Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition

Nikhil Churamani, Ozgur Kara, and Hatice Gunes

Abstract—As Facial Expression Recognition (FER) systems become integrated into our daily lives, these systems need to prioritise making *fair* decisions instead of only aiming at higher individual accuracy scores. From surveillance systems, to monitoring the mental and emotional health of individuals, these systems need to balance the *accuracy vs fairness* trade-off to make decisions that do not unjustly discriminate against specific under-represented demographic groups. Identifying *bias* as a critical problem in facial analysis systems, different methods have been proposed that aim to mitigate bias both at data and algorithmic levels. In this work, we propose the novel use of Continual Learning (CL), in particular, using Domain-Incremental Learning (Domain-IL) settings, as a potent bias mitigation method to enhance the *fairness* of FER systems. We compare different non-CL-based and CL-based methods for their *performance* and *fairness scores* on expression recognition and Action Unit (AU) detection tasks using two popular benchmarks, the RAF-DB and BP4D datasets, respectively. Our experimental results show that CL-based methods, on average, outperform other popular bias mitigation techniques on both *accuracy* and *fairness* metrics.

Index Terms—Fairness, Continual Learning, Bias Mitigation, Affective Computing, Facial Expression Recognition, Facial Action Units.

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) systems are increasingly becoming an integral part of human life, monitoring and controlling several aspects of our daily lives with little to no human oversight. From security and surveillance systems that deploy several ML models such as face detection and recognition systems [1], social media platforms that *auto-tag* pictures of our friends and family [2], recommender systems that track our digital footprints to advertise products that we might like to indulge in [3], to banking and finance applications that work on credit approvals based on socio-economic backgrounds, AI systems are ubiquitous, making 'smart' decisions about several critical aspects of our lives [4], [5]. It is thus important to ensure that these systems make fair and unbiased decisions to avoid potentially catastrophic consequences that affect individuals [6]. In this work, we focus on one such application of AI in real life; facial affect analysis systems.

Facial affect analysis systems (see [7]–[9] for a survey) aim to analyse facial expressions either by encoding facial muscle activity as Facial Action Units (AUs) [10] or determining individual expressions [11], [12]. Analysing large datasets of human faces, annotated for facial expressions, these models are heavily data-dependent and may be prone to *biases* originating from imbalances in the training data distribution. For a large variety of Facial Expression Recognition (FER) datasets, attributes such as gender, race, age or skin colour are implicitly encoded in the data which may also be learnt by a (deep) learning model [13]. If these attributes are not balanced across the entire distribution of the dataset, the model may learn to associate such *confounding* attributes with the task of FER. For example, if the training data has a disproportionate number of images of *Males* expressing '*Happy*' than *Females*, the model may learn to associate gender with the expression, leading to a lot of 'happy female' samples being potentially misclassified.

While the most effective method for preventing biases in FER datasets would be to ensure a balanced and representative data collection, this also turns out to be the most challenging problem. Owing to restrictions with respect to data recording settings, personal preferences, geographic location as well as several social and cultural constraints, it may not always be possible to ensure a balanced data collection. Most recent datasets try to ensure the data collection is fair and unbiased or at the least provide demographic annotations, along with affective labels, that enable researchers to make informed decisions while using these datasets for training ML models [13]. Yet, to ensure fairness despite the inherent imbalances in data distributions, several methods have been proposed that handle these imbalances at the *pre-processing*, *in-processing* or *post-processing* levels [14], [15].

Pre-processing methods focus on *strategically sampling* training data, that is, given the distribution of data with respect to a selected demographic attribute, samples belonging to under-represented groups are either over-sampled compared to dominant groups [16], or scaled penalties are applied when a model incorrectly classifies these samples [17]. Yet, these methods are not perfect, and some bias might still 'creep in'. To handle this, changes to the

- *N. Churamani and H. Gunes are with the Department of Computer Science and Technology, University of Cambridge, United Kingdom.*
  E-mail: {nikhil.churamani, hatice.gunes}@cl.cam.ac.uk
- *O. Kara is with the Electrical & Electronics Engineering Department at the Bogazici University, Istanbul, Turkey.*
  E-mail: ozgur.kara@boun.edu.tr

model architecture or the training regime need to be made. Algorithmic or *In-Processing* methods achieve this either by explicitly learning domain-specific information and discounting it from the model's learning [18] or by learning to completely ignoring domain-specific information by omitting these features from the learnt representations [19]. Post-processing methods, on the other hand, are mostly used to quantify bias in trained algorithms [14] and offer effective tools to evaluate the *fairness* of an ML model.

Interestingly, the underpinning principle behind all the above-mentioned methods is essentially to focus on learning and adapting to the inherent imbalances in data distribution, either by *synthetically* balancing it or adjusting the learning algorithm itself to account for these imbalances. This principle is shared by Continual Learning (CL) methods [20]–[22] that aim to balance learning in the model by being sensitive to shifts in data distributions, ensuring that one particular domain or task does not dominate the model's learning. Their ability to *continually* learn and adapt to novel information, aggregating new knowledge without impacting previously acquired information, may allow them to balance learning across the different learning domains. Domain-Incremental CL (Domain-IL) settings [23] particularly focus on managing shifts in input data distribution while the task remains the same. This can be considered analogous to solving FER tasks where input data belongs to different domains of gender (male, female) and race (black, white, asian, latino). The challenge for CL models will thus be to maintain FER performance with respect to one domain while acquiring information about new domains.

Motivated by this notion, we propose the novel use of CL as a learning paradigm, well-suited for developing *fairer* FER models that balance learning with respect to different attributes of gender and race. We formulate expression recognition and AU detection across these domain attributes as CL problems and compare several popular CL approaches with state-of-the-art bias mitigation methods. To the best of our knowledge, this is the first application of CL as a bias mitigation strategy for facial affect analyses. We explore Domain-IL where the models learn FER tasks across different domains, defined by the demographic attributes of *gender* and *race*. For each attribute, the data is split into different domains; gender into *male* and *female* domains, while race into *White/Caucasian, Black/African-American, Asian* and *Latino* domains. We primarily focus on regularisation-based CL methods as these do not require setting up additional memory or computational resources, allowing a fair and direct comparison with other bias-mitigation methods. Our experimental results show that CL-based approaches, on average, outperform other bias mitigation strategies, both in terms of *accuracy* as well as fairness scores for both FER and AU detection tasks across the domain splits.

2 Background and Related Work

2.1 Understanding Bias

Bias, both in human perception and behaviour as well as ML algorithms can be characterised as an inclination or prejudice towards a person or a group, that may be considered unfair. This may result from an over or under-exposure towards a certain group of individuals usually characterised by their gender, racial identity, social or economic background or age, amongst other factors. This exposure results in people considering individuals that share similar characteristics as themselves as 'in-group' members and others different from them as 'out-group' members [24]. People tend to be *biased* in favour of in-group members, evaluating them more positively on dimensions of judgement while being *biased* or *prejudiced* against out-group members [25].

Understanding how humans consider in-group and out-group members [24] in their immediate surroundings and base their decisions on aspects such as gender, race or age is important to view ML models in the right perspective when applied to real-world settings. Such an understanding will allow researchers to assess what may be considered 'fair' and how to achieve such fairness in algorithms.

2.1.1 Bias in Human Perception

Following the Perception-Action model of Empathy [26], an individual's behaviour, particularly their facial expressions and body gestures stimulate a similar neural activation in the observer, enabling them to empathise with and understand their actions, intentions and emotions. Gutsell et al. [27], through a series of experiments with participants (30 white university students) interacting with in-group (in this case, participants with a *Caucasian* ethnic identity) and out-group (excluded from this circle; in this case, *African–Canadian, East-Asian, South-Asian* ethnic identities) members, concluded that such perception-action couplings are reserved only for in-group members. In-group identification, that is, identifying other individuals to be sharing similar characteristics as oneself, causes a *positive association* with them [25]. Furthermore, people have a harder time recognising the faces of out-group members and interpreting their facial expressions [28], [29].

In ML algorithms, we may understand such 'inter-group bias' to result from imbalances in data distributions where certain groups may be considered to constitute 'in-group' attributes due to their dominance in the data, while under-represented attributes may be considered 'out-group'. Thus, having seen a lot of samples from certain groups, models may be more capable of correctly classifying such samples while performing poorly for the 'out-group' samples.

2.1.2 Bias in Machine Learning

Owing to similar reasons as in the case of human perception, over or under-exposure to experiences (or data) characterised by specific features, ML models can acquire biases that prejudice model performance for one or more data attributes. Facial analysis models may be affected by biases with respect to demographic attributes of gender, race or age, where samples belonging to one group dominate the data distribution. In such situations, the under-represented groups get adversely impacted by the model misclassifying samples from these groups. Buolamwini et al. in their seminal work [30], highlighted how popular face recognition algorithms disproportionately misclassified darker females, either misgendering them or not being able to detect their faces. Klare et al. [31] highlighted how face recognition algorithms employed by some law-enforcement agencies significantly under-perform for people labelled as black or female compared to other demographics. Such biases in critical systems may lead to unnecessary targeting and exploitation of people from under-represented groups, further disadvantaging their opportunities in society.

2.2 Mitigating Bias in Facial Analyses

The origins of bias in most ML-based facial analyses algorithms can be traced back to imbalances in data distributions. Collating balanced datasets that enable a fair evaluation of ML models [32] despite being the most effective solution for mitigating biases, may not be as straightforward to achieve as a varied and diverse subject-pool may not always be available. Thus, several strategies have been proposed for mitigating the effects of bias in ML algorithms. We use a similar nomenclature as [14] to discuss these strategies.

