4.482
Fold 4	89.922	86.797	98.656	90.796	95.841	91.05	95.326	93.896	1.137	92.786	3.58
Finetuned Model 1											
Fold 2	91.687	90.153	96.179	93.21	94.495	92.593	94.3	94.16	1.067	93.347	1.748
Finetuned Model 2											
Fold 3	89.55	89.24	96.56	93.67	92.39	96	94.25	93.56	1.082	93.15	2.504
Finetuned Model 3											
Fold 4	91.473	89.18	97.7	93.097	94.974	93.29	94.585	93.9	1.0955	93.523	2.348

overall accuracy of Fold 2 increased to 93.347% from 93%, similarly from 92.139% to 93.15% on Fold 3, and from 92.786% to 93.523% on Fold 4, respectively on Finetuned Model 2 and 3. Moreover, a significant reduction in bias was observed, and it was observed as the DoB was reduced from 3.181 to 1.748 on Fold 2 on Finetuned Model 1, followed by 2.504 from 4.482 on Fold 3 and 2.348 from 3.58 on Fold 4, respectively on Finetuned Model 2 and 3. Furthermore, the ratio of max–min accuracy was reduced to 1.067 from 1.122 on Fold 2 on Finetuned Model 1, followed by 1.082 from 1.205 on Fold 3, and 1.0955 from 1.137 on Fold 4, respectively on Finetuned Model 2 and 3. The gap between the maximum and minimum has reduced to around 8% (initially, it was around 15%), with a maximum accuracy of 97.7% for Black Males and a minimum of 89.18% for Black Females.

In summary, the human–machine partnership with the continual learning process with BPCOD as an outlier detection on UTKFace improved the classification accuracy by up to 1%. This slight improvement could be because fewer iterations of continual learning have been considered for our evaluation. Additionally, there was a reduction in the degree of bias as evidenced by the decreased ratios of max–min accuracy and DoB by up to 30% and 74% respectively on the UTKFace dataset.

DiveFace: Furthermore, on the DiveFace dataset with the gender classification task, from Table 10, the evaluation with the Baseline model, Fold 1, 2, 3 & 4 respectively obtained 97.6%, 97.118%, 97.529%, and 97.6% as the overall classification accuracy, and 1.027 & 0.79, 1.047 & 1.51, 1.035 & 1.079, and 1.0179 & 0.74 as the ratio of max–min & the DoB. Similarly, as we observed from Table 9, on fine-tuning with the outliers, it was observed that the overall classification accuracy was improved and the ratio of max–min accuracy and the DoB were reduced.

On fine-tuning with the outliers of Fold 1, Finetuned Model 1 obtained an increment in overall accuracy on Fold 2 from 97.118% to 98.4%. Similarly, Finetuned Model 2 and 3 obtained an increment in the overall accuracy from 97.529% to 98.287%, and from 97.6% to 98.249% respectively on Fold 3 and Fold 4. Moreover, we have observed a decline in bias, as the fine-tuning process progressed, the DoB reduced from 1.51 to 0.564 on Fold 2 using Finetuned Model 1, followed by 0.49 from 1.076 on Fold 3 and 0.663 from 0.74 on Fold 4 using Finetuned Model 2 and 3, respectively. Further, the ratio of max–min accuracy reduced to 1.0137 from 1.047 on Fold 2 using Finetuned Model 1, followed by 1.015 from 1.035 on Fold 3 using Finetuned Model 2, and no further improvement was found on Fold 4 using Finetuned Model 3. The gap between the maximum and minimum has reduced to less than 2% (initially, it was around 5%), with a maximum accuracy of 99.13% for White Males and a minimum of 97.84% for White Females.

In summary, the human–machine partnership with the continual learning process with BPCOD as outlier detection on DiveFace improved the classification accuracy by up to 1%. This slight improvement could be because of the higher performance of the dataset even on the baseline. Additionally, there was a reduction in the degree of bias as evidenced by the decreased DoB by up to 87%. Additionally, the framework performed better in the cross-dataset evaluation, obtaining a difference of up to 60% in bias reduction.

EBOD. **FairFace:** Table 11 showed the performance of the baseline and fine-tuned gender classifier when trained on FairFace and evaluated on the FairFace test set with EBOD used as a means to detect the outliers. On fine-tuning with subsequent outliers of each fold/timestamp using EBOD, it was observed that overall classification accuracy was improved and the ratio of max–min accuracy and Degree of Bias (DoB) were reduced across all the folds and the sub-groups. The Black Female and Male subgroups performed the least for all the baseline models.

As discussed before the baseline model obtained an overall classification of 91.54%, 93.446%, 92.58%, and 92.58% across four-folds, and the ratio of max–min accuracy and the DoB of 1.134 & 3.108, 1.14 & 3.018, 1.194 & 3.86, and 1.12 & 2.67 across Fold 1, Fold 2, Fold 3 and Fold 4 respectively. Fine-tuning with outliers of Fold 1 (Finetuned Model 1) did not improve the overall accuracy on Fold 2, but Finetuned Model 2 and 3 improved the overall accuracy on Fold 3 and Fold 4 from 92.58% to 92.95% and from 92.58% to 92.9%, respectively. Moreover, the DoB was reduced from 3.86 to 2.493 on Fold 3 using Finetuned Model 2, and on Fold 4, it was reduced from 2.67 to 2.02 using Finetuned Model 3. Subsequently, the ratio of max–min accuracy was also found to be reduced from 1.194 to 1.089 on Fold 3 using Finetuned Model 2, and from 1.12 to 1.095 on Fold 4 using Finetuned Model 3. The gap between the maximum and minimum has reduced to around 8% (initially, it was 17.7%), with a maximum accuracy of 96.875% for Middle Eastern Males and a minimum of 88.44% for Black Females.

In summary, the human–machine partnership with the continual learning process with EBOD as outlier detection on FairFace slightly improved the classification accuracy by around 0.5%. This slight improvement could be because fewer iterations of continual learning have been considered for our evaluation. Further, there was a reduction in the degree of bias as evidenced by the decrement in DoB by up to 60%.

On a similar task but on two different datasets, UTKFace and DiveFace, Tables 12 and 13 showed the increment in classification accuracy, and the decrement in the bias was observed on the cross-dataset evaluation.

UTKFace: From Table 12, as discussed before, the baseline model obtained an overall classification accuracy of 93.078%, 93%, 92.139%, and 92.786%, and the ratio of max–min and the degree of bias as 1.092 & 2.607, 1.122 & 3.181, 1.205 & 4.482 and 1.137 & 3.58, respectively, for Fold 1, 2, 3 and 4. On fine-tuning the model with the outliers of Fold

Table 10
Gender classification accuracy (%) on DiveFace testset for ResNet-18 across different folds and gender-racial groups using a continual learning framework with prior preservation loss. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model i is the fine-tuned model with the outliers from Fold i identified using BPCOD.

Race	East Asian		Sub-Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Baseline									
Fold 1	97.51	98.84	97.642	96.237	98.074	97.287	1.027	97.6	0.79
Fold 2	97.94	96.953	97.82	94.97	99.46	95.56	1.047	97.118	1.51
Fold 3	98.03	99.33	97.02	96	98.053	96.739	1.035	97.529	1.079
Fold 4	97.1	98.66	96.92	97.3	98.61	97.015	1.0179	97.6	0.74
Finetuned Model 1									
Fold 2	97.94	99.283	97.987	98.025	99.09	98.046	1.0137	98.4	0.564
Finetuned Model 2									
Fold 3	98.208	99.33	97.95	97.82	98.23	93.19	1.015	98.287	0.49
Finetuned Model 3									
Fold 4	98.19	98.848	97.1	98.378	99.13	97.84	1.021	98.249	0.663

Table 11
Gender classification accuracy (%) on FairFace testset for ResNet-18 across different folds, and gender-racial groups using a continual learning framework with prior preservation loss. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model i is the fine-tuned model with the outliers from Fold i identified using EBOD.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Baseline																	
Fold 1	85.643	85.96	91.453	91.1	94.54	93.846	89.44	92.54	97.087	94.624	91.379	91.011	93.662	89.431	1.134	91.54	3.108
Fold 2	89.64	84.7	93.01	93.264	94.5	94.38	94.86	96.55	96.208	94.9	93.407	94.631	93.214	94.98	1.14	93.446	3.018
Fold 3	90.164	80.2	92.353	94.413	95.19	94.15	93.9	94.089	95.098	95.495	90.206	95.745	93.725	91.428	1.194	92.58	3.86
Fold 4	93.229	87.437	94.118	93.33	93.989	88.442	95.161	92.453	97.92	90.426	90.27	91.52	94.719	93.133	1.12	92.58	2.67
Finetuned Model 1																	
Fold 2	88.74	84.153	91.4	92.746	97.5	94.382	90.65	95.57	95.734	94.9	92.308	91.275	92.143	93.724	1.159	92.516	3.227
Finetuned Model 2																	
Fold 3	87.98	89.34	92.94	92.18	95.67	93.617	95.775	94.09	95.098	95.5	89.69	94.68	94.118	90.61	1.089	92.95	2.493
Finetuned Model 3																	
Fold 4	90.625	88.44	92.513	92.38	92.35	91.96	95.16	94.34	96.875	92.553	91.35	94.54	94.389	93.133	1.095	92.9	2.02

1 (Finetuned Model 1), the overall accuracy was improved from 93% to 93.053%. Similarly, with Finetuned Model 2 & 3, the overall accuracy was improved to 93.29% from 92.14% on Fold 3, and to 93.66% from 92.786% on Fold 4 respectively. On considering the reduction in bias, the DoB reduced from 3.181 to 2.104 in Fold 2 with Finetuned Model 1, similarly on Fold 3, and Fold 4, the DoB reduced from 4.482 to 2.66 and from 3.58 to 1.562 with Finetuned Model 2 & 3 respectively. The gap between the maximum and minimum has reduced to around 5% (initially, it was 15%), with a maximum accuracy of 96.106% for Black females and a minimum of 91.473% for Asian males.

In summary, the human–machine partnership with the continual learning process with EBOD as outlier detection on UTKFace slightly improved the classification accuracy by around 2%. This slight improvement could be because fewer iterations of continual learning have been considered for our evaluation. Further, there was a reduction in the degree of bias as evidenced by the decrement in DoB by up to 85%.

DiveFace: From Table 13, as discussed prior, the baseline gender classification accuracy on Fold 1, Fold 2, Fold 3, and Fold 4 was obtained as 97.6%, 97.118%, 97.529% and 97.6% respectively. The baseline ratio of max–min accuracy and the degree of bias were obtained as 1.027 & 0.79, 1.047 & 1.51, 1.035 & 1.079, and 1.0179 & 0.74 for Fold 1, 2, 3, and 4 respectively.

On fine-tuning as observed in Tables 11 and 12 the overall classification accuracy was improved and the rate of disparity across race and gender was reduced. On fine-tuning with the outliers from Fold 1 (Finetuned Model 1), obtained improved overall classification accuracy on Fold 2 to 98.128% from 97.12%. Similarly on Fold 3, and Fold 4 with Finetuned Model 2 & 3 respectively, the overall classification accuracy increased from 97.53% to 98.287% and from 97.6% to 98.276% respectively. Moreover, a continuous reduction in bias was observed in terms of the DoB. On Fold 2, the DoB reduced from 1.51 to 0.477 with Finetuned Model 1, followed by 0.629 from 1.079 on Fold 3 with Finetuned Model 2 and from 0.74 to 0.668 on Fold 4 with Finetuned Model 3. In terms of the ratio of max–min accuracy, on Fold 2, it was reduced to 1.013 from 1.047 with Finetuned Model 1, and on Fold 3, from 1.035 to 1.018 with Finetuned Model 2. No further improvements were observed on Fold 4 with Finetuned Model 3. The gap between the maximum and minimum has reduced to less than 2% (initially, it was 5%), with a maximum accuracy of 99.46% for Sub-Saharan & South Indian females and a minimum of 97.464% for Sub-Saharan & South Indian males.

In summary, the human–machine partnership with the continual learning process with EBOD as outlier detection on DiveFace improved the classification accuracy by up to 1%. This slight improvement could be because fewer iterations of continual learning have been considered for our evaluation. Additionally, there was a reduction in the degree of bias as evidenced by the decreased DoB by up to 80%. Additionally, the framework performed better in the cross-dataset evaluation, achieving a difference of up to 20% in bias reduction.

Visualizing the model's decision-making process is crucial for gaining insights into its behavior and the impact of our proposed approach. Furthermore, Fig. 7 provides compelling insights into the changes observed in the Grad-CAM (Selvaraju et al., 2017) representations for gender classifiers. Grad-CAM is utilized to visualize the specific

Table 12
Gender classification accuracy (%) on UTKFace testset for ResNet-18 across different folds and gender-racial groups using a continual learning framework with prior preservation loss. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model i is the fine-tuned model with the outliers from data Fold i identified using EBOD.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Baseline											
Fold 1	91.036	88.387	96.078	91.058	96.516	93.12	94.815	93.615	1.092	93.078	2.607
Fold 2	91.198	86.433	97	92.16	96.697	93.056	94.95	92.427	1.122	93	3.181
Fold 3	91.686	81.435	98.107	92.13	93.45	92.732	94.968	92.602	1.205	92.139	4.482
Fold 4	89.922	86.797	98.656	90.796	95.841	91.05	95.326	93.896	1.137	92.786	3.58
Finetuned Model 1											
Fold 2	90.71	88.62	94.68	94.6	94.862	93.056	94.37	93.52	1.07	93.053	2.104
Finetuned Model 2											
Fold 3	91.45	87.34	95.7	96	94.96	93.484	94.896	92.776	1.1	93.288	2.66
Finetuned Model 3											
Fold 4	91.473	91.558	94.817	96.106	92.374	94.407	94.139	94.4	1.05	93.66	1.562

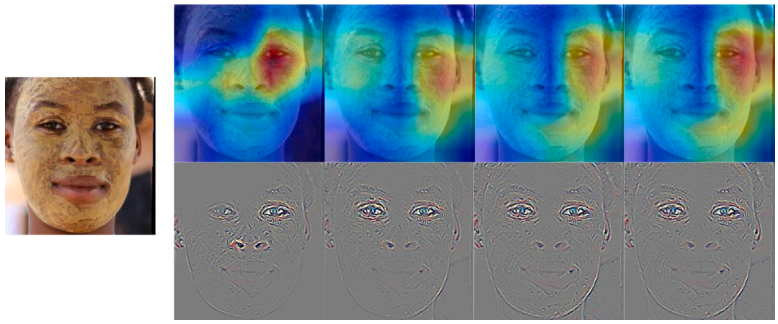

Fig. 7. Visualization of Grad-CAM (Top) and Combined Grad-CAM with Guided Back-propagation(Bottom) for Gender Classifiers: Baseline and Fine-tuned Models[1–3] (from left to right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

image regions utilized by the model for prediction. It generates a coarse localization map highlighting distinctive image regions crucial for decision-making by using gradients of a target concept. This method visually explains the decision-making process of deep neural networks, revealing which regions of an input image contribute most to the network's prediction. The map indicates highly activated regions through red, followed by green and blue zones.

Augmenting Grad-CAM with Guided Back-propagation enhances our result analysis by identifying specific features or patterns influencing the model's decision. By combining these techniques, a deeper understanding of the relationship between input features and predictions can be achieved, enabling more informed interpretations of the model's behavior. To this front, we generated the Grad-CAM visualization for both the baseline and iteratively fine-tuned gender classifiers.

Upon analyzing the Grad-CAM visualizations for the baseline model and the models obtained after each iteration of our proposed continual learning approach, which used prior preservation loss and EBOD for outlier detection, we noticed the baseline model primarily focused on the right eye area when making gender predictions. However, through the continual learning process, the model's attention expanded to a wider range of facial features, indicating an improved understanding of gender attributes beyond just the eye region.

Additionally, experimental results (See details at A.1) revealed a significant decline in accuracy up to 21% and increased discrimination up to 200% without prior preservation loss for gender classification, indicating catastrophic forgetting and underscoring its critical role in preserving knowledge and enhancing fairness in continual learning scenarios.

5.2. Smile attribute classification

5.2.1. Outlier detection

Table 14 presents the percentages of outliers identified by various outlier detection methods discussed in Section 5.2.1, including BPCOD, EBOD, MD, EUE, and DUE on smile attribute classification tasks. Among the listed methods, BPCOD obtained the lowest percentage of outliers at 0.5%, followed by EBOD at 1.136%. On the other hand, the remaining methods, namely MD, EUE, and DUE, exhibit increasing percentages ranging from 1.5% to 9.4%. Our comprehensive analysis is based on the BPCOD and the EBOD methods for outlier detection in congruence with our objective of a minimum trade-off between human involvement and performance.

5.2.2. Expert labeling

Table 15 presented the performance analysis of the baseline smile attribute classifier and the human–machine partnership via expert labeling. The classifier was trained on the CelebA dataset and evaluated on the LFW test set. BPCOD and EBOD methods were utilized for outlier detection. By combining the efforts of humans and machines, it was observed that the overall classification accuracy improved for all folds or timestamps. However, in most cases, the ratio of maximum–minimum accuracy and the Degree of Bias (DoB) did not decrease significantly across the sub-groups.

From Table 15, it was observed that on different non-overlapping folds of the LFW test set, the Baseline model obtained an overall accuracy of 83.48%, 83.318%, 83.739%, and 84% across Folds 1, 2, 3, & 4 respectively. When it comes to bias metrics, the ratio of max–min accuracy & the DoB obtained the highest values of 1.793 & 13.823 on Fold 1, 1.732 & 13.76 on Fold 2, 1.754 & 13.194 on Fold 3 and finally, 1.78 & 13.778 on Fold 4. Additionally, we observed there is a notable gap between the maximum accuracy (100% for Asian and Indian females on multiple folds) and the minimum accuracy (56.25%

Table 13
Gender classification accuracy (%) on DiveFace testset for ResNet-18 across different folds, and gender-racial groups using continual learning framework using prior preservation loss. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model i is the fine-tuned model with the outliers from Fold i identified using EBOD.

Race	East Asian		Sub-Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Baseline									
Fold 1	97.51	98.84	97.642	96.237	98.074	97.287	1.027	97.6	0.79
Fold 2	97.94	96.953	97.82	94.97	99.46	95.56	1.047	97.118	1.51
Fold 3	98.03	99.33	97.02	96	98.053	96.739	1.035	97.529	1.079
Fold 4	97.1	98.66	96.92	97.3	98.61	97.015	1.0179	97.6	0.74
Finetuned Model 1									
Fold 2	97.56	98.75	97.48	98.56	98.37	98.046	1.013	98.128	0.477
Finetuned Model 2									
Fold 3	98.208	99.33	97.58	97.82	97.876	98.913	1.018	98.287	0.629
Finetuned Model 3									
Fold 4	97.645	98.46	97.464	99.46	98.61	98	1.02	98.276	0.668

Table 14
Overall outlier detection performance on smile attribute classification test bench.

Method	BPCOD	EBOD	MD	EUE	DUE
% outliers	0.5	1.136	1.5	8.9	9.4

for Asian males on Fold 4). This large gap of around 54% indicates a high degree of bias in the baseline model's performance across different demographic groups.

When using BPCOD as the outlier detection method in the human–machine partnership with expert labeling, there were slight improvements. The overall accuracy slightly improved by 0.5%, and for each fold, it increased to 83.66%, 83.45%, 83.91%, and 84.052%, respectively. However, the ratio of maximum–minimum accuracy did not change across the subsequent folds, while the degree of bias slightly increased by 0.74%, with values of 14, 13.87, 13.29, and 13.82.

Better enhancements were obtained by utilizing EBOD as the outlier detection method. The overall accuracies improved by up to 4%, and for each fold, they increased to 85.78%, 85.5%, 87.38%, and 85.44%. At the same time, the ratio of maximum–minimum accuracy slightly decreased to 1.76, 1.702, 1.724, and 1.78 compared to the baseline ratio. Similarly, the degree of bias increased to 15.117, 14.97, 14.118, and 14.93.

Additionally, the expert labeling framework could not reduce the gap between the maximum and minimum accuracy compared to the baseline. This may be because there were too few uncertain samples to significantly improve the accuracy for the underperforming subgroups.

5.2.3. Continual learning

This section presents the evaluation of the smile attribute classifier in a continual learning setting (For details, refer to Section 4.2). Furthermore, this section showcases the significance of prior preservation loss in the continual learning framework.

BPCOD. On the smile attribute classifier, the human–machine partnership with continual learning, wherein the outliers were identified using BPCOD, reduced the extent of bias and improved the overall classification accuracy. From Table 16, it was observed that on different non-overlapping folds of the LFW test set, the Baseline model achieved an overall accuracy of 83.48%, 83.318%, 83.739%, and 84% across Fold 1, 2, 3, & 4, respectively. When it comes to bias metrics, the ratio of max–min accuracy & the Degree of Bias (DoB) attained the highest values of 1.793 & 13.823 on Fold 1, 1.732 & 13.76 on Fold 2, 1.754 & 13.194 on Fold 3, and finally, 1.78 & 13.778 on Fold 4.

On the human–machine partnership with continual learning framework, fine-tuning with the outliers identified using BPCOD, the Finetuned Model 1 model showed improved overall accuracy on Fold 2 from 83.32% to 85.19%. Followed by 83.75% from 83.74% on Fold 3 with Finetuned Model 2, and from 84% to 84.532% on Fold 4 with Finetuned Model 3. Moreover, a significant reduction in bias was observed. As in terms of DoB, on Fold 2, it was reduced to 10.82 from 13.76 with Finetuned Model 1, then 7.42 from 13.194 on Fold 3 with Finetuned Model 2, and finally from 13.778 to 6.675 on Fold 4 with Finetuned Model 3. Further, the ratio of max–min accuracy was found to be reduced, on Fold 2 with Finetuned Model 1, the ratio of max–min accuracy reduced from 1.732 to 1.571, then on Fold 3 with Finetuned Model 2, the ratio reduced from 1.754 to 1.523, and on Fold 4 with Finetuned Model 3, the ratio reduced from 1.78 to 1.4. Additionally. the gap between the maximum and minimum has reduced to around 27% (initially, it was 54% (maximum of 100% for Asian and Indian Females on multiple folds and a minimum of 56.25% for Asian males on Fold 4)), with a maximum accuracy of 93.548% for Black Females and a minimum of 66.67% on Indian females.

In summary, the human–machine partnership with the continual learning process with BPCOD as outlier detection on LFW led to a slight improvement in overall accuracy by up to 2%. This slight improvement could be because fewer outliers, as well as fewer iterations of continual learning, have been considered for our evaluation. Also, it is worth noting the difference in the distribution of training data, CelebA, and test distribution, LFW. Additionally, there was a reduction in the degree of bias as evidenced by the decreased DoB by up to 70%.

EBOD. For the same task of smile attribute classification, from Table 17, we could observe the baseline overall classification accuracy was obtained as 83.48%, 83.318%, 83.739%, and 84% after evaluating on Fold 1, Fold 2, Fold 3 and Fold 4 as discussed prior in Section 4.2. Also, the ratio of max–min accuracy and the degree of bias were obtained as 1.793 & 13.823, 1.732 & 13.76, 1.754 & 13.194, and 1.78 & 13.778 when evaluated on Fold 1, Fold 2, Fold 3, and Fold 4, respectively.

On the human–machine partnership with continual learning framework, fine-tuning the Baseline with the outliers from Fold 1 (Finetuned Model 1), we observed an increment in overall classification on Fold 2, it improved to 84.526% from 83.318%. Similarly, on Folds 3 and 4, the overall classification accuracy increased to 89% from 83.739%, and to 89.44% from 84% respectively with Finetuned Model 2, and 3. Moreover, the DoB of Fold 2 reduced to 8.861 from 13.76 with Finetuned Model 1, followed by 7.075 from 13.194 on Fold 3 with Finetuned Model 2, and finally 5.908 from 13.778 on Fold 4 with Finetuned Model 4. Further, the ratio of max–min accuracy was also reduced to 1.354 from 1.732 on Fold 2 with Finetuned Model 1, followed by 1.44 to 1.754 on Fold 3 with Finetuned Model 2, and finally 1.26 from 1.78 on Fold 4 with Finetuned Model 3. Further, the gap between the maximum and minimum has reduced to around 21% (initially, it was 54% (maximum of 100% for Asian and Indian

Table 15
Smile attribute classification accuracy (%) on LFW testset across different folds and gender-racial groups using expert labeling framework. M stands for Male, and F stands for Female. NS stands for Non-Smiling and S stands for Smiling Face. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race.