2.2.1 Pre-Processing Approaches

A simplistic strategy can be selectively sampling training data in a manner that balances learning. Samples from under-represented domains are over-sampled while dominant domains are under-sampled to balance learning the training data [19], [33]. This results in the training set to effectively have a balanced distribution. However, this may not be possible in small-scale datasets as under-sampling already limited data might not be efficient. An alternative approach is to use data-augmentation techniques to synthetically generate additional data for the under-represented groups [34]–[36], to balance training data distribution.

2.2.2 In-Processing Approaches

Another popular approach to mitigate the effects of imbalances in data distributions is to weight the model prediction loss differently for different domain attributes. A weighting factor is applied to the training loss based on the occurrence rate of the different classes or domains [16], [17], [37] penalising misclassifications for the under-represented groups more than others. This reduces the effect of these imbalances, mitigating biases in learning.

More recently, several learning strategies have been proposed that, while handing imbalances in data distributions using the above-mentioned techniques, also deal with biases in ML models at the algorithm-level. Howard et al. [5] propose a hierarchical approach that combines outputs from the cloud-based Microsoft Emotion API with a specialised learner, offering a 17.3% improvement in recognition results on a minority domain, in this case, children's facial expressions. Other approaches focus on explicitly separating decision boundaries with respect to sensitive domain attributes ensuring that imbalances in data are not perpetuated while training the model, achieving *'fairness through awareness'* [18]. Alternatively, the model can be trained to ignore domain-specific information, making it *unaware* or *blind* towards domain differences and focus only on the task at hand. Adversarial learning has been used to achieve such *'fairness through blindness'* [19] using a min-max training regime that maximises sensitivity towards the task at hand while minimising learning of domain-specific information. Xu et al. [38] implement a disentangled approach [39] that uses a similar strategy to mitigate bias with respect to sensitive domain attributes of gender and race for FER by ensuring that the feature representations learnt by the model do not contain any domain-specific information. The model is split into two parts with a shared feature extraction subnetwork. The first part focuses on facial expression analysis, while the other part consists of separate branches for each domain, designed to suppress domain-specific information.

2.2.3 Post-Processing Approaches

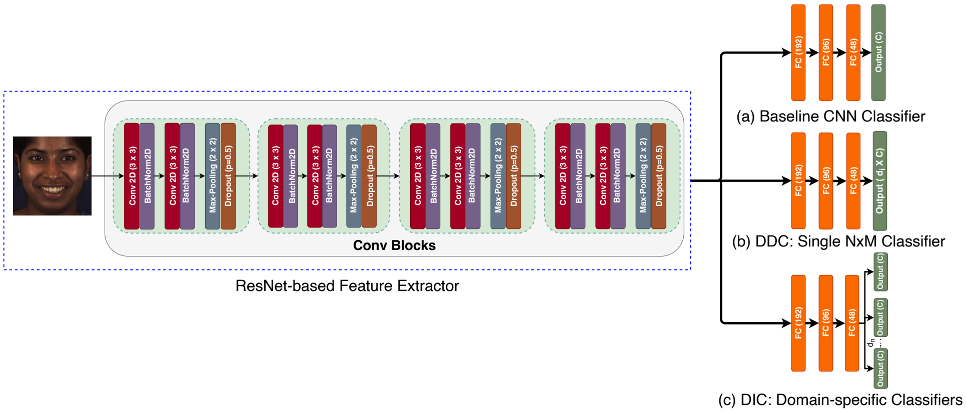
Despite several methods proposed for training *fair* ML systems, it may not always be possible to completely eradicate bias from the model. In such cases, it is still important to quantify the bias to mitigate it and make fairer decisions. Post-processing approaches (see [14], [40] for a general discussion) focus on quantifying bias in existing algorithms and attempt to counter the effects on classification tasks.

2.3 Continual Learning

Learning to detect and manage shifts in data distributions, Continual Learning (CL) methods (see [20]–[22]) can effectively learn with incrementally acquired data, offering an improvement over traditional ML models, especially for real-world application. Typically, CL models are evaluated on 3 different learning scenarios [23]. The first scenario is *Task-Incremental Learning (Task-IL)* where models incrementally learn to solve tasks, explicitly being informed about the task identity. Learning is split into different tasks, each corresponding to learning some sub-tasks or classes. Models are evaluated on their ability to preserve knowledge across the tasks. The second scenario focuses on *Domain-Incremental Learning (Domain-IL)* where the task to be learnt does not change but the input data distribution changes and models are evaluated on their ability to manage such shifts. The third learning scenario is *Class-Incremental Learning (Class-IL)* where models need to learn one class at a time, sequentially receiving input data for only that class.

In recent years, several CL approaches have been proposed that employ (deep) ML architectures and equip them with learning capabilities such that they can incrementally integrate novel information while preserving past knowledge [20]. The most common and straightforward approach to achieve this is by regulating model updates in a manner that enables the preservation of knowledge. Such *regularisation-based* methods minimise *destructive interference* by freezing those parts of the model that are most sensitive to previous tasks [41] and updating the rest of the model, selectively. Alternatively, weight-update constraints and penalties are applied that discourage changes in network parameters that deteriorate model performance on previous tasks [42], [43]. A priority or importance term may also be applied to network parameters based on their relevance to a given task and only those parameters are allowed to be updated which have lower importance [44]. Despite the competitive performance of regularisation-based methods, they become computationally expensive as the number of tasks or classes grow, limiting their performance and applicability in Task-IL (in extreme cases) and Class-IL scenarios.

Other CL-based approaches include rehearsal-based methods [45] that aim to simulate offline batch-learning based settings by either physically storing previously encountered data samples or learning a generative or probabilistic model that learns data statistics to simulate *pseudo-*



ResNet-based Feature Extractor

(a) Baseline CNN Classifier

(b) DDC: Single NxM Classifier

(c) DIC: Domain-specific Classifiers

Fig. 1. Model Architectures for (a) the Baseline CNN, (b) Domain Discriminative Classification (DDC) [19] with an $N \times M$ classifier where N is the number of *domains* and M is the number of *classes* within each domain, (c) Domain Independent Classification (DIC) [19] with N independent classifiers with M classes each. The Baseline CNN is also used to implement all CL methods for a fair comparison.

samples for previously seen tasks [46]–[48]. Yet, as the number of tasks increase, it becomes extremely difficult to train these models. Furthermore, additional memory and computational resources need to be allocated to either store data samples or generate simulated pseudo-samples making it challenging to implement these approaches.

In this work, we focus on regularisation-based methods evaluated under Domain-IL settings where these models are required to learn to solve expression recognition and AU detection tasks across different domains of gender and race.

3 Methodology

In order to understand bias in FER models, it is important to determine how the implicit data distribution affects model performance. For this, we need to understand which domain attributes dominate the data and how an algorithm performs with respect to these attributes. In this section, we present the problem formulation, the learning scenario as well as the different methods employed in this work, comparing them with popular CL-based methods.

3.1 Problem Formulation

We aim to measure the variance in model performance on a specific task with respect to *gender* and *race* as domain attributes and compare model performances for expression recognition and AU detection. Given a set of input images x_i with task labels y_i and domain label d_i, we wish to determine how the performance of an algorithm $\mathcal{A}(y_i|x_i, d_i)$, varies with respect to different domain labels d_i.

To enable a fair comparison between different bias mitigation methods, for our experiments, we implement the same ResNet-based [49] CNN for all the methods consisting of 4 convolutional blocks, each with 2 conv layers, a max-pooling layer and implementing drop-out with batch-normalisation. The output of the last conv block is connected to 3 dense layers and a classification layer making model predictions. *ReLU* activation is used for each conv and dense layer. The same architecture is used to implement all the approaches compared in this work (see Fig. 1) with the exception of the Disentangled Approach for which the results from the original paper [38] are used directly.

3.1.1 *The Baseline*

For baseline evaluations, we split the datasets into different subsets based on the domain attributes. For example, for gender, the datasets are split into *male* and *female* splits and model performance is reported when trained incrementally on these data splits. This is sometimes also referred to as *fine-tuning* [50]. The model (see Fig. 1a), without any explicit mechanism to preserve knowledge, is expected *forget* old tasks while preference is given to the new tasks.

3.1.2 *Off-line Training*

Providing another baseline, the above-described Convolutional Neural Network (CNN) model (see Fig. 1a) is trained on all the training data, *off-line*, at once but its performance is reported individually on domain-specific test-splits. Off-line training provides a fair comparison with traditional ML-based learning models and is a popularly used benchmark for evaluating the performance of CL-based methods.

3.2 Non-CL-based Bias Mitigation Strategies

We investigate popular bias mitigation strategies from literature and implement 5 different methods for comparison. We group them under 'non-CL-based' strategies to differentiate them from the baselines and CL methods.

3.2.1 *Focal Loss for Mitigating Bias*

In-processing approaches modulate the training loss for the model by weighting the training samples based on how the different classes are distributed. One such approach is the use of Focal Loss (FL) [51] for training models to handle class-imbalances by determining 'hard samples' that may easily be misclassified by the model and assigning them a higher importance while loss contributions from 'easy samples' are down-weighted. FL is computed as follows:

$$\mathrm{FL}(p_\mathrm{t}) = -\alpha_\mathrm{t}(1 - p_\mathrm{t})^{\gamma} \log(p_\mathrm{t}) \qquad (1)$$

where $log(p_t)$ is the cross-entropy loss, α_t balances the importance of positive or negative samples, and γ is a tuneable *focusing* factor adjusting the rate at which easy samples are downweighted. We set $\alpha = 0.25$ and $\gamma = 2.0$ following the recommendations from [51]. As FL determines easy or hard samples based on the class distribution, when learning across domains, if the class distribution remains similar, samples from same classes may be considered easy or hard for all domain groups. Thus, FL reduces differences in model performance on individual classes instead of actively balancing model performance across domain groups.

3.2.2 Domain Discriminative Classification (DDC)

A popular method for mitigating bias is to focus on achieving 'fairness through awareness' [18] where information about sensitive attributes (or domains) is explicitly learnt in feature encodings. This information later allows models to account for bias in learning by being more 'aware'. One way to achieve this to create an $N \times M$ discriminative classifier where N denotes the number of domains and M is the number of classes to be learnt [19]. For example, for FER classifying 7 different expression classes for samples encoding 3 different race labels, a classifier is used with each output unit corresponding to a unique expression-race label pair (in this case, $7 \times 3 = 21$ label pairs). This allows the model to be more 'aware' of the different domains in order to learn discriminative features for each of them. For our experiments, we use the same model architecture (see Fig 1b) only replacing the output layer.

3.2.3 Domain Independent Classification (DIC)

A major concern with the DDC method is that the network may implicitly learn decision boundaries within the same class across different domains. This may be redundant as, despite the different domain attributions, the class-boundaries may remain the same and the network may be unnecessarily penalised due to incorrect domain predictions even if it predicts the task correctly. Wang et al. [19] offer a solution to this by training separate classifiers for each domain, sharing the feature extraction layers. For our experiments, we make use of the same model architecture (see Fig. 1c), connecting separate classifiers for each of the domains. The DIC model consists of different *heads*, each consisting of the same number of output units but corresponding to different domain attribute labels.

3.2.4 Strategic Sampling (SS)

A simple approach for handling bias arising from imbalanced data distributions is to strategically sample data [16] for each domain-class mapping such that the resultant data distribution 'appears to be balanced'. Samples from under-represented distributions can be sampled more often during training or equivalently, prediction loss can be appropriately weighted to account for under-represented classes. In our experiments, samples (s) for each of the N domains (d_i) are assigned a weight (w_i) *inversely proportional* to the occurrence rate of samples for that domain, scaling the loss function to account for imbalances in the training set distribution. The scaled cross-entropy loss is given as:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{i=1}^{N} w_i \sum_{s=1}^{S} y_s^{(i)} \log \hat{y}_s^{(i)} \tag{2}$$

3.2.5 Disentangled Feature Learning (DA)

Xu et al. [38] implement the Disentangled Feature Learning (DA) approach [39] for facial expression recognition that ensures that the feature representations learnt by a model do not contain any domain-specific information. The two sub-parts of the model focus on analysing facial expressions while learning to suppress domain-specific information. Here, we use their results [38] directly for comparison.

3.3 Continual Learning Approaches

Domain-Incremental CL deals with scenarios where the structure of the tasks remains the same albeit with changing input distribution [23]. We explore expression recognition and AU detection in a domain-incremental manner where the models learn to solve these tasks as the input data distributions change with respect to domain attributes of gender and race. For example, for gender, the models first learn to classify expression classes or predict activated AUs for 'male' samples and then, sequentially, learn to solve these tasks for 'female' samples (or vice-versa), without forgetting the previous task. Each method is implemented using the Baseline CNN architecture as shown in Fig 1a.

3.3.1 Elastic Weight Consolidation (EWC)

The EWC approach [43] imposes a quadratic penalty on parameter updates between old and new tasks in order to avoid forgetting previously learnt information. For each parameter θ, its relevance is calculated with respect to the task's training data $\mathcal{D}$, modelled as the posterior distribution $p(\theta|\mathcal{D})$. Thus, for two data distributions $\mathcal{D}_A$ and $\mathcal{D}_B$, corresponding to two independent tasks A and B, according to Bayes' rule, the posterior probability is given as:

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B), \tag{3}$$

such that $\log p(\theta|\mathcal{D}_A)$ embeds all information about previously learnt tasks. As this term becomes intractable, Laplace approximation is used to approximate it as a Gaussian Distribution with its mean given by θ^*_A (referring to parameters of task A) and the importance of the parameters determined by the diagonal of the Fischer Information Matrix. The loss function for the EWC method thus becomes:

$$L(\theta) = L_B(\theta) + \frac{1}{2}\lambda \sum_i F_i(\theta_i - \theta^*_{A,i})^2, \tag{4}$$

where L_B is the loss for task B, λ is the regularisation coefficient that determines the relevance of old tasks with respect to the new one, i denotes the index of the parameter θ and F_i is the i^{th} diagonal element of the Fischer Matrix. More generally, for each new task, an additional quadratic penalty is added with the loss function as:

$$L(\theta) = L_T(\theta) + \frac{1}{2}\lambda \sum_t^{T-1} \sum_i F_i(\theta_i - \theta^*_{t,i})^2, \tag{5}$$

where L_T is the loss for task T and $t \in [1, T-1]$ corresponds to previously seen tasks.

3.3.2 EWC-Online

A disadvantage for the EWC method is that as the number of tasks increase, the number of quadratic terms in the regularisation term grows linearly. To handle this, Schwarz et al. [52] propose a modification to EWC where instead of many quadratic terms, a single quadratic penalty is applied, determined by a running sum of the Fischer Information Matrices of previous tasks. The updated regularisation term of EWC-online is given as:

$$L_{reg}^{T} = \sum_{i} \tilde{F}_i^{(T-1)} (\theta_i - \theta_i^{(T-1)})^2, \tag{6}$$

where $\theta_i^{(T-1)}$ is the i^{th} parameter after learning task $T-1$ and $\tilde{F}_i^{(T-1)}$ is the running sum of the diagonal elements of the Fischer Matrices of all previous tasks calculated as:

$$\tilde{F}_i^{(T)} = \gamma \tilde{F}_i^{(T-1)} + F_i^T, \tag{7}$$

where γ controls the contribution of previously learnt tasks.

3.3.3 Synaptic Intelligence (SI)

Similar to EWC, SI also penalises changes to relevant weight parameters (synapses) in a manner that new tasks can be learnt without forgetting the old [44]. To alleviate forgetting, the importance for solving a learned task is computed for each individual synapse and changes in the most important synapses are discouraged. A modified cost function L_n^* is used with a surrogate loss term which approximates the summed loss functions of all the previous tasks L_o^*:

$$L_n^* = L_n + c \sum_{i} \Omega_k^n (\theta_k^* - \theta_k)^2, \tag{8}$$

where θ_k represents the parameters for the new task, θ_k^* represents the parameters at the end of the previous task, Ω_k^n is the parameter regulation strength and c is the weighting factor balancing new vs. old learning.

3.3.4 Memory Aware Synapses (MAS)

MAS also calculates the importance of each parameter by looking at the sensitivity of the output function instead of the loss [50]. For each new sample, MAS updates the importance of each parameter by evaluating how sensitive the model prediction is to the changes in that parameter. For an input x, the model prediction is given by $F(x;\theta)$, where θ represents model parameters. Introducing a small perturbation δ_{ij} in θ may result in a change in model output:

$$F(x;\theta+\delta) - F(x;\theta) \approx \sum_{i,j} g_{i,j}(x)\delta_{ij} \tag{9}$$

where $g_{i,j}(x) = \frac{\partial F(x;\theta)}{\partial \theta_{ij}}$ is the change in the output of F with respect to θ_{ij} evaluated for input x, and δ_{ij} is the perturbation introduced for $\theta_{i,j}$. The importance weight $\Omega_{i,j}$ for parameter $\theta_{i,j}$ can thus be determined by the magnitude of the gradient g_{ij}, that is, change in model output for an input x based on the change (δ_{ij}) in model parameters:

$$\Omega_{i,j} = \frac{1}{N} \sum_{k=1}^{N} \| g_{i,j}(x_k) \| \tag{10}$$

where N is the number of samples. Thus, parameters that have the most impact on model predictions are given high importance and changes to these parameters are penalised. Different from EWC and SI, parameter importance is computed only using unlabelled data by measuring changes in model performance. For each new task T_n, in addition to the task-loss $L_n(\theta)$, changes to parameters important for previous tasks are penalised:

$$L(\theta) = L_n(\theta) + \lambda \sum_{i,j} \Omega_{i,j} (\theta_{i,j} - \theta_{i,j}^*)^2, \tag{11}$$

where λ is the hyperparameter balancing new vs. old task losses, $\Omega_{i,j}$ is the importance computed for parameter θ_{ij} and θ^* denotes the old network parameters.

3.3.5 Naive Rehearsal (NR)

For Naive Rehearsal (NR) [53], we implement a straightforward rehearsal-based method that combines new data with previously seen data while training the model. A small replay buffer of size N is implemented to randomly store a fraction of previously seen data samples that can be replayed to the model. Each mini-batch of data is constructed using an equal number of samples from the new as well as previously seen data. This interleaving of data pertaining to previously learnt tasks with new data ensures that old knowledge is not overwritten by new data.

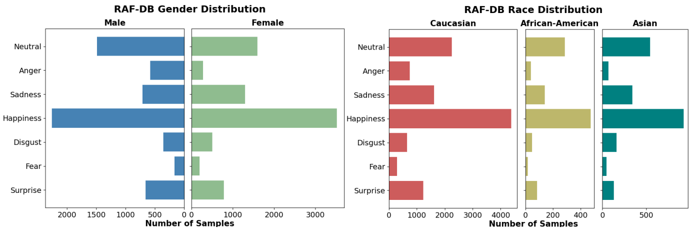
4 Experiment Set-Up

4.1 Datasets

For evaluating the different bias mitigation strategies and comparing them with CL-based methods, we use two popular benchmark datasets; the RAF-DB dataset for FER *in-the-wild* and the BP4D dataset recorded for AU detection in controlled settings. These datasets are selected due to (i) the diversity in their data acquisition settings, (ii) providing labels not only for expression/AU recognition but also gender and race attributes, and (iii) containing notable imbalances in the data distributions with respect to class and domain attribute labels. These factors make the RAF-DB and the BP4D datasets a good choice for our evaluation.