Race	Asian				Black				Indian				White				Max/Min↓	Overall↑	DoB↓
Gender	M		F		M		F		M		F		M		F				
	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S			
Baseline																			
Fold 1	98.46	55.769	100	76.67	91.025	71.428	85.714	81.48	95.918	58.14	100	77.78	96.01	71.023	94.34	81.98	1.793	83.48	13.823
Fold 2	95.21	57.73	100	76.47	92.42	63.49	85.71	85.71	96.552	71.053	100	63.63	96.35	74.65	95.94	78.16	1.732	83.318	13.76
Fold 3	97.059	57	93.33	68.33	87.037	74.194	69.231	93.33	95.652	70	100	93.75	96.635	74.213	95.109	74.937	1.754	83.739	13.194
Fold 4	98.413	56.25	100	73.08	91.549	68.254	100	74.19	92.98	72.093	100	72.22	96.976	72	93.72	82.258	1.78	84	13.778
Expert labeling-BPCOD																			
Fold 1	98.46	55.77	100	76.67	91.025	71.429	85.714	81.48	97.96	58.14	100	77.78	96.3	71.024	94.81	81.984	1.793	83.659	13.986
Fold 2	95.21	57.732	100	76.47	93.94	63.492	85.714	85.714	96.552	71.053	100	63.64	96.49	74.654	96.446	78.158	1.732	83.454	13.868
Fold 3	97.794	57	93.33	68.33	88.89	74.193	69.231	93.33	95.65	70	100	93.75	96.854	74.213	95.11	74.94	1.754	83.914	13.291
Fold 4	98.413	56.25	100	73.077	91.549	68.254	100	74.194	92.982	72.093	100	72.22	97.336	72	94.203	82.258	1.78	84.052	13.821
Expert labeling-EBOD																			
Fold 1	100	56.73	100	78.33	98.718	71.429	100	81.48	97.96	58.14	100	77.78	99.644	71.18	98.585	82.506	1.76	85.78	15.117
Fold 2	100	58.763	100	76.471	96.97	65.079	100	85.714	98.276	71.053	100	63.636	99.713	74.654	98.985	78.68	1.702	85.5	14.966
Fold 3	100	58	100	68.33	100	74.194	92.308	93.33	98.551	70	100	93.75	99.781	74.363	100	75.439	1.724	87.378	14.118
Fold 4	100	56.25	100	73.077	100	68.254	100	74.194	96.491	72.093	100	72.22	99.712	72	100	82.8	1.78	85.443	14.927

Table 16
Smile attribute classification accuracy (%) on LFW testset for ResNet-18 across different folds, and gender-racial groups using continual learning framework using prior preservation loss. M stands for Male, and F stands for Female. NS stands for Non-Smiling and S stands for Smiling Face. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from Fold *i* identified using BPCOD.

Race	Asian				Black				Indian				White				Max/Min↓	Overall↑	DoB↓
Gender	M		F		M		F		M		F		M		F				
	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S			
Baseline																			
Fold 1	98.46	55.769	100	76.67	91.025	71.428	85.714	81.48	95.918	58.14	100	77.78	96.01	71.023	94.34	81.98	1.793	83.48	13.823
Fold 2	95.21	57.73	100	76.47	92.42	63.49	85.71	85.71	96.552	71.053	100	63.63	96.35	74.65	95.94	78.16	1.732	83.318	13.76
7Fold 3	97.059	57	93.33	68.33	87.037	74.194	69.231	93.33	95.652	70	100	93.75	96.635	74.213	95.109	74.937	1.754	83.739	13.194
Fold 4	98.413	56.25	100	73.08	91.549	68.254	100	74.19	92.98	72.093	100	72.22	96.976	72	93.72	82.258	1.78	84	13.778
Finetuned Model 1																			
Fold 2	90.419	77.32	97.059	86.27	90.91	65.079	85.71	85.71	98.276	73.684	100	63.64	94.628	76.96	92.893	84.474	1.571	85.19	10.82
Finetuned Model 2																			
Fold 3	86.029	79	83.33	78.33	83.33	87.1	61.538	86.67	88.406	75	90.91	93.75	90.93	86.207	83.7	85.714	1.523	83.746	7.42
Finetuned Model 3																			
Fold 4	83.33	84.375	89.286	80.769	77.465	85.714	83.33	93.548	85.945	93.023	66.67	83.33	86.825	88.563	77.295	93.01	1.4	84.532	6.675

Fig. 8. Visualization of Grad-CAM(Top) and Combined Grad-CAM with Guided Back-propagation(Bottom) for Smile Attribute Classifiers: Baseline and Fine-tuned Models[1–3] (from left to right).[Best viewed in color.].

Females on multiple folds and a minimum of 56.25% for Asian males on Fold 4)), with a maximum accuracy of 100% for Indian Females and a minimum of 79.365% on Black males.

In summary, the human–machine partnership approach involving a continual learning process with EBOD as the outlier detection method on the LFW dataset led to an improvement in overall accuracy by up to 5%. This improvement could be due to the identification of more outliers compared to the previous continual learning framework that used BPCOD for outlier detection. It is also worth noting the difference in the distribution of the training data (CelebA) and the test distribution (LFW). Additionally, there was a significant reduction in the degree of bias, as evidenced by a decrease of up to 80% in the DoB (degree of bias) metric and the ratio of max–min accuracy.

Furthermore, for the smile attribute classification task, we observed improvements in the Grad-CAM visualizations when comparing the baseline model and the fine-tuned models obtained after each iteration of our proposed continual learning approach using prior preservation loss and EBOD for outlier detection (Fig. 8). Initially, the baseline model primarily focused on the bottom right part of the face when making smile predictions. However, through the continual learning framework, the iteratively trained models gradually shifted their attention to the regions near the mouth. These visualizations demonstrate that the continual learning process enabled the model better to identify relevant facial features for smile attribute classification. Specifically, the model's enhanced ability to focus on the mouth region, which is a crucial indicator of smiles, provides compelling evidence of the effectiveness of our continual learning approach in improving performance for this task.

Additionally, experimental results (See details at A.2) revealed a significant decline in accuracy up to 34% and increased discrimination

Table 17
Smile attribute classification accuracy (%) on LFW testset for ResNet-18 across different folds and gender-racial groups using a continual learning framework using prior preservation loss. M stands for Male, and F stands for Female. NS stands for Non-Smiling and S stands for Smiling Face. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model i is the fine-tuned model with the outliers from Fold i identified using EBOD.

Race	Asian				Black				Indian				White				Max/Min↓	Overall↑	DoB↓
Gender	M		F		M		F		M		F		M		F				
	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S			
Baseline																			
Fold 1	98.46	55.769	100	76.67	91.025	71.428	85.714	81.48	95.918	58.14	100	77.78	96.01	71.023	94.34	81.98	1.793	83.48	13.823
Fold 2	95.21	57.73	100	76.47	92.42	63.49	85.71	85.71	96.552	71.053	100	63.63	96.35	74.65	95.94	78.16	1.732	83.318	13.76
Fold 3	97.059	57	93.33	68.33	87.037	74.194	69.231	93.33	95.652	70	100	93.75	96.635	74.213	95.109	74.937	1.754	83.739	13.194
Fold 4	98.413	56.25	100	73.08	91.549	68.254	100	74.19	92.98	72.093	100	72.22	96.976	72	93.72	82.258	1.78	84	13.778
Finetuned Model 1																			
Fold 2	93.413	88.66	94.118	90.2	86.36	69.84	71.429	89.286	93.103	71.053	71.429	81.82	94.56	78.96	91.878	86.316	1.354	84.526	8.861
Finetuned Model 2																			
Fold 3	93.382	88	100	83.33	88.89	90.32	69.231	93.33	91.304	82.5	90.91	93.75	97.293	82.459	93.478	85.714	1.44	89	7.075
Finetuned Model 3																			
Fold 4	93.65	81.25	85.714	90.385	92.958	79.365	100	83.871	87.72	90.7	100	88.89	94.89	82.55	87.44	91.67	1.26	89.44	5.908

up to 42% for smile attribute classification without prior preservation loss, indicating catastrophic forgetting and underscoring its critical role in preserving knowledge and enhancing fairness in continual learning scenarios.

5.3. Comparative analysis with SOTA bias mitigation techniques

To comprehensively evaluate the effectiveness of our proposed approach, we conducted a comparative analysis against popular state-of-the-art (SOTA) bias mitigation techniques for gender classification. As most bias mitigation strategies are applied to the gender classification problem, we chose this task for our comparative analysis. We evaluated our approach alongside techniques based on multi-tasking (Das et al., 2018), adversarial debiasing (Zhang et al., 2018), deep generative views (Ramachandran & Rattani, 2023), and consistency regularization (Krishnan & Rattani, 2023), using the FairFace and UTKFace datasets. The algorithms were trained on the FairFace dataset and tested on both FairFace and UTKFace datasets. We compared these techniques against our proposed continual learning approach using the ResNet-18 model, which was trained with prior preservation loss and utilized EBOD for outlier detection. For evaluation, we used the fine-tuned model after the third iteration. We chose EBOD as the outlier detection mechanism because it strikes a good balance in terms of human intervention compared to other methods.

For the comparative analysis, we utilized overall classification accuracy, the ratio of maximum and minimum accuracy values, and the Degree of Bias (DoB) as evaluation metrics, as shown in Table 18. As evident from Table 18, our proposed human–machine partnership model obtained the lowest DoB and max–min ratio, and achieved the highest accuracy on the FairFace test set, while on the UTKFace dataset, it obtained the highest overall accuracy.

On the FairFace test set, our proposed human–machine partnership method achieved the lowest DoB of 1.164, the ratio of 1.037, and the highest overall accuracy of 96.47%. On the UTKFace dataset, our proposed method obtained the highest overall accuracy of 95.91%, while Krishnan and Rattani (2023) achieved the lowest DoB of 0.95 and the ratio of maximum and minimum accuracy values of 1.02.

It is noteworthy that an existing technique based on adversarial debiasing (Zhang et al., 2018) exhibited a trade-off between accuracy and fairness, attributable to the addition of the adversarial component, which reduced the model's generalization capacity. Additionally, adversarial debiasing and multi-tasking (Das et al., 2018; Zhang et al., 2018) based bias mitigation techniques require demographically annotated data. Furthermore, the technique based on deep generative views is computationally expensive and limited in its ability to synthesize images of the 3D scene with multi-view consistency. Consequently, in comparison to the existing State-of-the-Art (SOTA) methodologies which are in-processing techniques, our proposed technique is applied during test time. It can be used with already deployed models. Our approach has the advantage of reducing bias even when protected attributes are not available, and it can be applied across different domains and applications. Most importantly, it incorporates a human subject matter expert, making it a reliable and trustworthy approach during testing. Worth mentioning, our proposed methodology offers the benefit of improved fairness without compromising classification accuracy.

6. Key findings

In this section, we will address the important findings and observations from the experiments conducted:

- The visual examples in Fig. 9 highlight outlier instances, detected using the EBOD method, that deviates substantially from normal cases, posing challenges for accurate gender and smile attribute classification.
- Systematic approach to harness sample prediction uncertainty (outliers), and human–machine partnership at test time, enhances the fairness of the system without compromising its classification performance.
- With our proposed human–machine partnership approach, the groups that had the lowest performance (Black ethnicity in FairFace & Asian ethnicity in UTKFace for gender classification, and Asian Males in LFW for smile attribute classification) when using the baseline framework improved their classification accuracy. Importantly, this improvement did not negatively impact the groups that performed best (Middle Eastern ethnicity in FairFace, & White ethnicity in UTKFace for gender classification, and White Females in LFW for smile attribute classification) with the baseline framework.
- With the human–machine partnership approach involving expert labeling, the accuracy improved by up to 6% for the Black demographic group in the FairFace dataset, & 5% for the Asian demographic group in the UTKFace dataset for gender classification, and 1% for Asian Males in the LFW dataset for smile attribute classification. Similarly, through the third iteration of continual learning, the accuracy further increased by up to 3% for the Black group in FairFace & 3% for the Asian group in UTKFace for gender classification and 10% for Asian Males in LFW for smile attribute classification.
- Overall, the human–machine partnership with a continual learning framework improved classification accuracy (up to 3% for gender and 5% for smile attribute) and significantly reduced bias (up to 60% for gender and 80% for smile attribute) as shown in Figs. 10–12.
- The GradCAM visualizations in Figs. 7 and 8 compellingly demonstrate that incorporating the continual learning framework enabled the models to focus on more relevant facial regions, resulting in enhanced discrimination capabilities for both the gender and smile attribute classification tasks.