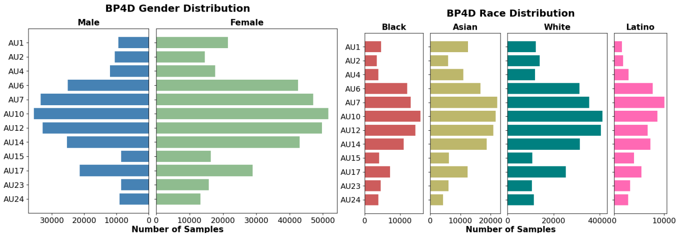
4.1.1 RAF-DB Dataset

The RAF-DB dataset [54] consists of $\approx 15K$ facial images labelled for six expression classes namely, *Surprise, Fear, Disgust, Happy, Sad* and *Anger* along with *Neutral* to denote absence of any expression. Additionally, it provides demographic attribute labels such as gender (Male, Female, Unsure) and race (Caucasian, African-American, Asian) labels. For our experiments, we split the dataset using multiple grouping strategies based on the gender and race Labels. For gender-based grouping, we exclude images labelled as 'Unsure' and only use the 'Male' and 'Female' samples. As shown in Fig. 2, not only is the dataset imbalanced with respect to the different expression categories, there also exist stark imbalances with respect to different demographic attributes as well. The majority of the samples in the training set represent the 'Happy' expression class and belong to 'Female' and 'Caucasian' categories.



(a) Gender (b) Race

Fig. 2. RAF-DB data distribution for Gender and Race Attributes.



(a) Gender (b) Race

Fig. 3. BP4D data distribution for Gender and Race Attributes.

TABLE 1
Regularisation Coefficient values and Buffer Size used for FER experiments with the RAF-DB dataset ($\times 10^3$).

Method	W/O Data-augmentation		W/ Data-augmentation	
	Gender	*Race*	*Gender*	*Race*
EWC (λ)	5	10	10	5
EWC-Online (γ)	10	1	10	5
SI (c)	1	10	1	5
MAS (λ)	0.5	5	1	0.2
NR (N)	0.4	1.1	0.1	1.1

TABLE 2
Regularisation Coefficient values and Buffer Size used for AU detection experiments with the BP4D dataset ($\times 10^3$).

Method	W/O Data-augmentation		W/ Data-augmentation	
	Gender	*Race*	*Gender*	*Race*
EWC (λ)	0.05	5	5	1
EWC-Online (γ)	0.01	0.5	0.01	0.1
SI (c)	5	0.1	10	0.01
MAS (λ)	0.5	0.01	5	5
NR (N)	0.2	0.2	1.1	0.4

4.1.2 BP4D Dataset

The BP4D dataset [55] consists of video sequences from 41 subjects performing 8 different affective tasks to elicit emotional reactions. Each video is annotated frame-wise for the occurrence and intensity of the activated AUs. In our experiments, we only use occurrence labels for 12 most frequent AU resulting in $\approx 150K$ labelled frames, in total. Other than the frame-wise AU labels, demographic attribute labels for gender (Male, Female) and race (Black, White, Latino, Asian) have been provided to us, specifically for this research. Fig. 3 shows the data distribution of the BP4D dataset for the 12 AU labels with respect to the gender and race attributes. As can be seen, the majority of the samples in the dataset represent 'White' and 'Female' attribute labels.

4.2 Pre-processing and Data-Augmentation

Even though both RAF-DB and BP4D datasets provide face-centred RGB images, we use the dlib Python Library to crop-out only the face region and resize the images to $(100 \times 100 \times 3)$ while also normalising them to be used as input for all the models. The image size is inspired from other approaches in literature [8], [37] and limitations in terms of the available GPU resources. Training deep neural networks requires a lot of training data for each of the classes to be learnt. Due to the inherent imbalances in the dataset with respect to the different expression classes (RAF-DB) or the AU labels (BP4D), we increase the overall training data by performing data-augmentation by randomly ($p = 0.5$) flipping images horizontally to create additional samples. For each experiment, we present the results *with* and *without* data-augmentation separately, for clarity.

4.3 Experiment Settings

4.3.1 Evaluation Metrics

To compare the different methods on their ability to balance classification performance within individual domain-splits while remaining consistent across the domains, we evaluate them both in terms of their performance, using accuracy scores for FER and F1-Scores for AU detection, as well as *fairness*. Furthermore, for the CL methods, we also report Catastrophic Forgetting (CF) scores, measuring the ability of the models to maintain performance on previously seen tasks while learning new tasks.

Accuracy (Acc): Accuracy is defined as the fraction of correctly classified samples. Given that TP = True Positives, FP = False Positives, TN = True Negatives and FN = False Negatives, Accuracy (Acc) can be computed as:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

F1-Score (F1): F1-Score is defined as harmonic mean of the precision (P) and recall (R) scores and is computed as:

$$F1 = \frac{2RP}{R + P} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{13}$$

In our experiments, we report accuracy and F1-scores separately for different gender and race attributes to highlight differences in model performance for these domains, underlining bias in the models' performance.

Fairness Measure ($\mathcal{F}$): To evaluate different approaches for their *fairness* with respect to model performance for gender and race attributes, we use the 'equal opportunity' definition of *fairness*, as proposed by Hardt et al. [56].

Let $\mathbf{x}$, $\mathbf{y}$, $\hat{\mathbf{y}}$ be the variables denoting input, ground truth label and the predicted label, respectively, $s \in S_i$ be the sensitive (domain) attribute (for example, ($S_i = \{$male, female$\}$), f be a function computing the *accuracy score* for a given sensitive attribute s and d be the dominant attribute which has the highest *accuracy score*, then the *Fairness Measure* $\mathcal{F}$ of a model is defined as the largest accuracy gap among all sensitive attributes computed as the minimum of the ratios of the accuracy scores of each sensitive attribute with respect to the dominant attribute.

TABLE 3
Experiment 1: Gender-wise Accuracy and Fairness Scores on RAF-DB dataset. Accuracy scores are reported after training the models on both Male and Female subsets. **Bold** values denote best while [*bracketed*] denote second-best values for each column.

Method	Accuracy W/O Data-Augmentation		Fairness	Accuracy W/ Data-Augmentation		Fairness
	Male	*Female*		*Male*	*Female*	
Baseline	0.596±0.025	0.714±0.014	0.834	0.596±0.017	0.730±0.008	0.816
Offline	0.704±0.011	0.746±0.007	0.944	0.724±0.006	0.759±0.007	0.954
Non-CL-based Bias Mitigation Methods						
Focal Loss [51]	0.696±0.008	0.736±0.002	0.945	0.716±0.010	0.750±0.005	0.954
DDC [18]	0.699±0.013	0.722± 0.008	0.968	0.717±0.013	0.746±0.007	0.961
DIC [19]	0.698±0.014	0.744± 0.006	0.938	0.729±0.008	0.758±0.002	0.962
SS [16]	0.716±0.010	[*0.750±0.008*]	0.955	0.729±0.013	[*0.764±0.011*]	0.954
DA [38]	0.625	0.610	0.975	[*0.742*]	0.744	[*0.997*]
Continual Learning Methods						
EWC [43]	**0.723± 0.006**	0.744±0.006	0.972	0.735±0.007	0.748±0.012	0.983
EWC-Online [52]	0.721± 0.008	0.743±0.006	0.970	0.736±0.003	0.756±0.010	0.974
SI [44]	0.718± 0.007	0.725±0.004	**0.990**	0.739±0.008	0.739±0.005	**0.999**
MAS [50]	0.721± 0.008	0.735±0.012	[*0.980*]	**0.745±0.006**	0.753±0.009	0.990
NR [53]	[*0.722± 0.001*]	**0.778±0.006**	0.928	0.738±0.004	**0.799±0.005**	0.923

TABLE 4
Experiment 1: Catastrophic Forgetting (CF) and Overall Accuracy (previous tasks) after each task for Gender-ordered learning on RAF-DB dataset. **Bold** values denote the best while [*bracketed*] denote second-best values for each column.

Method	W/O Data-Augmentation				W/ Data-Augmentation			
	Task 1		Task 2		Task 1		Task 2	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [43]	**0.730**	X	[*0.734*]	**-0.028**	**0.746**	X	0.742	**-0.032**
EWC Online [52]	0.728	X	0.733	-0.015	[*0.745*]	X	0.746	[*-0.024*]
SI [44]	0.721	X	0.728	-0.003	0.741	X	0.738	-0.004
MAS [50]	0.721	X	0.731	[*-0.024*]	0.743	X	[*0.749*]	-0.013
NR [53]	[*0.729*]	X	**0.753**	-0.010	0.736	X	**0.772**	-0.010

$$\mathcal{F} = \min(\frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_0, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})}, ..., \frac{f(\hat{\mathbf{y}}, \mathbf{y}, s_n, \mathbf{x})}{f(\hat{\mathbf{y}}, \mathbf{y}, d, \mathbf{x})}) \tag{14}$$

In other words, $\mathcal{F}$ is defined as the ratio of the lowest accuracy for a sensitive attribute with respect to the highest accuracy value for that sensitive attribute. We adapt the Fairness Measure $\mathcal{F}$ to use F1-Scores instead of accuracy scores when evaluating *fairness* for the BP4D dataset.

Catastrophic Forgetting (CF): Catastrophic forgetting occurs when learning a new task negatively impacts previously learnt information. We report the *CF* metric score [57] for the CL methods, measuring the average change in the *accuracy scores* for each previous task right after learning a new task. This is computed as follows:

$$CF = \frac{\sum_{j=1}^{i-1} a_{j,j} - a_{i,j}}{i-1}, A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}$$

where $a_{i,j}$ denotes the accuracy of i^{th} task after learning j^{th} task, A is the matrix storing accuracy scores with dimensions $(n \times n)$ and n is the number of tasks or domains.

4.3.2 Implementation Details

All models are trained using the *adam* optimiser with a learning rate of $1.0e^{-4}$ and a batch-size of 24. For the experiments with RAF-DB, all models are trained for 25 epochs while for BP4D, due to a higher number of data samples, training converged after only 10 epochs. For RAF-DB experiments, we use the training and test-splits provided with the dataset [54] while for the BP4D dataset, we randomly select subjects to be included in the test-set maintaining the overall data distribution with respect to *gender* and *race* attributes, respectively. All experiments are *repeated* 3 times and the results are *averaged* across the repetitions to account for the random seeds. All models are implemented with PyTorch based on the CL benchmarks provided by [23], [53].

To ensure that the CL methods are optimised for FER and Facial AU detection, we run separate grid-based searches to optimise the regularisation coefficient values for the different models that drive their performance. Table 1 and 2 report the optimised regularisation coefficient values for experiments with the RAF-DB and BP4D datasets, respectively.

5 Experiments and Results

5.1 Experiment 1: Mitigating Bias in FER

We compare state-of-the-art bias mitigation approaches (see Section 3.2) with popular CL-based methods (see Section 3.3) on their ability to classify facial expressions without being affected by imbalances in data distributions. To evaluate the applicability of CL strategies as *'fair'* FER systems, we train and test these approaches on the RAF-DB dataset (both without and with data-augmentation) to learn 7 expression classes, namely, *surprise, sadness, happiness, fear, anger, disgust* and *neutral*, with respect to 2 different domain groups; gender (Male, Female) and race (Caucasian, African-American, Asian).

5.1.1 Bias Across Gender Attributes

For RAF-DB, about 53.4% of the samples are labelled as 'Female' while the 'Male' group constitutes about 40.3% of the total samples. The rest of the samples are labelled as 'Unsure' and omitted from our evaluations (see Fig. 2a). As a

TABLE 5
Experiment 1: Race-wise Accuracy and Fairness Scores on RAF-DB dataset. Accuracy scores are reported after training the models on all the subsets. **Bold** values denote best while [*bracketed*] denote second-best values for each column.

Method	Accuracy W/O Data-Augmentation			Fairness	Accuracy W/ Data-Augmentation			Fairness
	Caucasian	*African American*	*Asian*		*Caucasian*	*African American*	*Asian*	
Baseline	0.750±0.019	0.764±0.029	**0.795±0.010**	0.943	0.758±0.004	0.778±0.023	**0.809±0.008**	0.937
Offline	0.727±0.010	0.750±0.011	0.735±0.025	0.969	0.743±0.006	0.762±0.017	0.763±0.007	0.974
Non-CL-based Bias Mitigation approaches								
Focal Loss [51]	0.714±0.010	0.723±0.002	0.743±0.013	0.961	0.737±0.006	0.757±0.023	0.752±0.013	0.972
DDC [18]	0.714±0.009	0.710±0.009	0.721±0.009	0.985	0.729±0.006	0.736±0.001	0.747±0.007	0.976
DIC [19]	0.724±0.004	0.730±0.015	0.732±0.016	0.989	0.745±0.007	0.768±0.012	0.772±0.013	0.965
SS [16]	0.734±0.005	0.728±0.015	0.757±0.014	0.961	0.748±0.002	0.752±0.019	0.767±0.023	0.975
DA [38]	0.634	0.584	0.544	0.858	0.756	0.766	0.704	0.919
Continual Learning approaches								
EWC [43]	0.764±0.011	0.758±0.002	0.768±0.016	0.987	**0.796±0.006**	[*0.788±0.009*]	0.794±0.007	0.990
EWC-Online [52]	[*0.773±0.018*]	0.763±0.003	0.773±0.002	0.987	0.777±0.017	0.780±0.002	0.785±0.014	0.990
SI [44]	0.769±0.010	[*0.766±0.006*]	0.769±0.008	**0.996**	0.785±0.013	0.783±0.003	0.782±0.009	**0.996**
MAS [50]	0.762±0.007	0.756±0.001	0.764±0.010	[*0.990*]	0.781±0.017	0.776±0.012	0.781±0.007	[*0.994*]
NR [53]	**0.779±0.015**	**0.772±0.017**	[*0.793±0.002*]	0.974	[*0.787±0.012*]	**0.796±0.005**	[*0.808±0.014*]	0.974

TABLE 6
Experiment 1: Catastrophic Forgetting (CF) and Overall Accuracy (previous tasks) after each task for Race-ordered learning on RAF-DB dataset. **Bold** values denote best while [*bracketed*] denote second-best values for each column.

Method	W/O Data-Augmentation						W/ Data-Augmentation					
	Task 1		Task 2		Task 3		Task 1		Task 2		Task 3	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [43]	0.777	X	**0.754**	[*0.025*]	0.764	0.019	0.797	X	**0.768**	**0.030**	**0.795**	[*0.000*]
EWC-Online [52]	[*0.778*]	X	0.738	0.046	[*0.779*]	0.012	0.798	X	[*0.763*]	[*0.038*]	0.779	0.026
SI [44]	**0.782**	X	0.736	0.047	0.769	**-0.003**	**0.802**	X	0.759	0.044	0.784	0.007
MAS [50]	0.766	X	[*0.744*	**0.023**	0.762	[*0.003*]	[*0.801*]	X	0.762	0.039	0.780	0.004
NR [53]	[*0.778*]	X	0.734	0.049	**0.781**	0.009	0.784	X	0.744	0.043	[*0.790*]	**-0.006**

result, the effective split of the dataset with respect to gender is somewhat balanced, 56.3% Female against 43.7% Male samples. Furthermore, the class distribution between Male and Female sub-sets is also similar with 'Happy' samples dominating both distributions. For non-CL-based methods, the models are trained on the entire dataset and tested individually on the Male and Female subsets. For the CL evaluations, however, the learning is split into two tasks corresponding to expression recognition for the Male (Task 1) followed by expression recognition for the Female (Task 2) sub-sets. For each of the tasks, individually, the models learn all the 7 expression classes together. The effect of domain-ordering, that is, whether to learn with *male* samples first or vice-versa, is discussed further in Section 6.1.

Table 3 presents the results comparing the different methods on their Accuracy and Fairness scores. CL methods, overall, outperform all other methods both in accuracy as well as fairness scores while the baseline method is the most unfair. Furthermore, although the accuracy scores of all the approaches increase when data-augmentation is used, not all of them are able to maintain fairness. CL methods (with the exception of NR) on the other hand, improve upon their fairness scores, with SI [44] performing the best both without and with data-augmentation. Individual class-wise performance for the CL methods is presented in Tables 7 of the *supplementary material* provided with the overall model accuracy being the worst for *disgust* due to the overall low number of samples, in line with what was reported in [38].

To fully appreciate how CL enables models to retain their performance across the two tasks, it is important to understand how learning each new task impacts the models' performance on previously learnt tasks. Table 4 reports the overall accuracy and CF scores for the CL methods after each task, evaluating the performance of these methods in terms of their ability to classify expressions for both male (Task 1) and female (Task 2) sub-sets. We observe that both with and without augmentation, NR [53] achieves the highest overall accuracy after both tasks are learnt, while EWC experiences the least forgetting. Furthermore, negative CF scores for all CL methods indicate that after learning Task 2, that is, to predict expressions on Female samples, the overall accuracy increased for both Male and Female samples, without any forgetting occurring in the model.

5.1.2 Bias Across Race Attributes

The data distribution for RAF-DB dataset is highly imbalanced with respect to Race with a majority (77.4%) of the samples labelled as Caucasian, while the African-American and Asian subsets correspond to only 7.1% and 15.5% of the samples, respectively. However, the relative class-distributions within individual race groups remains similar (see Fig. 2b). Similar to gender experiments, for the non-CL-based methods, models are trained on the entire dataset and evaluated individually for the different race attributes. For CL evaluations, the learning is split into three tasks corresponding to learning to predict expressions for Caucasian (Task 1), African-American (Task 2) and Asian (Task 3) faces.

Table 5 presents the results of the experiments comparing Accuracy and Fairness scores. The imbalances in the data distribution with respect to the domain attributes affect all the approaches such that the model accuracy varies across the race groupings. Yet, the CL methods seem to handle this best, achieving comparable accuracy with high fairness scores. Even though CL is not always the best

TABLE 7

Experiment 2: Gender-wise F1-Scores and Fairness Scores on BP4D dataset. F1-scores are reported after training the models on both Male and Female subsets. **Bold** values denote best while [*bracketed*] denote second-best values for each column.

Method	**F1 Scores W/O Data-Augmentation**		**Fairness**	**F1 Scores W/ Data-Augmentation**		**Fairness**
	Male	***Female***		***Male***	***Female***	
Baseline	0.428±0.020	**0.626±0.016**	0.683	**0.545±0.035**	**0.646±0.025**	0.843
Offline	0.495±0.024	[*0.573±0.021*]	0.865	[*0.530±0.027*]	0.601±0.024	0.881
Non-CL-based Bias Mitigation approaches						
Focal Loss [51]	0.453±0.012	0.486±0.009	0.932	0.459±0.008	0.480±0.021	0.956
DDC [18]	0.453±0.015	0.507±0.017	0.893	0.492±0.025	0.561±0.026	0.877
DIC [19]	0.429±0.028	0.497±0.026	0.863	0.435±0.063	0.502±0.054	0.866
SS [16]	0.492±0.009	0.564±0.015	0.872	0.521±0.012	0.574±0.021	0.907
DA [38]	0.462	0.533	0.868	0.541	[*0.621*]	0.871
Continual Learning approaches						
EWC [43]	**0.524±0.021**	0.523± 0.032	**0.997**	0.501±0.025	0.507±0.069	[*0.988*]
EWC-Online [52]	0.492±0.026	0.523± 0.030	0.941	0.467±0.043	0.496±0.028	0.942
SI [44]	0.469±0.020	0.472± 0.082	[*0.993*]	0.512±0.014	0.519±0.066	[*0.988*]
MAS [50]	[*0.496±0.069*]	0.511±0.024	0.970	0.511±0.026	0.507±0.023	**0.992**
NR [53]	0.474±0.026	0.498± 0.023	0.952	0.520±0.020	0.529±0.015	0.983

TABLE 8

Experiment 2: CF and Overall Accuracy (previous tasks) after each task for Gender-ordered learning on BP4D dataset. **Bold** values denote the best while [*bracketed*] denote second-best values for each column.