Table 18
Comparative analysis of gender classification task. A: Multi-Tasking (Das et al., 2018), B: Adversarial debiasing (Zhang et al., 2018), D: Deep Generative Views (Ramachandran & Rattani, 2023). E: Consistency Regularization (Krishnan & Rattani, 2023). The top performance results are highlighted in bold.

Method	Accuracy								DoB↓	Max/Min↓
	Black	East Asian	Indian	Latino Hispanic	Middle Eastern	Southeast Asian	White	Overall↑		
FairFace										
A	91.26	94.45	95.05	95.19	97.35	94.2	94.96	94.64	1.81	1.067
B	87.66	91.93	93.67	93.8	95.96	91.81	93.96	92.69	2.62	1.095
D	91.64	95.29	95.38	95.32	97.11	93.5	94.92	94.72	1.72	1.06
E	90.83	93.6	94.48	94.7	95.94	93.64	94.57	94	1.59	1.056
Ours	94.25	96.34	97.59	96.65	97.76	96.03	96.67	**96.47**	**1.164**	**1.037**
UTKFace										
B	94.62	–	93.65	–	–	91.89	94.97	93.78	1.38	1.03
E	95.85	–	95.43	–	–	93.67	95.16	95.03	**0.95**	**1.02**
Ours	97.17	–	96.17	–	–	93.49	96.82	**95.91**	1.67	1.039

Fig. 9. Illustrates examples of normal and uncertain/outlier cases for gender classification and smile attribute classification tasks. The top left shows normal examples for gender classification, and the bottom left shows normal examples for smile attribute classification. The top right displays outlier examples identified by the EBOD method for gender classification, while the bottom right shows outlier examples detected by EBOD for smile attribute classification. These outlier examples on the right exhibit unique characteristics like occlusion, varying poses, and facial expressions, which can make classification particularly challenging for the model.

- Finally, experimental results (A.1 and A.2) revealed a significant decline in accuracy (up to 21% for gender and 34% for smile attribute) and increased discrimination (up to 200% for gender and 42% for smile attribute) without prior preservation loss using a continual learning framework, indicating catastrophic forgetting and underscoring its critical role in preserving knowledge and enhancing fairness in continual learning scenarios.
- We observed that incorporating the uncertain samples via expert labeling or continual learning did not significantly enhance the overall classification accuracy, this is due to the nature of incremental learning challenge, as the existing weights and parameters are heavily biased towards the initial training data distribution, and outliers in different data folds could be in small numbers to overpower the existing distribution. However, we observed a performance gain of up to 5% in the third iteration of continual learning. The performance of the classifier is bound to improve with the number of adaptations in an iterative manner.

6.1. Explanation of accuracy vs. Fairness trade-off

To explain the modest improvement in overall accuracy but significant reduction in bias, we can refer to Table 12 as an example. The accuracy boost is incremental (around 1%), while the substantial reduction in bias (around 50%) is a notable achievement. This indicates that the fine-tuned model is better calibrated and less discriminatory towards specific demographic subgroups.

From the baseline results on Fold 4, there is a notable gap of around 12% between the highest accuracy (98.656 for Black Males) and the lowest accuracy (86.797% for Asian Females). This large gap indicates a high degree of bias in the baseline model's performance across different demographic groups. However, after the third iteration of re-training (Finetuned 3) on Fold 4, the highest accuracy is 96.106% for Black Females, and the lowest is 91.473% for Asian Males. The gap between the highest and lowest accuracy has been reduced to around 4.6%. This reduction in the gap between the maximum and minimum accuracy values across demographic groups is reflected in the ratio of max–min accuracies, which decreases from 1.137 for the baseline to 1.05 for FineTuned 3 (a lower value indicates less disparity).

Additionally, the fine-tuned model has improved the accuracy for underperforming groups like Asian Females (from 86.797% to 91.558%) and Indian Males (from 95.841% to 92.374%). Consequently, the standard deviation of accuracy values across gender and race groups, represented by the DoB, has decreased substantially from 3.58 for the baseline to 1.562 for FineTuned 3. Therefore, it is evident that the continual learning process has helped to uplift the performance of underperforming demographic groups while maintaining or slightly improving the performance of well-performing groups. This has led to a more uniform and fair distribution of accuracy across different gender and race categories, resulting in a significant reduction in the degree of bias while still providing a modest overall accuracy gain.

The key inferences suggest that a human–machine partnership with continual learning exhibits a higher potential for mitigating bias without compromising the generalization capacity, thereby enhancing the fairness of facial attribute classification compared to a static expert labeling framework.

7. Conclusion and future work

Much of the existing machine learning-based fairness literature is in-processing techniques that assume the presence of protected attributes during the training stage, such as ethnicity and sex, for bias mitigation. However, in practice, the collection of protected features, or their use for training or inference is often precluded due to privacy and regulation. This severely limits the applicability of traditional fairness

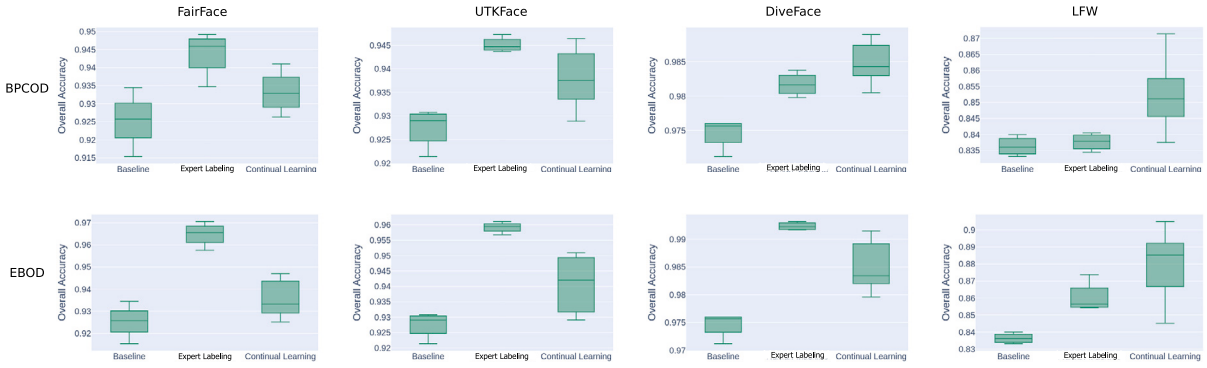

Fig. 10. Boxplots of overall classification accuracy for ResNet18 model: Baseline, expert labeling, and continual learning frameworks across different folds for gender classification (FairFace, UTKFace & DiveFace) and smile attribute classification (LFW).

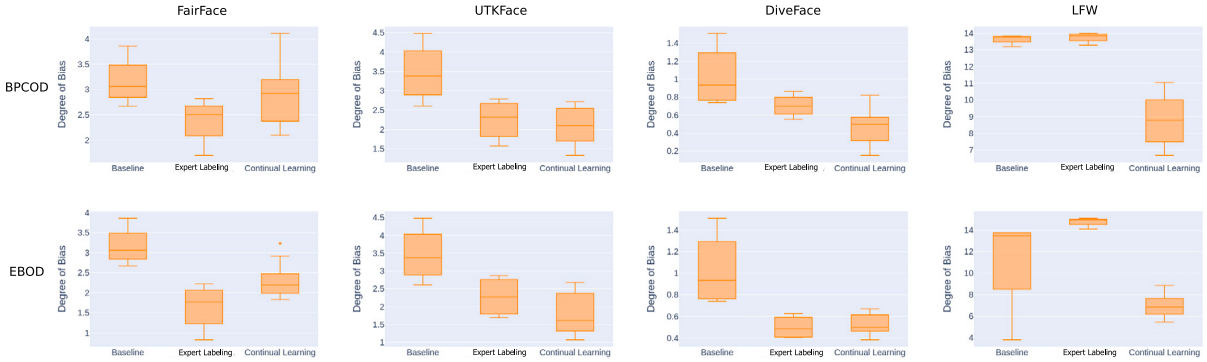

Fig. 11. Boxplots of degree of bias (DoB) for ResNet18 model: Baseline, expert labeling, and continual learning frameworks across different folds for gender classification (FairFace, UTKFace & DiveFace) and smile attribute classification (LFW).

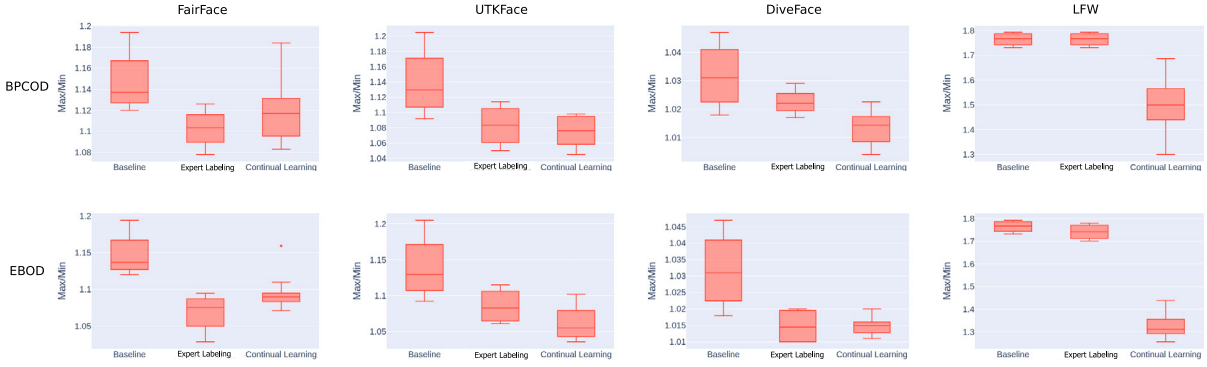

Fig. 12. Boxplots of the ratio of max–min accuracies for ResNet18 model: Baseline, expert labeling, and continual learning frameworks across different folds for gender classification (FairFace, UTKFace & DiveFace) and smile attribute classification (LFW).

research. Further, existing approaches to mitigating bias may offer a trade-off between fairness and classification performance. Furthermore, the existing in-processing bias mitigation techniques cannot be utilized for already deployed models.

We have proposed a novel approach for mitigating bias at test-time by incorporating minimal human–machine partnership with expert labeling and continual learning for facial attribute classifiers as a case study. By leveraging the expertise of human experts to label the outliers/uncertain data samples, we have demonstrated the better effectiveness of fine-tuning a deep neural network through an iterative process that combines human guidance with machine learning over expert labeling. Through extensive experimentation on gender and smile attribute classification tasks, our approach has shown significant improvements in both accuracy and fairness. The results of our experiments indicate that our method achieves a noteworthy 2% improvement in gender classification accuracy and a substantial 5% improvement in smile attribute classification accuracy when compared to baseline models. Furthermore, our approach has demonstrated its potential in reducing bias at test-time, with an impressive over 80% reduction in bias for both gender and smile attribute classification tasks.

As a part of future work, the efficacy of our proposed mitigation techniques will be evaluated for multi-attribute classifiers across several protected attributes. Lastly, we will extend our work to address demographic bias mitigation in other domains such as general image classification and natural language processing.

CRediT authorship contribution statement

Anoop Krishnan Upendran Nair: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Formal analysis, Investigation, Writing – reviewing & Editing. **Ajita Rattani:** Funding acquisition, Investigation, Project administration, Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part by National Science Foundation (NSF), United States award no. 2129173. This work was done when Anoop was a PhD. student at Wichita State University.

Appendix

This section discusses the performance of the gender classifier and the smile attribute classifier when using a continual learning framework without the prior preservation loss during iterative training on ResNet-18 architecture. The gender classifier was initially trained on the FairFace training set and then evaluated on different non-overlapping test folds from the FairFace, UTKFace, and DiveFace datasets. The smile attribute classifier was trained on the CelebA training set and evaluated on different non-overlapping test folds from the LFW dataset. The performance was assessed across different genders and races. The protocol for retraining the models followed the same approach as described in the paper. Additionally, we have presented the results of gender classifiers trained on the Vision Transformer (ViT) architecture. We observed a similar trend as the models trained on the ResNet-18 architecture. Therefore, due to the straightforward and expected nature of the results, we have omitted the evaluation of the smile attribute classifier trained on ViT. However, we want to emphasize that the trends observed for the smile attribute classification task were equally compelling and aligned with the results of the one trained on ResNet-18, further solidifying the efficacy of our approach.