Method	**W/O Data-Augmentation**				**W/ Data-Augmentation**			
	Task 1		Task 2		Task 1		Task 2	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [43]	[*0.745*]	X	[*0.768*]	[*-0.020*]	0.729	X	[*0.761*]	**-0.024**
EWC Online [52]	0.744	X	**0.769**	**-0.022**	0.728	X	0.753	[*-0.016*]
SI [44]	**0.753**	X	0.764	0.002	[*0.736*]	X	0.749	-0.008
MAS [50]	0.734	X	0.765	-0.011	**0.745**	X	**0.766**	-0.002
NR [53]	0.740	X	0.759	0.000	[*0.736*]	X	0.758	-0.006

performing, particularly for the *Asian* subset, all of them achieve high fairness scores, with SI performing the best. This underlines their ability to balance learning across the different tasks. They are able to give preference to being consistent and *fair*, trading-off higher accuracy scores for any individual race label. The NR approach, on the other hand, achieves the highest accuracy scores on Task 1 and Task 2, owing to the explicit replay mechanism, but sacrifices fairness across all groups in the process. Additionally, as RAF-DB is a relatively small dataset, data-augmentation has a positive effect on accuracy scores of all the models, but the fairness scores do not change significantly. Individual class-wise performance for the CL models is presented in Tables 8 of the *supplementary material*.

In Table 6, the accuracy and CF scores can be seen for all the CL methods reporting model performance on all previous tasks computed at the end of each new task. We see that all models tend to forget as they learn new tasks, yet the SI method is able to mitigate forgetting the best. When data-augmentation is used, the individual accuracy scores are enhanced but the CF scores do not improve.

5.2 Experiment 2: Mitigating Bias in AU Detection

As more than one AU may be *activated* at the same time (for example, AU $1, 2$ and 26 together may depict *surprise*), predicting AUs poses a multi-label classification problem. Imbalances in data distributions for different *gender* and *race* attributes become even more prominent with certain AUs having much more data samples than others (see Fig. 3). With the Domain-IL protocol for CL evaluations, we focus on how splitting learning based on domain-attributes can impact model performance in terms of F1-Scores and Fairness evaluations. We compare different bias mitigation strategies (see Section 3.2) with CL-based methods (see Section 3.3) to understand how they cope with imbalances in data distributions with respect to domain attributes (rather than AU labels, explicitly) while retaining model performance. Similar to Experiment 1, we compare the performance of models (without and with data-augmentation) on detecting activations for 12 AUs, together, for *gender* (Male, Female) and *race* (White, Black, Asian, Latino) groups.

5.2.1 Bias Across Gender Attributes

Similar to Experiment 1, for the non-CL-based methods, we train the individual models on the entire dataset but evaluate them individually for Male and Female subsets. The BP4D data distribution is skewed in favour of Female samples constituting 60.96% of the data while only 39.04% samples belong to the Male sub-set (see Fig. 3a). However, within each sub-set the relative data distribution with respect to AU labels largely remains the same. For CL methods, the learning is split into two *tasks*: Task 1: Male and Task 2: Female, incrementally learning to detect all AU activations in the two data splits. The effect of domain-ordering, that is, whether to learn with Male samples first or vice-versa, is discussed in Section 6.2.

Table 7 compares the different methods on their F1-Scores and Fairness for both Male and Female subsets. CL methods are shown to consistently out-perform all other methods in terms of the fairness scores while offering competitive performance on F1-scores. Despite offering some of the best individual F1-scores, baseline evaluations perform the worst in terms of model *fairness* with the skew in data

TABLE 9

Experiment 2: Race-wise F1-Scores and Fairness Scores on BP4D dataset. F1-scores are reported after training the models on all the subsets. **Bold** values denote best while [*bracketed*] denote second-best values for each column.

Method	F1 Scores W/O Data-Augmentation				Fairness	F1 Scores W/ Data-Augmentation				Fairness
	Black	*Asian*	*White*	*Latino*		*Black*	*Asian*	*White*	*Latino*	
Baseline	0.372±0.035	0.303±0.032	**0.597±0.041**	0.528±0.024	0.507	0.370±0.051	0.331±0.041	**0.654±0.062**	[*0.565±0.032*]	0.506
Offline	0.497±0.025	[*0.520±0.036*]	0.507±0.028	0.517±0.063	0.923	0.503±0.062	[*0.521±0.052*]	0.529±0.055	0.553±0.047	0.884
Non-CL-based Bias Mitigation Approaches										
Focal Loss [51]	**0.526±0.009**	0.453±0.009	0.459±0.008	0.453±0.016	0.860	0.516±0.012	0.459±0.021	0.480±0.008	0.410±0.029	0.793
DDC [18]	0.492±0.032	0.481±0.036	0.532±0.028	[*0.541±0.047*]	0.889	0.501±0.008	0.510±0.006	0.541±0.023	0.542±0.031	0.924
DIC [19]	0.485±0.062	0.479±0.060	0.504±0.038	0.520±0.091	0.921	0.495±0.017	0.502±0.043	0.524±0.029	0.503±0.039	0.944
SS [16]	0.510±0.028	**0.524±0.036**	[*0.561±0.047*]	**0.542±0.083**	0.909	0.513±0.059	0.504±0.061	0.482±0.091	0.491±0.071	0.939
DA [38]	0.501	0.502	0.495	0.511	**0.968**	**0.540**	**0.609**	[*0.572*]	**0.624**	0.866
Continual Learning Approaches										
EWC [43]	0.484±0.019	0.508±0.024	0.512±0.011	0.507±0.059	[*0.944*]	[*0.517±0.028*]	0.486±0.010	0.486±0.018	0.476±0.022	0.920
EWC-Online [52]	[*0.518±0.061*]	0.484±0.043	0.520±0.028	0.520±0.029	0.931	0.463±0.052	0.464±0.021	0.464±0.016	0.484±0.024	0.957
SI [44]	0.455±0.029	0.468±0.052	0.470±0.064	0.461±0.111	**0.968**	0.471±0.059	0.476±0.019	0.466±0.009	0.465±0.065	**0.976**
MAS [50]	0.516±0.056	0.489±0.010	0.492±0.037	0.525±0.047	0.931	0.432±0.026	0.429±0.081	0.429±0.017	0.446±0.049	[*0.961*]
NR [53]	0.491±0.041	0.441±0.016	0.471±0.044	0.487±0.067	0.898	0.469±0.023	0.456±0.054	0.491±0.011	0.481±0.024	0.928

TABLE 10

Experiment 2: CF and Overall Accuracy (previous tasks) after each task for Race-ordered learning on BP4D dataset. **Bold** values denote the best while [*bracketed*] denote second-best values for each column.

Method	W/O Data-Augmentation								W/ Data-Augmentation							
	Task 1		Task 2		Task 3		Task 4		Task 1		Task 2		Task 3		Task 4	
	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF	Acc.	CF
EWC [43]	0.763	X	0.707	0.059	0.692	[*0.125*]	0.761	[*-0.042*]	0.763	X	0.686	0.087	0.673	**0.129**	0.748	-0.026
EWC-Online [52]	**0.773**	X	0.693	0.079	0.682	**0.122**	0.757	-0.032	0.766	X	0.697	0.081	0.682	[*0.140*]	0.754	-0.026
SI [44]	[*0.765*]	X	**0.734**	**0.044**	**0.714**	0.154	[*0.764*]	-0.027	[*0.774*]	X	0.702	0.091	[*0.686*]	0.147	[*0.768*]	-0.032
MAS [50]	0.759	X	[*0.724*]	[*0.048*]	[*0.707*]	0.144	0.762	-0.035	0.754	X	**0.717**	**0.056**	**0.697**	0.162	0.762	[*-0.033*]
NR [53]	0.759	X	0.710	0.055	0.690	0.143	**0.777**	**-0.053**	**0.775**	X	[*0.705*]	[*0.079*]	0.685	0.156	**0.780**	**-0.035**

distribution between male and female samples impacting model performance. CL models are able to balance learning across both the domain groups, resulting in the best *fairness* scores, although trading-off higher F1-scores on individual tasks. EWC performs the best in terms of *fairness* while also providing the best F1-scores for *male* samples. Data-augmentation, overall, has a positive impact on model performance but does not impact model fairness, significantly for all the approaches. Yet, for MAS the additional data allows for the model to assign importance to relevant features, resulting in the highest fairness scores. Individual AU-wise results between Male and Female splits do not vary significantly for the 12 AU labels with AU 2 and AU 12 achieving the lowest and highest F1-scores, respectively, across the models for both the splits. These results are provided in Table 11 of the supplementary material. This can be due to these classes consisting of the lowest and highest number of samples across both gender and race splits (see Fig 3).

Comparing different CL methods on their ability to maintain performance across the tasks, Table 8 reports the overall accuracy and CF scores for the models. We use accuracy here instead of F1-scores as CF is defined to use accuracy scores and may not be directly adapted to use F1-scores. Owing to the complex multi-label nature of the tasks as well as the high gender disparity in the data distribution, we see a high variation in the performance of the different CL methods. While EWC-Online performs the best without employing data-augmentation achieving a negative CF score, the MAS model performs the best with data-augmentation. EWC is the most consistent maintaining model performance while alleviating forgetting.