A.1. Gender classifier

Tables 19, 20, and 21 present the evaluation results of the baseline and fine-tuned gender classification models without the prior preservation loss on the FairFace, UTKFace, and DiveFace test sets, respectively, across different gender and race subgroups. The outliers in the datasets were identified using the Boundary Proximity Confidence-based Outlier Detection (BPCOD) method. Similarly, Tables 22, 23, and 24 exhibit the same evaluation results but with outliers identified using the Ensemble-Based Outlier Detection (EBOD) method. It is evident from the above-mentioned tables that the overall classification dropped up to 21% over each iteration, accompanied by an increasing rate of bias up to 200%.

A.2. Smile attribute classifier

Table 25 presents the evaluation results of the baseline and fine-tuned smile attribute classification models without the prior preservation loss on the LFW test sets, across different gender and race subgroups. The outliers in the datasets were identified using the Boundary Proximity Confidence-based Outlier Detection (BPCOD) method. Similarly, Table 26 exhibits the same evaluation results but with outliers identified using the Ensemble-Based Outlier Detection (EBOD) method. Similarly, on the gender classification task, we observed decreasing classification performance up to 34% as each fine-tuning took place, on top of that, the decreased fairness up to 42% was also observed.

A.3. Classifiers trained on ViT architecture

Tables 27–44 show the evaluation of gender classifiers trained on the Vision Transformer (ViT) architecture. From the Tables 28, 34, and 40, by applying our proposed expert learning framework, the overall classification accuracy on the FairFace test set improved by up to 4% when using BPCOD for outlier detection and 5% when using EBOD. Similarly, on the UTKFace dataset, improvements of 2% and 4% were observed when using BPCOD and EBOD, respectively. Additionally, an improvement of almost 1% was observed on the DiveFace test set when using both BPCOD and EBOD, as the DiveFace test set already exhibited fair performance even with the baseline models.

Regarding bias reduction, the FairFace test set showed a bias reduction of up to 20% when using BPCOD as the outlier detection method and up to 48% when using EBOD. Similarly, on the UTKFace dataset, a bias reduction of up to 36% was observed when utilizing BPCOD, and up to 45% when using EBOD. Finally, on the DiveFace dataset, a bias reduction of up to 50% was observed when using BPCOD, and up to 40% when using EBOD.

Additionally from the Tables 29, 30, 35, 36, 41 and 42, when applying continual learning with prior preservation loss, as shown in the tables, the overall accuracy on the FairFace test set slightly improved by up to 2% for both BPCOD and EBOD, as this is an iterative process. The bias was reduced by up to 12% when using BPCOD and up to 22% when using EBOD. Similarly, on the UTKFace dataset, the overall accuracy improved by up to 1%–2% for both BPCOD and EBOD, and the bias reduction was up to 15% for BPCOD and 28% for EBOD. Finally, on the DiveFace dataset, a modest increment of 1% in overall accuracy was observed when using both BPCOD and EBOD, and the bias reduction was up to 22% when using BPCOD and 62% when using EBOD.

Moreover from the Tables 31, 32, 37, 38, 43 and 44, when continual learning was performed without using prior preservation loss, the overall accuracy decreased by up to 20%, and the bias increased up to six times compared to the baseline model's performance. Similar trends were observed for the smile attribute classification task on the ViT architecture, where our proposed continual learning framework using EBOD for outlier detection significantly reduced bias and modestly improved classification performance.

In summary, by leveraging the expertise of human experts to label outliers and uncertain data samples, and fine-tuning deep neural networks through an iterative process that combines human guidance with machine learning based on expert labeling, we can improve the fairness of facial-based gender and smile attribute classification tasks without negatively affecting the model's ability to generalize to new data.

Data availability

Data will be made available on request.

Table 19
Gender classification accuracy (%) on FairFace testset on ResNet18 architecture across different folds and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using BPCOD without the prior preservation loss.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min(↓)	Overall(↑)	DoB(↓)
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Baseline																	
Fold 1	85.643	85.96	91.453	91.1	94.54	93.846	89.44	92.54	97.087	94.624	91.379	91.011	93.662	89.431	1.134	91.54	3.108
Fold 2	89.64	84.7	93.01	93.264	94.5	94.38	94.86	96.55	96.208	94.9	93.407	94.631	93.214	94.98	1.14	93.446	2.91
Fold 3	90.164	80.2	92.353	94.413	95.19	94.15	93.9	94.089	95.098	95.495	90.206	95.745	93.725	91.428	1.194	92.58	3.86
Fold 4	93.229	87.437	94.118	93.33	93.989	88.442	95.161	92.453	97.92	90.426	90.27	91.52	94.719	93.133	1.12	92.58	2.67
Finetuned Model 1																	
Fold 2	75.225	74.863	74.731	89.119	78	89.89	78.505	93.596	84.834	91.837	79.121	79.866	80	92.47	1.252	83	6.774
Finetuned Model 2																	
Fold 3	74.863	68.02	87.647	68.156	81.25	78.723	79.343	82.76	86.765	87.387	83.505	61.702	80.39	72.653	1.42	78.083	7.7
Finetuned Model 3																	
Fold 4	64.583	77.89	77.54	71.905	63.388	79.9	73.66	83.02	70.31	86.17	77.838	63.03	69.637	83.69	1.367	74.47	7.35

Table 20
Gender classification accuracy (%) on UTKFace testset on ResNet18 architecture across different folds and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using BPCOD without the prior preservation loss.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Baseline											
Fold 1	91.036	88.387	96.078	91.058	96.516	93.12	94.815	93.615	1.092	93.078	2.607
Fold 2	91.198	86.433	97	92.16	96.697	93.056	94.95	92.427	1.122	93	3.181
Fold 3	91.686	81.435	98.107	92.13	93.45	92.732	94.968	92.602	1.205	92.139	4.482
Fold 4	89.922	86.797	98.656	90.796	95.841	91.05	95.326	93.896	1.137	92.786	3.58
Finetuned Model 1											
Fold 2	77.017	90.591	87.874	93.554	84.404	94.68	85.137	93.978	1.23	88.404	5.682
Finetuned Model 2											
Fold 3	63.183	91.983	86.231	91.55	79.823	93.484	84.975	90.95	1.48	85.273	9.372
Finetuned Model 3											
Fold 4	68.217	87.88	88.676	78.23	79.896	82.327	84.125	78.344	1.3	80.96	6.094

Table 21
Gender classification accuracy (%) on DiveFace testset on ResNet18 architecture across different folds and different demographics. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using BPCOD without the prior preservation loss.

Race	East Asian		Sub-Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Baseline									
Fold 1	97.51	98.84	97.642	96.237	98.074	97.287	1.027	97.6	0.79
Fold 2	97.94	96.953	97.82	94.97	99.46	95.56	1.047	97.118	1.51
Fold 3	98.03	99.33	97.02	96	98.053	96.739	1.035	97.529	1.079
Fold 4	97.1	98.66	96.92	97.3	98.61	97.015	1.0179	97.6	0.74
Finetuned Model 1									
Fold 2	90.449	98.387	89.094	97.67	92.014	97.513	1.104	94.187	3.773
Finetuned Model 2									
Fold 3	88.89	99.162	93.296	95.636	90.973	97.1	1.116	94.176	3.525
Finetuned Model 3									
Fold 4	84.783	95.97	84.42	93.33	85.417	95.854	1.137	89.963	5.17

Table 22
Gender classification accuracy (%) on FairFace testset on ResNet18 architecture across different folds and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using EBOD without the prior preservation loss.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min(↓)	Overall(↑)	DoB(↓)
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Baseline																	
Fold 1	85.64	85.96	91.45	91.1	94.54	93.85	89.44	92.54	97.09	94.62	91.38	91.01	93.66	89.43	1.13	91.54	3.11
Fold 2	89.64	84.7	93.01	93.26	94.5	94.38	94.86	96.55	96.21	94.9	93.41	94.63	93.21	94.98	1.14	93.45	2.91
Fold 3	90.16	80.2	92.35	94.41	95.19	94.15	93.9	94.09	95.1	95.5	90.21	95.74	93.72	91.43	1.19	92.58	3.86
Fold 4	93.23	87.44	94.12	93.33	94	88.44	95.16	92.45	97.92	90.43	90.27	91.52	94.72	93.13	1.12	92.58	2.67
Finetuned Model 1																	
Fold 2	74.32	77.6	74.73	89.12	74	88.76	80.84	89.66	85.78	91.84	73.08	79.87	78.21	90.8	1.26	82.04	6.75
Finetuned Model 2																	
Fold 3	48.09	73.1	87.06	56.98	73.08	79.79	77.46	75.86	81.86	73.87	69.59	62.77	85.49	55.51	1.81	71.46	11.26
Finetuned Model 3																	
Fold 4	61.98	73.87	76.47	67.14	65.57	82.41	75.27	79.72	75.52	78.72	74.59	73.33	76.9	68.24	1.33	73.55	5.59

Table 23
Gender classification accuracy (%) on UTKFace testset on ResNet18 architecture across different folds and different demographics. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using EBOD without the prior preservation loss.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Baseline											
Fold 1	91.04	88.39	96.08	91.06	96.52	93.12	94.82	93.62	1.09	93.08	2.61
Fold 2	91.2	86.43	97	92.16	96.7	93.06	94.95	92.43	1.12	93	3.18
Fold 3	91.69	81.44	98.11	92.13	93.45	92.73	94.97	92.6	1.2	92.14	4.48
Fold 4	89.92	86.8	98.66	90.8	95.84	91.05	95.33	93.9	1.14	92.79	3.58
Finetuned Model 1											
Fold 2	67.48	91.25	74.42	95.3	76.15	93.29	81.89	92.24	1.41	84	9.78
Finetuned Model 2											
Fold 3	60.1	89.45	66.95	92.71	74.16	89.72	77.35	86.34	1.54	79.6	11.13
Finetuned Model 3											
Fold 4	58.14	89.39	71.78	90.44	63.6	90.6	69.29	88.54	1.5	77.72	12.6

Table 24
Gender classification accuracy (%) on DiveFace testset on ResNet18 architecture across different folds and different demographics. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using EBOD without the prior preservation loss.

Race	East Asian		Sub-Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Baseline									
Fold 1	97.51	98.84	97.642	96.237	98.074	97.287	1.027	97.6	0.79
Fold 2	97.94	96.953	97.82	94.97	99.46	95.56	1.047	97.118	1.51
Fold 3	98.03	99.33	97.02	96	98.053	96.739	1.035	97.529	1.079
Fold 4	97.1	98.66	96.92	97.3	98.61	97.015	1.0179	97.6	0.74
Finetuned Model 1									
Fold 2	90.075	97.31	90.1	98.025	93.103	97.51	1.088	94.355	3.42
Finetuned Model 2									
Fold 3	78.136	99.33	89.013	95.45	82.12	97.283	1.27	90.22	7.89
Finetuned Model 3									
Fold 4	74.82	98.273	82.246	95.135	81.076	97.512	1.313	88.18	9.142

Table 25
Smile attribute classification accuracy (%) on LFW testset on ResNet18 architecture across different folds and different demographics. M stands for Male, and F stands for Female. NS stands for Non-Smiling and S stands for Smiling Face. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using BPCOD without the prior preservation loss.

Race	Asian				Black				Indian				White				Max/Min↓	Overall↑	DoB↓
Gender	M		F		M		F		M		F		M		F				
	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S			
Baseline																			
Fold 1	98.46	55.77	100	76.67	91.025	71.43	85.71	81.48	95.918	58.14	100	77.78	96.01	71.023	94.34	81.98	1.79	83.48	13.82
Fold 2	95.21	57.73	100	76.47	92.42	63.49	85.71	85.71	96.55	71.05	100	63.63	96.35	74.65	95.94	78.16	1.73	83.32	13.76
Fold 3	97.06	57	93.33	68.33	87.04	74.19	69.23	93.33	95.65	70	100	93.75	96.63	74.21	95.11	74.94	1.75	83.74	13.19
Fold 4	98.41	56.25	100	73.08	91.55	68.25	100	74.19	92.98	72.09	100	72.22	96.98	72	93.72	82.26	1.78	84	13.78
Finetuned Model 1																			
Fold 2	59.28	86.6	61.76	94.12	63.64	73.02	28.57	92.86	75.86	86.84	42.86	63.64	64.54	87.3	53.81	94.21	3.3	70.56	18.71
Finetuned Model 2																			
Fold 3	50	58	33.33	63.33	42.59	38.71	53.85	43.33	52.17	40	54.54	25	57.06	53.82	55.98	52.88	2.53	48.41	9.94
Finetuned Model 3																			
Fold 4	59.52	42.71	67.86	42.31	60.56	33.33	83.33	32.26	52.63	39.53	50	33.33	63	46.77	66.67	38.71	2.58	50.78	14.37

Table 26
Smile attribute classification accuracy (%) on LFW testset on ResNet18 architecture across different folds and different demographics. M stands for Male, and F stands for Female. NS stands for Non-Smiling and S stands for Smiling Face. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using EBOD without the prior preservation loss.

Race	Asian				Black				Indian				White				Max/Min↓	Overall↑	DoB↓
Gender	M		F		M		F		M		F		M		F				
	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S			
Baseline																			
Fold 1	98.46	55.769	100	76.67	91.025	71.428	85.714	81.48	95.918	58.14	100	77.78	96.01	71.023	94.34	81.98	1.793	83.48	13.823
Fold 2	95.21	57.73	100	76.47	92.42	63.49	85.71	85.71	96.552	71.053	100	63.63	96.35	74.65	95.94	78.16	1.732	83.318	13.76
Fold 3	97.059	57	93.33	68.33	87.037	74.194	69.231	93.33	95.652	70	100	93.75	96.635	74.213	95.109	74.937	1.754	83.739	13.194
Fold 4	98.413	56.25	100	73.08	91.549	68.254	100	74.19	92.98	72.093	100	72.22	96.976	72	93.72	82.258	1.78	84	13.778
Finetuned Model 1																			
Fold 2	82.036	85.567	64.71	92.157	71.21	71.429	71.429	96.429	72.414	76.316	57.143	81.82	84.24	88.48	71.066	92.895	1.687	78.71	10.66
Finetuned Model 2																			
Fold 3	77.206	78	86.67	83.33	83.33	79.03	69.231	80	78.261	87.5	63.64	87.5	88.66	80.21	82.609	85.714	1.393	80.68	6.547
Finetuned Model 3																			
Fold 4	56.349	89.58	71.429	98.077	80.282	74.603	66.67	90.322	64.912	86.046	50	100	77.754	85.63	68.116	93.817	2	78.349	14.262

Table 27
Gender classification accuracy (%) on FairFace testset on ViT architecture across different folds, and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and races; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across genders and races.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Baseline																	
Fold 1	88.24	88.64	94.44	94.24	95.15	95.9	92.22	93.87	97.09	97.85	94.25	92.13	94.72	95.53	1.11	93.88	2.68
Fold 2	90.22	86.67	95.7	95.85	96	94.38	94.86	97.54	97.63	97.96	93.96	94.63	94.29	95.4	1.13	94.65	2.89
Fold 3	91.89	87.69	93.53	93.86	94.66	96.32	95.31	94.09	95.59	98.2	89.69	94.68	95.67	92.65	1.12	93.845	2.61
Fold 4	90.67	89.9	93.05	93.33	96.15	94.5	94.09	94.34	97.4	90.43	94.05	93.33	97.03	94.85	1.08	93.79	2.21

Table 28
Gender classification accuracy (%) on FairFace testset on ViT architecture across different folds, and gender-racial groups on expert labeling framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and races; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across genders and races.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Expert labeling-BPCOD																	
Fold 1	90.41	92.69	94.88	97.49	95.76	99.3	95.2	96.17	97.57	98.7	95.44	95.55	96.14	99.3	1.1	96.04	2.46
Fold 2	92.34	92.58	99.34	98.87	97.524	96.63	95.33	99.41	98.59	99.15	93.96	95.97	95.37	97.08	1.08	96.58	2.43
Fold 3	94.73	98.24	98.91	99.35	96.1	99.54	96.26	99.87	96.573	99	93.64	96.26	95.69	94.31	1.07	97.03	2.09
Fold 4	92.78	99.63	99.18	99.78	97.27	99.54	96.2	98.71	97.4	99.87	95.2	97.032	97.37	95.73	1.076	97.55	2.08
Expert labeling-EBOD																	
Fold 1	94.36	97.9	97.09	99.11	97.6	99.3	96.23	97.61	97.57	98.7	97.81	97.25	98.28	99.3	1.052	97.722	1.31
Fold 2	94.64	95.06	99.37	99.87	98.54	96.07	97.2	99.41	99.07	99.15	96.72	96.64	98.26	97.92	1.05	97.66	1.62
Fold 3	96.9	99.35	98.84	99.2	98.01	99.54	96.73	99.41	97.559	99.15	93.28	97.31	97.69	98.03	1.07	97.97	1.68
Fold 4	97.18	99.63	99.76	98.92	98.95	99.54	95.68	99.41	97.914	99.15	99.69	98.89	99.74	98.35	1.04	98.76	1.15

Table 29
Gender classification accuracy (%) on FairFace testset on ViT architecture across different folds, and gender-racial groups on a continual learning framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from Fold *i* identified using BPCOD.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Finetuned Model 1																	
Fold 2	91.25	90.1	97.91	97.45	98.03	95.7	94.7	98.04	98.11	94.8	94.8	95.7	94.7	96.2	1.09	95.53	2.47
Finetuned Model 2																	
Fold 3	92.3	89.2	94.5	95.16	95.52	94.5	95.33	97.03	98.13	97.96	93.42	94.1	95.06	93.69	1.1	94.71	2.29
Finetuned Model 3																	
Fold 4	91.63	91.1	95.3	95.31	97.2	97.95	95.31	96.48	98.21	97	95.84	97.81	97.56	95.7	1.08	95.88	2.17

Table 30
Gender classification accuracy (%) on FairFace testset on ViT architecture across different folds, and gender-racial groups on a continual learning framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from Fold *i* identified using EBOD.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Finetuned Model 1																	
Fold 2	91.71	92.3	94.04	95.32	99.05	95.7	94.7	96.55	97.15	97.96	94.8	95.7	94.7	96.2	1.08	95.42	2
Finetuned Model 2																	
Fold 3	92.3	91.89	99.48	95.61	97.47	94.5	95.8	96.528	97.63	96.15	96.16	94.63	96.29	94.12	1.08	95.61	2.03
Finetuned Model 3																	
Fold 4	92.45	91.3	94.75	95.89	97.2	97.38	94.77	94.56	96.1	97.08	94.27	95.91	95.352	95.7	1.07	95.19	1.73

Table 31
Gender classification accuracy (%) on FairFace testset on ViT architecture across different folds and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using BPCOD without the prior preservation loss.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Finetuned Model 1																	
Fold 2	75.71	76.6	76.89	91.59	79.24	89.9	78.51	94.553	86.09	94.8	79.59	79.86	80.92	92.88	1.25	84.09	7.21
Finetuned Model 2																	
Fold 3	76.3	74.37	88.76	67.75	80.8	80.53	80.53	82.76	87.21	89.86	83.03	61.02	82.07	73.63	1.47	79.19	8.02
Finetuned Model 3																	
Fold 4	62.81	80.08	76.66	71.9	64.85	85.37	72.828	84.71	69.94	86.17	81.1	64.28	71.34	85.23	1.37	75.52	8.38

Table 32
Gender classification accuracy (%) on FairFace testset on ViT architecture across different folds and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using EBOD without the prior preservation loss.

Race	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Finetuned Model 1																	
Fold 2	74.8	79.4	76.89	91.59	75.17	88.77	80.84	90.58	87.05	94.87	73.51	79.86	79.11	91.2	1.29	83.11	7.262
Finetuned Model 2																	
Fold 3	49	79.92	88.17	56.65	72.67	81.62	78.62	75.86	82.28	75.97	69.19	62.07	87.28	56.25	1.8	72.54	12.23
Finetuned Model 3																	
Fold 4	60.28	75.95	75.6	67.14	67.08	88.05	74.42	81.34	75.12	78.72	77.71	74.78	78.77	69.5	1.46	74.61	6.89

Table 33
Gender classification accuracy (%) on UTKFace testset on ViT architecture across different folds, and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and races; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across genders and races.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Baseline											
Fold 1	91.43	91.84	98.39	89.28	98.18	97.56	96.18	96.19	1.1	94.88	3.28
Fold 2	91.43	88.37	100	91.8	92.59	97.67	92.75	95.93	1.13	93.82	3.53
Fold 3	91.84	93.88	100	92.86	96.43	92.1	95.56	95.83	1.1	94.81	2.55
Fold 4	92.1	90.91	95.56	97.87	88.52	93.88	97.9	91.17	1.1	93.49	3.19

Table 34
Gender classification accuracy (%) on UTKFace testset on ViT architecture across different folds, and gender-racial groups on expert labeling framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and races; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across genders and races.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Expert labelings-BPCOD											
Fold 1	92.836	95.636	100	91.971	98.369	99.007	96.858	97.521	1.087	96.525	2.875
Fold 2	92.409	93.07	100	94.059	92.944	99.617	93.318	98.776	1.082	95.524	3.311
Fold 3	92.79	99.13	100	94.986	97.89	95.838	96.355	97.902	1.078	96.861	2.346
Fold 4	94.487	94.31	95.56	98.2	92.61	95.724	98.664	94.75	1.065	95.538	2.025
Expert labeling-EBOD											
Fold 1	95.93	96.082	99.726	96.13	99.776	99.007	97.987	99.825	1.041	98.058	1.769
Fold 2	97.306	95.81	100	96	95.61	99.374	95.009	99.724	1.053	97.354	2.051
Fold 3	97.786	99.13	100	96.919	98.254	95.092	97.441	99.794	1.052	98.052	1.627
Fold 4	98.722	94.537	95.932	98.2	98.83	96.187	99.655	97.56	1.054	97.453	1.746

Table 35
Gender classification accuracy (%) on UTKFace testset on ViT architecture across different folds, and gender-racial groups on a continual learning framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from Fold *i* identified using BPCOD.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Finetuned Model 1											
Fold 2	91.919	92.175	99.154	94.56	93.68	97.188	92.119	97.734	1.079	94.817	2.859
Finetuned Model 2											
Fold 3	90.637	95.457	96.73	92.1	95.696	93.46	96.148	95.564	1.0679	94.474	2.169
Finetuned Model 3											
Fold 4	93.963	91.135	95.94	97.317	92	94.5	97.825	92	1.0739	94.335	2.539

Table 36
Gender classification accuracy (%) on UTKFace testset on ViT architecture across different folds, and gender-racial groups on a continual learning framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from Fold *i* identified using EBOD.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Finetuned Model 1											
Fold 2	92.58	91.87	97.608	94.234	94.19	97.674	94.81	97.069	1.063	95	2.241
Finetuned Model 2											
Fold 3	94.36	93.878	99.467	95.74	97.8	95.83	96.512	95.2	1.059	96.1	1.827
Finetuned Model 3											
Fold 4	94.23	93.111	98.49	98.416	96.98	94.5	98.367	96.76	1.058	96.36	2.134

Table 37
Gender classification accuracy (%) on UTKFace testset on ViT architecture across different folds, and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using BPCOD without the prior preservation loss.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Finetuned Model 1											
Fold 2	77.212	92.62	90.592	93.19	80.82	99.38	83.17	97.544	1.287	89.316	8.044
Finetuned Model 2											
Fold 3	63.28	91.57	87.895	92.272	82.37	92.852	85.505	94.12	1.487	86.23	10.11
Finetuned Model 3											
Fold 4	69.87	92.04	85.89	84.327	73.8	84.883	86.4	76.074	1.317	81.66	7.537

Table 38
Gender classification accuracy (%) on UTKFace testset on ViT architecture across different folds, and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using EBOD without the prior preservation loss.

Race	Asian		Black		Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F	M	F			
Finetuned Model 1											
Fold 2	67.651	93.294	76.721	94.927	72.915	97.916	80	95.741	1.447	84.895	11.893
Finetuned Model 2											
Fold 3	60.194	90.57	68.245	93.438	76.522	89.117	77.836	89.35	1.552	80.659	12
Finetuned Model 3											
Fold 4	59.55	93.63	69.532	97.49	58.749	93.417	71.161	85.98	1.659	78.69	15.817

Table 39
Gender classification accuracy (%) on DiveFace testset on ViT architecture across different folds, and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and races; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across genders and races.

Race	East Asian		Sub Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Baseline									
Fold 1	97.865	99.67	97.249	97.491	97.723	99.457	1.025	98.242	1.0249
Fold 2	98.127	98.746	98.154	98.205	99.0926	98.4	1.01	98.454	0.35
Fold 3	98.028	99.665	96.462	98.73	98.053	98.551	1.033	98.248	0.967
Fold 4	97.283	99.04	96.739	98.74	99.132	98.839	1.025	98.295	0.93

Table 40
Gender classification accuracy (%) on DiveFace testset on ViT architecture across different folds, and gender-racial groups on expert labeling framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among genders and races; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across genders and races.