5.2.2 Bias Across Race Attributes

The majority of the samples in the BP4D dataset are labelled as White (approximately 46.76%) with other samples corresponding to Asian (26.08%), Black (16.56%) and Latino (10.6%) groups (see Fig. 3b). For our evaluations, we split the dataset into 4 sub-sets based on these labels, representing the 4 tasks, that is, Task 1: Black, Task 2: Asian, Task 3: White and Task 4: Latino, for the CL models. Within each of the 4 tasks, the relative AU distribution remains largely the same (see Fig. 3b). The effect of domain-ordering, that is, which race group to start with and what order to follow, is discussed in detail in Section 6.2.

Table 9 presents race-wise F1-scores and Fairness evaluations. The baseline, despite offering the best F1-Scores for *White* samples, performs the worst in terms of model *fairness* both with and without data-augmentation. Non-CL methods improve upon baseline evaluations offering improved F1-scores and Fairness scores with DA achieving the best Fairness scores. CL methods, overall, achieve high Fairness scores across all methods with SI performing the best. CL approaches balance learning across the race groups, trading-off higher F1-Scores for better Fairness evaluations. Data-augmentation improves the F1-scores and Fairness across most evaluations with SI still remaining the *fairest*. AU-wise results for the CL models for race-ordered experiments are presented in Table 12 of the supplementary material.

Different CL models handle the high variance in data distribution with respect to racial identity labels with varying levels of success. Table 10 shows how, at different points during the learning, different models perform better than others, while NR achieves the highest accuracy and CF scores after all tasks are learnt, both with and without data-

TABLE 11
Experiment 1: Fairness Measure Scores across Gender and Race distributions for the RAF-DB Dataset. **Bold** values denote best while [*bracketed*] denote second-best values for each column.

Method	W/O Data-Augmentation		W/ Data-Augmentation	
	Gender	*Race*	*Gender*	*Race*
Baseline	0.834	0.943	0.816	0.937
Offline Training	0.944	0.925	0.954	0.974
Non-CL-based Bias Mitigation Methods				
Focal Loss [51]	0.945	0.961	0.954	0.972
DDC [19]	0.968	0.985	0.961	0.976
DIC [19]	0.938	0.989	0.962	0.965
SS [16]	0.955	0.961	0.954	0.975
DA [38]	0.975	0.858	[*0.997*]	0.919
Continual Learning Methods				
EWC [43]	0.972	0.987	0.983	0.990
EWC-Online [52]	0.970	0.987	0.974	0.990
SI [44]	**0.990**	**0.996**	**0.999**	**0.996**
MAS [50]	[*0.980*]	[*0.990*]	0.990	[*0.994*]
NR [53]	0.928	0.974	0.923	0.974

TABLE 12
Experiment 2: Fairness Measure Scores across Gender and Race distributions for the BP4D Dataset. **Bold** values denote best while [*bracketed*] denote second-best values for each column.

Method	W/O Data-Augmentation		W/ Data-Augmentation	
	Gender	*Race*	*Gender*	*Race*
Baseline	0.683	0.507	0.843	0.506
Offline	0.865	0.923	0.881	0.884
Non-CL-based Bias Mitigation Approaches				
Focal Loss [51]	0.932	0.860	0.956	0.793
DDC [19]	0.893	0.889	0.877	0.924
DIC [19]	0.863	0.921	0.866	0.944
SS [16]	0.872	0.909	0.907	0.939
DA [38]	0.868	**0.968**	0.871	0.866
Continual Learning Approaches				
EWC [43]	**0.997**	[*0.944*]	[*0.988*]	0.920
EWC-Online [52]	0.941	0.931	0.942	0.957
SI [44]	[*0.993*]	**0.968**	[*0.988*]	**0.976**
MAS [50]	0.970	0.931	**0.992**	[*0.961*]
NR [53]	0.952	0.898	0.983	0.928

augmentation. The negative CF scores for all the approaches at the end of Task 4 signifies that all the CL models were able to mitigate forgetting and the overall model performance improved as they incrementally learnt new tasks.

6 Discussion

Our experiments on FER (see Section 5.1) and AU detection (see Section 5.2) tasks are motivating as they highlight how adopting CL strategies may enable *fairer* facial affect analysis algorithms. Consistently achieving high accuracy as well as fairness measure scores, CL offers an improvement over existing learning strategies for bias mitigation in ML algorithms. Robustly managing imbalances in data distributions, both without and with data-augmentation, CL methods are better equipped to deal with biases owing to their learning strategy of focusing on one domain group at a time. Here, we discuss each task individually and highlight how CL provides a solution towards *fairer* facial affect analyses.

6.1 Facial Expression Recognition

When applied to FER, CL methods aim to sequentially learn to predict the 7 expression categories for the different gender and race groups. The models are trained with all the classes, one domain group at a time, and as the model experiences samples from other groups, they actively try to maintain performance at previously seen groups without forgetting. As a result, for both gender and race groups, CL models achieve high fairness scores by balancing performance across the domain splits while also offering competitive accuracy scores, with the SI model performing the best in terms of fairness (see Table 11). Yet, NR achieves a competitive average accuracy score (across domain groups) of 0.771 for gender and 0.807 for race evaluations, highest amongst the compared methods, with the benchmark evaluations on RAF-DB at 0.853 [58]. Selective updates of model parameters to mitigate forgetting allows CL models to maintain high accuracy scores across the different gender and race attributes. This makes them distinct from other approaches, directly focusing on balancing model performance across different domain distributions instead of deciding whether to capture domain-specific features or not. In comparison, non-CL-based methods rely on becoming 'aware' of domain attributes to predict expressions according to the subjects *sharing* gender or race attributes or learning feature representations that actively 'block' domain discriminative features [38]. Furthermore, for most non-CL methods, with the exception of DA, we need to know the domain groupings, for example, how many race groups exist, a priori which may not always be possible in real-world scenarios. For CL methods, however, as models learn sequentially, there is no need to provide any domain information a priori and learning can be extended to new domains.

Domain-Ordering: One concern when applying CL methods to FER tasks is the task or class-ordering effect where model performance is seen to be sensitive to the order in which it learns different expression classes [48]. In our experiments, as we implement the Domain-IL scenario where all classes are learnt at the same time, albeit one domain-group at a time, class-ordering does not play any role in the learning. Instead, we explore whether different domain-orderings, that is, learning with different sequencing of gender or race group splits has any effect on the models' ability to maintain performance. For both gender and race domains, we experiment with different orders of learning the tasks, but no significant effect of domain-ordering is witnessed on the models' performance. Individual results for these experiments can be found in Tables $1-6$ of the supplementary material provided.

6.2 Action Unit Detection

Action Unit (AU) detection poses a *harder* multi-label classification problem where the models need to predict all the AUs activated in a given sample. The inherent class-imbalances in the BP4D dataset are further accentuated by the imbalances with respect to gender and race attributes, making it extremely difficult for models to maintain performance across the different groups. The under-represented classes are reduced to even fewer samples per class when split across gender or race, making it even more difficult for these models to cope with data imbalances. CL-based methods are able to balance learning across the different gender and race groups, trading-off higher individual F1-Scores for *fairness*. For gender evaluations, EWC offers a competitive average F1-Score of 0.542, with the state-of-the-art evaluation at 0.645 [37] using spatio-temporal features, while also being the *fairest*. For race evaluations, the non-CL-based DA approach performs the best (without data augmentation) on fairness scores, tied with SI, despite achieving relatively low individual F1-Scores. Although, individual F1-Scores increase for DA when using data-augmentation, achieving the best average F1-Score of 0.584. The additional data allows the model to learn disentangle features enhancing performance on individual groups, however, it has a relatively lower *fairness* score compared to CL-based methods. Similar to race evaluations on the RAF-DB dataset, the SI approach remains the fairest, both with and without data-augmentation, prioritising balancing learning across the different race groups instead of maximising individual performance. Owing to the highly imbalanced class-distributions, the performance of all models is poor for under-represented classes such as AU $1, 2$ and 4, across all gender and race splits. On the other hand, the highest model performances are achieved for dominant classes such as AUs 10 and 12. These results are in line with other AU prediction approaches [37], [59], [60] that report similar differences in performance across these AUs.

Domain-Ordering: Due to the multi-label settings, all classes are learnt together with no ordering of the classes required. Furthermore, domain-ordering, that is, in which order the gender and race domains should be learnt, does not have any significant effect on model performance for the CL methods. Results from the gender and race-ordering experiments, reporting average model performance are provided in Tables $9 - 10$ of the supplementary material. For the race orderings, potentially 4 factorial ($4! = 24$) orderings are possible and evaluating all of them was intractable, thus, we select 2 different orderings, once staring with the domain with lowest number of test subjects (black) and once with the highest (white). No substantial difference can be concluded in model performances across the different race or gender domain orderings.

6.3 Limitations of CL-based Bias Mitigation

Our benchmark experiments with the RAF-DB and BP4D datasets highlight the potential of CL-based models for creating *fairer* facial affect analyses systems. CL-based models outperform other bias mitigation strategies for evaluations across gender and race domains, managing shifts in data distributions well. However, more work is needed to optimise CL-based models for multi-label settings where they under-perform on F1-scores(see Table 7 and 9). We perform hyper-parameter optimisation for the different CL-methods used, exploring the regularisation coefficients used in these approaches. Yet, more work is needed to develop customised algorithms that can tackle multi-label classification. Recent work by Kim et al. [61] proposes a new replay-based strategy, the Partitioning Reservoir Sampling (PRS), that aims to tackle CL for multi-label classification, balancing both intra- and inter-task imbalances. Yet, they benchmark their approach on classification settings with little-to-no overlap between the tasks. This is not the case for AU detection where the different domains, as well as the classes within each domain, share feature representations, making it even harder for the models.