Race	East Asian		Sub Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Expert labeling-BPCOD									
Fold 1	99.7	100	98.032	99.2	97.897	100	1.021	99.138	0.95
Fold 2	99.13	100	98.658	100	99.269	99.51	1.013	99.42	0.523
Fold 3	99.54	100	98.51	100	98.053	99.3	1.02	99.2	0.798
Fold 4	99.79	100	99.18	100	99.132	100	1	99.68	0.417
Expert labeling-EBOD									
Fold 1	99.471	100	98.618	99.2	98.596	100	1.0142	99.31	0.629
Fold 2	99.253	100	99.338	100	99.451	99.51	1.007	99.592	0.3283
Fold 3	99.102	100	97.574	100	99.47	99.3	1.025	99.241	0.895
Fold 4	99.099	100	97.817	100	99.832	100	1.022	99.458	0.877

Table 41
Gender classification accuracy (%) on DiveFace testset on ViT architecture across different folds, and gender-racial groups on a continual learning framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from Fold *i* identified using BPCOD.

Race	East Asian		Sub Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Finetuned Model 1									
Fold 2	98.127	99.3	98.321	98.7	98.724	99.5	1.014	98.779	0.536
Finetuned Model 2									
Fold 3	98.5	100	98	99.105	99.57	98.74	1.02	98.986	0.728
Finetuned Model 3									
Fold 4	99.13	99.233	97.3	98.74	99.132	98.84	1.02	98.729	0.725

Table 42
Gender classification accuracy (%) on DiveFace testset on ViT architecture across different folds, and gender-racial groups on a continual learning framework. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers from Fold *i* identified using EBOD.

Race	East Asian		Sub Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Finetuned Model 1									
Fold 2	99.56	99.3	98.85	98.7	99.71	99.5	1.01	99.27	0.408
Finetuned Model 2									
Fold 3	99.764	100	99.24	99.231	98.2	99.13	1.018	99.261	0.623
Finetuned Model 3									
Fold 4	98.5	98.46	98.54	99.279	99.132	98.658	1	98.761	0.353

Table 43
Gender classification accuracy (%) on DiveFace testset on ViT architecture across different folds and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using BPCOD without the prior preservation loss.

Race	East Asian		Sub Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Finetuned Model 1									
Fold 2	90.622	99	89.4	98.7	91.6744	99.6	1.114	94.832	4.739
Finetuned Model 2									
Fold 3	88.89	99.496	92.759	98.356	90.973	98.92	1.119	94.9	4.59
Finetuned Model 3									
Fold 4	84.943	96.34	84.262	94.711	85.869	97.656	1.16	90.63	6.232

Table 44
Gender classification accuracy (%) on DiveFace testset on ViT architecture across different folds and gender-racial groups. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and race; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and race. Finetuned Model *i* is the fine-tuned model with the outliers identified from Fold *i* using EBOD without the prior preservation loss.

Race	East Asian		Sub Saharan & South Indian		White		Max/Min↓	Overall↑	DoB↓
Gender	M	F	M	F	M	F			
Finetuned Model 1									
Fold 2	90.247	99.11	90.407	95.61	92.759	94.32	1.098	93.74	3.374
Finetuned Model 2									
Fold 3	78.134	99.665	88.501	98.16	82.12	99.105	1.276	90.948	9.41
Finetuned Model 3									
Fold 4	74.96	98.651	82.092	96.543	81.505	99.345	1.325	88.85	10.563

References

Abdurrahim, S. H., Samad, S. A., & Huddin, A. B. (2018). Review on the effects of age, gender, and race demographics on automatic face recognition. *Visual Computer*, *34*(11), 1617–1630. http://dx.doi.org/10.1007/s00371-017-1428-z.

Albiero, V., S, K. K., Vangara, K., Zhang, K., King, M. C., & Bowyer, K. W. (2020). Analysis of gender inequality in face recognition accuracy. In *IEEE WACV workshops* (pp. 81–89). IEEE, http://dx.doi.org/10.1109/WACVW50321.2020.9096947.

Albiero, V., Zhang, K., King, M. C., & Bowyer, K. W. (2022). Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, *17*, 127–137. http://dx.doi.org/10.1109/TIFS.2021.3135750.

Almadan, A., Krishnan, A., & Rattani, A. (2020). Bwcface: Open-set face recognition using body-worn camera. In M. A. Wani, F. Luo, X. A. Li, D. Dou, & F. Bonchi (Eds.), *19th IEEE international conference on machine learning and applications, ICMLA 2020, miami, FL, USA, December 14-17, 2020* (pp. 1036–1043). IEEE, http://dx.doi.org/10.1109/ICMLA51294.2020.00168.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, *35*(4), 105–120. http://dx.doi.org/10.1609/aimag.v35i4.2513.

Amini, A., Schwarting, W., Soleimany, A., & Rus, D. (2020). Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, neurIPS 2020, December 6-12, 2020, virtual*.

Attenberg, J., Ipeirotis, P., & Provost, F. J. (2015). Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *ACM Journal of Data and Information Quality*, *6*(1), 1:1–1:17. http://dx.doi.org/10.1145/2700832.

Barlas, P., Kyriakou, K., Guest, O., Kleanthous, S., & Otterbacher, J. (2020). To "see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–31. http://dx.doi.org/10.1145/3432931.

Becker, D., Kenrick, D., Neuberg, S., Blackwell, K., & Smith, D. (2007). The confounded nature of angry men and happy women. *Journal of Personality and Social Psychology*, *92*, 179–190. http://dx.doi.org/10.1037/0022-3514.92.2.179.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. http://dx.doi.org/10.1177/0049124118782533, arXiv:https://doi.org/10.1177/0049124118782533.

Best-Rowden, L., & Jain, A. K. (2018). Longitudinal study of automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(1), 148–162. http://dx.doi.org/10.1109/TPAMI.2017.2652466.

Bostan, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics, COLING 2018, santa fe, new Mexico, USA, August 20-26, 2018* (pp. 2104–2119). Association for Computational Linguistics.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler, & C. Wilson (Eds.), *Proceedings*

of machine learning research: vol. 81, Conference on fairness, accountability and transparency, FAT 2018, 23-24 February 2018, new york, NY, USA (pp. 77–91). PMLR.

Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In A. Roychoudhury, C. Cadar, & M. Kim (Eds.), *Proceedings of the 30th ACM joint European software engineering conference and symposium on the foundations of software engineering, ESEC/FSE 2022, Singapore, Singapore, November 14-18, 2022* (pp. 1122–1134). ACM, http://dx.doi.org/10.1145/3540250.3549093.

Chiu, C., Chung, H., Chen, Y., Shi, Y., & Ho, T. (2023). Fair multi-exit framework for facial attribute classification. http://dx.doi.org/10.48550/arXiv.2301.02989, arXiv:2301.02989.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. http://dx.doi.org/10.1089/big.2016.0047.

Chuang, C., & Mroueh, Y. (2021). Fair mixup: Fairness via interpolation. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021* (p. 15). OpenReview.net.

Correia, A. H. C., & Lécué, F. (2019). Human-in-the-loop feature selection. In *The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, honolulu, hawaii, USA, January 27 - February 1, 2019* (pp. 2438–2445). AAAI Press, http://dx.doi.org/10.1609/aaai.v33i01.33012438.

Das, A., Dantcheva, A., & Brémond, F. (2018). Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In L. Leal-Taixé, & S. Roth (Eds.), *Lecture notes in computer science: vol. 11129, Computer vision - ECCV 2018 workshops - munich, Germany, September 8-14, 2018, proceedings, part i* (pp. 573–585). Springer, http://dx.doi.org/10.1007/978-3-030-11009-3_35.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A. S., Nemade, G., & Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, online, July 5-10, 2020* (pp. 4040–4054). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.ACL-MAIN.372.

Denton, E., Hutchinson, B., Mitchell, M., & Gebru, T. (2019). Detecting bias with generative counterfactual face attribute augmentation. arXiv:1906.06439.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021* (p. 22). OpenReview.net.

FaceX (2022). FaceX face API.

Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th international conference on learning representations, ICLR 2019, new orleans, la, USA, May 6-9, 2019*. OpenReview.net.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. http://dx.doi.org/10.1016/S1364-6613(99)01294-2.

Georgopoulos, M., Oldfield, J., Nicolaou, M. A., Panagakis, Y., & Pantic, M. (2021). Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, *129*(7), 2288–2307. http://dx.doi.org/10.1007/s11263-021-01448-w.

Gildenblat, J., & contributors (2021). Pytorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam.

Grother, P., Quinn, G. W., & Phillips, P. J. (2011). Report on the evaluation of 2D still-image face recognition algorithms. In *NIST report* (p. 61).

Han, L., Dong, X., & Demartini, G. (2021). Iterative human-in-the-loop discovery of unknown unknowns in image datasets. In E. Kamar, & K. Luther (Eds.), *Proceedings of the ninth AAAI conference on human computation and crowdsourcing, HCOMP 2021, virtual, November 14-18, 2021* (pp. 72–83). AAAI Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition, CVPR* (pp. 770–778). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2016.90.

Huang, G. B., Mattar, M. A., Lee, H., & Learned-Miller, E. G. (2012). Learning to align from scratch. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012. proceedings of a meeting held December 3-6, 2012, lake tahoe, nevada, United states* (pp. 773–781).

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments: Technical Report 07–49*, University of Massachusetts, Amherst.

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819. http://dx.doi.org/10.1109/TAI.2022.3194503.

Joo, J., & Kärkkäinen, K. (2020). Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *FATE/MM '20, Proceedings of the 2nd international workshop on fairness, accountability, transparency and ethics in multimedia* (pp. 1–5). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3422841.3423533.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In P. A. Flach, T. D. Bie, & N. Cristianini (Eds.), *Lecture notes in computer science: vol. 7524, Machine learning and knowledge discovery in databases - European conference, ECML PKDD 2012, bristol, UK, September 24-28, 2012. proceedings, part II* (pp. 35–50). Springer, http://dx.doi.org/10.1007/978-3-642-33486-3_3.

Kärkkäinen, K., & Joo, J. (2021). FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE winter conference on applications of computer vision, WACV 2021, waikoloa, HI, USA, January 3-8, 2021* (pp. 1547–1557). IEEE, http://dx.doi.org/10.1109/WACV48630.2021.00159.

Kiruthika, S., & Masilamani, V. (2021). Retinal image quality assessment using sharpness and connected components. In B. Raman, S. Murala, A. S. Chowdhury, A. Dhall, & P. Goyal (Eds.), *Communications in computer and information science: vol. 1568, Computer vision and image processing - 6th international conference, CVIP 2021, rupnagar, India, December 3-5, 2021, revised selected papers, part II* (pp. 181–191). Springer, http://dx.doi.org/10.1007/978-3-031-11349-9_16.

Klare, B., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801. http://dx.doi.org/10.1109/TIFS.2012.2214212.

Klie, J., Webber, B., & Gurevych, I. (2023). Annotation error detection: Analyzing the past and present for a more coherent future. *Computational linguistics*, *49*(1), 157–198. http://dx.doi.org/10.1162/COLI_A_00464.

Kong, F., Yuan, S., Hao, W., & Henao, R. (2023). Mitigating test-time bias for fair image retrieval. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems 36: annual conference on neural information processing systems 2023, neurIPS 2023, new orleans, la, USA, December 10 - 16, 2023*.

Krishnan, A., Almadan, A., & Rattani, A. (2020a). Probing fairness of mobile ocular biometrics methods across gender on VISOB 2.0 dataset. In A. D. Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, & R. Vezzani (Eds.), *Lecture notes in computer science: vol. 12668, Pattern recognition. ICPR international workshops and challenges - virtual event, January 10-15, 2021, proceedings, part VIII* (pp. 229–243). Springer, http://dx.doi.org/10.1007/978-3-030-68793-9_16.

Krishnan, A., Almadan, A., & Rattani, A. (2020b). Understanding fairness of gender classification algorithms across gender-race groups. In *19th IEEE international conference on machine learning and applications, ICMLA* (pp. 1028–1035). IEEE, http://dx.doi.org/10.1109/ICMLA51294.2020.00167.

Krishnan, A., Almadan, A., & Rattani, A. (2021). Investigating fairness of ocular biometrics among Young, middle-aged, and older adults. In *2021 international carnahan conference on security technology, ICCST 2021, hatfield, United kingdom, October 11-15, 2021* (pp. 1–7). IEEE, http://dx.doi.org/10.1109/ICCST49569.2021.9717383.

Krishnan, A., Neas, B., & Rattani, A. (2022). Is facial recognition biased at near-infrared spectrum as well? In *2022 IEEE international symposium on technologies for homeland security* (pp. 1–7). http://dx.doi.org/10.1109/HST56032.2022.10025433.

Krishnan, A., & Rattani, A. (2023). A novel approach for bias mitigation of gender classification algorithms using consistency regularization. *Image and Vision Computing*, *137*, Article 104793. http://dx.doi.org/10.1016/j.imavis.2023.104793.

Kuzmin, A., Nagel, M., van Baalen, M., Behboodi, A., & Blankevoort, T. (2023). Pruning vs quantization: Which is better? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems 36: annual conference on neural information processing systems 2023, neurIPS 2023, new orleans, la, USA, December 10 - 16, 2023*.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *IEEE conf. on computer vision and pattern recognition* (pp. 34–42). Boston, MA.

Li, Z., & Xu, C. (2021). Discover the unknown biased attribute of an image classifier. In *2021 IEEE/CVF international conference on computer vision, ICCV 2021, montreal, QC, Canada, October 10-17, 2021* (pp. 14950–14959). IEEE, http://dx.doi.org/10.1109/ICCV48922.2021.01470.

Lin, X., Kim, S., & Joo, J. (2022). Fairgrape: Fairness-aware gradient pruning method for face attribute classification. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Lecture notes in computer science: vol. 13673, Computer vision - ECCV 2022 - 17th European conference, tel aviv, Israel, October 23-27, 2022, proceedings, part XIII* (pp. 414–432). Springer, http://dx.doi.org/10.1007/978-3-031-19778-9_24.

Lingenfelter, B., Davis, S. R., & Hand, E. M. (2022). A quantitative analysis of labeling issues in the celeba dataset. In G. Bebis, B. Li, A. Yao, Y. Liu, Y. Duan, M. Lau, R. Khadka, A. Crisan, & R. Chang (Eds.), *Lecture notes in computer science: vol. 13598, Advances in visual computing - 17th international symposium, ISVC 2022, san diego, CA, USA, October 3-5, 2022, proceedings, part i* (pp. 129–141). Springer, http://dx.doi.org/10.1007/978-3-031-20713-6_10.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *2015 IEEE international conference on computer vision, ICCV 2015, santiago, Chile, December 7-13, 2015* (pp. 3730–3738). IEEE Computer Society, http://dx.doi.org/10.1109/ICCV.2015.425.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2019,*

brighton, United kingdom, May 12-17, 2019 (pp. 2847–2851). IEEE, http://dx.doi.org/10.1109/ICASSP.2019.8682620.

Majumdar, P., Singh, R., & Vatsa, M. (2021). Attention aware debiasing for unbiased model prediction. In *IEEE/CVF international conference on computer vision workshops, ICCVW 2021, montreal, BC, Canada, October 11-17, 2021* (pp. 4116–4124). IEEE, http://dx.doi.org/10.1109/ICCVW54120.2021.00459.

Malmström, M., Skog, I., Axehill, D., & Gustafsson, F. (2022). Detection of outliers in classification by using quantified uncertainty in neural networks. In *2022 25th international conference on information fusion* (pp. 1–7). http://dx.doi.org/10.23919/FUSION49751.2022.9841376.

Marcinkevics, R., Ozkan, E., & Vogt, J. E. (2022). Debiasing deep chest X-Ray classifiers using intra- and post-processing methods. In Z. C. Lipton, R. Ranganath, M. P. Sendak, M. W. Sjoding, & S. Yeung (Eds.), *Proceedings of machine learning research: vol. 182, Proceedings of the machine learning for healthcare conference, MLHC 2022, 5-6 August 2022, durham, NC, USA* (pp. 504–536). PMLR.

Masood, S., Gupta, S., Wajid, A., Gupta, S., & Ahmed, M. (2018). Prediction of human ethnicity from facial images using neural networks. In S. C. Satapathy, V. Bhateja, K. S. Raju, & B. Janakiramaiah (Eds.), *Data engineering and intelligent computing* (pp. 217–226). Singapore: Springer Singapore.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *Psychology of Learning and Motivation: vol. 24, Title* (pp. 109–165). Academic Press, http://dx.doi.org/10.1016/S0079-7421(08)60536-8.

Monarch, R. M., & Manning, C. D. (2021). *Human-In-The-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Manning Publications.

Morales, A., Fiérrez, J., Vera-Rodríguez, R., & Tolosana, R. (2021). SensitiveNets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(6), 2158–2164. http://dx.doi.org/10.1109/TPAMI.2020.3015420.

Muthukumar, V. (2019). Color-theoretic experiments to understand unequal gender classification accuracy from face images. In *IEEE conference on computer vision and pattern recognition workshops, CVPR workshops 2019, long beach, CA, USA, June 16-20, 2019* (pp. 2286–2295). Computer Vision Foundation / IEEE, http://dx.doi.org/10.1109/CVPRW.2019.00282.

Nadimpalli, A. V., & Rattani, A. (2022). GBDF: gender balanced DeepFake dataset towards fair DeepFake detection. http://dx.doi.org/10.48550/arXiv.2207.10246, CoRR arXiv:2207.10246.

Ouyang, B., Song, Y., Li, Y., Sant, G., & Bauchy, M. (2021). EBOD: An ensemble-based outlier detection algorithm for noisy datasets. *Knowledge-Based Systems*, *231*, Article 107400. http://dx.doi.org/10.1016/j.knosys.2021.107400.

Palma, G. D., Kiani, B. T., & Lloyd, S. (2019). Random deep neural networks are biased towards simple functions. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, neurIPS 2019, December 8-14, 2019, vancouver, BC, Canada* (pp. 1962–1974).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. http://dx.doi.org/10.1109/TKDE.2009.191.

Park, S., Hwang, S., Kim, D., & Byun, H. (2021). Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence, EAAI 2021, virtual event, February 2-9, 2021* (pp. 2403–2411). AAAI Press.

Prabhu, V. U., Yap, D. A., Wang, A., & Whaley, J. (2019). Covering up bias in celeba-like datasets with Markov blankets: A post-hoc cure for attribute prior avoidance. CoRR arXiv:1907.12917.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 429–435). ACM, http://dx.doi.org/10.1145/3306618.3314244.

Ramachandran, S., & Rattani, A. (2023). Deep generative views to mitigate gender classification bias across gender-race groups. In J.-J. Rousseau, & B. Kapralos (Eds.), *Pattern recognition, computer vision, and image processing. ICPR 2022 international workshops and challenges* (pp. 551–569). Cham: Springer Nature Switzerland.

Ramaswamy, V. V., Kim, S. S. Y., & Russakovsky, O. (2021). Fair attribute classification through latent space de-biasing. In *IEEE conference on computer vision and pattern recognition, CVPR 2021, virtual, June 19-25, 2021* (pp. 9301–9310). Computer Vision Foundation / IEEE, http://dx.doi.org/10.1109/CVPR46437.2021.00918.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*(2), 285–308. http://dx.doi.org/10.1037/0033-295x.97.2.285.

Rattani, A., Derakhshani, R., & Ross, A. (2019). Introduction to selfie biometrics. In A. Rattani, R. Derakhshani, & A. Ross (Eds.), *Advances in computer vision and pattern recognition, Selfie biometrics - advances and challenges* (pp. 1–18). Springer, http://dx.doi.org/10.1007/978-3-030-26972-2_1.

Reiter, R. (1977). On closed world data bases. In H. Gallaire, & J. Minker (Eds.), *Advances in data base theory, Logic and Data Bases, Symposium on Logic and Data Bases, Centre D'ÉTudes Et de Recherches de Toulouse, France, 1977* (pp. 55–76). New York: Plemum Press, http://dx.doi.org/10.1007/978-1-4684-3384-5_3.

Rekognition, A. (2022). Amazon Rekognition face API.

Russakovsky, O., Li, L., & Fei-Fei, L. (2015). Best of both worlds: Human-machine collaboration for object annotation. In *IEEE conference on computer vision and pattern recognition, CVPR 2015, boston, MA, USA, June 7-12, 2015* (pp. 2121–2131). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2015.7298824.

Salim, N. R., Sankaranarayanan, N., & Jayaraman, U. (2021). Gender classification beyond visible spectrum using shallow convolution neural network. In *2021 IEEE madras section conference* (pp. 1–7). http://dx.doi.org/10.1109/MASCON51689.2021.9563425.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE international conference on computer vision, ICCV 2017, venice, Italy, October 22-29, 2017* (pp. 618–626). IEEE Computer Society, http://dx.doi.org/10.1109/ICCV.2017.74.

Sensoy, M., Kaplan, L. M., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, neurIPS 2018, December 3-8, 2018, montréAl, Canada* (pp. 3183–3193).

Services, M. A. C. (2022). Microsoft Azure Cognitive Services face API.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Siddiqui, H., Rattani, A., Ricanek, K., & Hill, T. J. (2022). An examination of bias of facial analysis based BMI prediction models. In *IEEE/CVF conference on computer vision and pattern recognition workshops, CVPR workshops 2022, new orleans, la, USA, June 19-20, 2022* (pp. 2925–2934). IEEE, http://dx.doi.org/10.1109/CVPRW56347.2022.00330.

Singh, R., Majumdar, P., Mittal, S., & Vatsa, M. (2022). Anatomizing bias in facial analysis. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, February 22 - March 1, 2022* (pp. 12351–12358). AAAI Press.

Steephen, J., Mehta, S., & Surampudi, B. (2017). Do we expect women to look happier than they are? A test of gender-dependent perceptual correction. *Perception*, *47*, Article 0301006617745240. http://dx.doi.org/10.1177/0301006617745240.

Tapia, J. E., Perez, C. A., & Bowyer, K. W. (2016). Gender classification from the same iris code used for recognition. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1760–1770. http://dx.doi.org/10.1109/TIFS.2016.2550418.

Vera-Rodríguez, R., Blázquez, M., Morales, A., Gonzalez-Sosa, E., Neves, J. C., & Proença, H. (2019). FaceGenderID: Exploiting gender information in DCNNs face recognition systems. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 2254–2260). Computer Vision Foundation / IEEE, http://dx.doi.org/10.1109/CVPRW.2019.00278.

Villa, M., Gofman, M. I., Mitra, S., Almadan, A., Krishnan, A., & Rattani, A. (2020). A survey of biometric and machine learning methods for tracking students' attention and engagement. In M. A. Wani, F. Luo, X. A. Li, D. Dou, & F. Bonchi (Eds.), *19th IEEE international conference on machine learning and applications, ICMLA 2020, miami, FL, USA, December 14-17, 2020* (pp. 948–955). IEEE, http://dx.doi.org/10.1109/ICMLA51294.2020.00154.

Vision, D. (2022). Deep Vision face API.

Wang, G. (2019). Humans in the loop: The design of interactive AI systems.

Wang, Z., Dong, X., Xue, H., Zhang, Z., Chiu, W., Wei, T., et al. (2022). Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, new orleans, la, USA, June 18-24, 2022* (pp. 10369–10378). IEEE, http://dx.doi.org/10.1109/CVPR52688.2022.01013.

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., et al. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, seattle, WA, USA, June 13-19, 2020* (pp. 8916–8925). Computer Vision Foundation / IEEE.

Wick, M. L., Panda, S., & Tristan, J. (2019). Unlocking fairness: a trade-off revisited. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, neurIPS 2019, December 8-14, 2019, vancouver, BC, Canada* (pp. 8780–8789).

Yao, A., Gall, J., Leistner, C., & Gool, L. V. (2012). Interactive object detection. In *2012 IEEE conference on computer vision and pattern recognition, providence, RI, USA, June 16-21, 2012* (pp. 3242–3249). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2012.6248060.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27: annual conference on neural information processing systems 2014, December 8-13 2014, montreal, quebec, Canada* (pp. 3320–3328).

Zhang, K., Gao, C., Guo, L., Sun, M., Yuan, X., Han, T. X., et al. (2017). Age group and gender estimation in the wild with deep ror architecture. *IEEE Access*, *5*, 22492–22503. http://dx.doi.org/10.1109/ACCESS.2017.2761849.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In J. Furman, G. E. Marchant, H. Price, & F. Rossi (Eds.), *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society, AIES 2018, new orleans, la, USA, February 02-03, 2018* (pp. 335–340). ACM, http://dx.doi.org/10.1145/3278721.3278779.

Zhang, Zhifei, Song, Yang, & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *IEEE conference on computer vision and pattern recognition* (p. 9). IEEE.