Furthermore, as regularisation-based CL models assign *importance* to different parameters based on their contribution towards previously learnt tasks, shared feature representations make it harder for models to incrementally learn different tasks or domains as model parameters may contribute to more than one task or domain. Rehearsal-based methods such as NR, on the other hand, require the models to physically store seen samples from previous tasks, interleaving them with new data to maintain performance. As the number of tasks, or in the case of Domain-IL, data-splits across domains such as gender or race increase, storing samples from all the domains becomes extremely expensive both in terms of its memory footprint as well as the computational power needed to train the algorithms.

Additionally, as the tasks increase, models may experience saturation [62] requiring stronger regularisation in the models to be able to preserve past knowledge [63]. The performance of the models also takes a hit where the model needs to re-prioritise whether to give more importance to the new task or remembering previous tasks. We see this in race-wise splits for both the datasets (see Table 5 and 9) where regularisation-based models focus on attaining higher performance scores for the last split, impacting the fairness scores for most models.

7 Conclusion and Future Work

In this work, we propose the novel use of Domain Incremental CL as a potent bias mitigation method for facial affect analysis tasks. In particular, we highlight how using Domain-IL settings, regularisation-based CL methods can help develop *fairer* expression recognition and AU detection algorithms. Our experiments with popular benchmark datasets, RAF-DB for expression recognition and BP4D for AU detection, showcase the superlative performance of CL methods at handling imbalances in data distributions with respect to demographic attributes of gender and race, while offering competitive model performance. In comparison with state-of-the-art bias mitigation approaches, these methods are able to balance learning across different domain splits, not only achieving high accuracy scores but also maintaining fairness across the different splits.

Yet, this proof-of-concept evaluation was limited to regularisation-based methods only and hence further experimentation is needed to fully understand the benefits of using CL as an effective bias mitigation strategy for facial expression and action unit recognition tasks. With harder problems, as in the case of multi-label AU detection, we see that even though most regularisation-based methods achieve high accuracy, they do so by sacrificing fairness across different domain attributes. While a simplistic and naive rehearsal mechanism is able to improve model performance, our future work will aim to investigate other, more complex, pseudo-rehearsal [47], [48], [61], [63] or neuro-inspired [62], [64], [65] methods for bias mitigation.

Acknowledgement

For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

Funding: N. Churamani is funded by the EPSRC under grant EP/R513180/1 (ref. 2107412). H. Gunes' work is supported by the EPSRC under grant ref. EP/R030782/1 and partially by the European Union's Horizon 2020 Research and Innovation program project WorkingAge under grant agreement No. 826232. O. Kara contributed to this work during a summer research study at the Department of Computer Science and Technology, University of Cambridge.

Data Access Statement: This study involved secondary analyses of pre-existing datasets. All datasets are described in the text and cited accordingly. Licensing restrictions prevent sharing of the datasets. The authors thank Prof Lijun Yin from the Binghamton University (USA) for providing access to the BP4D Dataset and relevant race attributes; and Shan Li, Prof Weihong Deng and JunPing Du from the Beijing University of Posts and Telecommunications (China) for providing access to RAF-DB.

Code Access: Code for the experiments to be made available at https://github.com/nchuramani/CLForBiasMitigation.

References

[1] S. Feldstein, "The Global Expansion of AI Surveillance," Carnegie Endowment for International Peace, Tech. Rep., 2019.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[3] S. Dey, B. R. Duff, N. Chhaya, W. Fu, V. Swaminathan, and K. Karahalios, "Recommendation for Video Advertisements Based on Personality Traits and Companion Content," in *25th International Conference on Intelligent User Interfaces*. ACM, 2020, p. 144–154.

[4] D. Roselli, J. Matthews, and N. Talagala, "Managing Bias in AI," in *Companion Proceedings of The 2019 World Wide Web Conference*. Association for Computing Machinery, May 2019, p. 539–544.

[5] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2017, pp. 1–7.

[6] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, Oct 2017.

[7] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, June 2015.

[8] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: a survey," *IEEE Transactions on Affective Computing*, June 2017.

[9] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, 2020.

[10] P. Ekman and W. V. Friesen, *Facial action coding systems*. Consulting Psychologists Press, 1978.

[11] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[12] P. Ekman, "Darwin's contributions to our understanding of emotional expressions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3449–3451, Dec. 2009.

[13] S. Li and W. Deng, "A Deeper Look at Facial Expression Dataset Bias," *IEEE Transactions on Affective Computing*, 2020.

[14] S. Yucer, S. Akcay, N. A. Moubayed, and T. Breckon, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation." in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[15] J. Cheong, S. Kalkan, and H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing: Overview and techniques," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 39–49, 2021.

[16] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *17th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'01, 2001, p. 973–978.

[17] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment," in *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 725–740.

[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through Awareness," in *3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. ACM, 2012, p. 214–226.

[19] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8919–8928.

[20] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[21] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and autonomous systems*, vol. 15, no. 1-2, pp. 25–46, 1995.

[22] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information Fusion*, vol. 58, pp. 52–68, 2020.

[23] G. M. van de Ven and A. S. Tolias, "Three scenarios for continual learning," *arXiv preprint arXiv:1904.07734*, 2019.

[24] H. Tajfel and J. Turner, "An integrative theory of intergroup conflict," *The Social Psychology of Inter-group Relations*, 1979.

[25] M. Hewstone, M. Rubin, and H. Willis, "Intergroup Bias," *Annual Review of Psychology*, vol. 53, no. 1, pp. 575–604, 2002.

[26] S. D. Preston, "A perception-action model for empathy," *Empathy in mental illness*, vol. 1, pp. 428–447, 2007.

[27] J. N. Gutsell and M. Inzlicht, "Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups," *Journal of experimental social psychology*, vol. 46, no. 5, pp. 841–845, 2010.

[28] S. L. Sporer, "Recognizing faces of other ethnic groups: An integration of theories." *Psychology, Public Policy, and Law*, vol. 7, no. 1, pp. 36–97, Mar. 2001.

[29] H. A. Elfenbein and N. Ambady, "Is there an in-group advantage in emotion recognition?" *Psychological Bulletin*, vol. 128, no. 2, pp. 243–249, 2002.

[30] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 77–91.

[31] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, "Face Recognition Performance: Role of Demographic Information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[32] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[33] V. Iosifidis and E. Ntoutsi, "Dealing with bias via data augmentation in supervised learning scenarios," in *International Workshop on Bias in Information, Algorithms, and Systems (BIAS)*. CEUR Workshop Proceedings, 2018, pp. 24–29.

[34] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 68–81, 2019.

[35] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using Synthetic Data to Improve Facial Expression Analysis with 3D Convolutional Networks," in *IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1609–1618.

[36] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, vol. 89, pp. 385–397, 2015.

[37] N. Churamani, S. Kalkan, and H. Gunes, "AULA-Caps: Lifecycle-Aware Capsule Networks for Spatio-Temporal Analysis of Facial Actions," in *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021.

[38] T. Xu, J. White, S. Kalkan, and H. Gunes, “Investigating Bias and Fairness in Facial Expression Recognition,” in *Computer Vision – ECCV 2020 Workshops*. Springer International Publishing, 2020, pp. 506–523.

[39] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, “Exploring Disentangled Feature Representation Beyond Face Identification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, “Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 2269–2277.

[41] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, “PathNet: Evolution Channels Gradient Descent in Super Neural Networks,” *arXiv: Neural and Evolutionary Computing*, 2017.

[42] Z. Li and D. Hoiem, “Learning without forgetting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[43] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[44] F. Zenke, B. Poole, and S. Ganguli, “Continual learning through synaptic intelligence,” *Proceedings of Machine Learning Research*, vol. 70, p. 3987, 2017.

[45] A. Robins, “Catastrophic forgetting in neural networks: the role of rehearsal mechanisms,” in *Proc. The 1st New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, Nov 1993, pp. 65–68.

[46] A. Robins, “Catastrophic Forgetting, Rehearsal and Pseudorehearsal,” *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[47] H. Shin, J. K. Lee, J. Kim, and J. Kim, “Continual learning with deep generative replay,” in *Advances in Neural Information Processing Systems*, 2017, pp. 2990–2999.

[48] N. Churamani and H. Gunes, “CLIFER: Continual Learning with Imagination for Facial Expression Recognition,” in *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020, pp. 322–328.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[50] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, “Memory aware synapses: Learning what (not) to forget,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[52] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, “Progress & Compress: A Scalable Framework for Continual Learning,” in *International Conference on Machine Learning*, 2018, pp. 4528–4537.

[53] Y.-C. Hsu, Y.-C. Liu, A. Ramasamy, and Z. Kira, “Re-evaluating continual learning scenarios: A categorization and case for strong baselines,” in *Workshop on Continual Learning, NeurIPS*, 2018.

[54] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2852–2861.

[55] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, “BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database,” *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014, best of Automatic Face and Gesture Recognition 2013.

[56] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[57] N. Díaz-Rodríguez, V. Lomonaco, D. Filliat, and D. Maltoni, “Don’t forget, there is more than forgetting: new metrics for Continual Learning,” in *Workshop on Continual Learning, NeurIPS*, 2018.

[58] Z. Wen, W. Lin, T. Wang, and G. Xu, “Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition,” *arXiv preprint arXiv:2109.07270*, 2021.

[59] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, “Semantic relationships guided representation learning for facial action unit recognition,” in *AAAI Conference on Artificial Intelligence*, 2019, pp. 8594–8601.

[60] Z. Shao, L. Zou, J. Cai, Y. Wu, and L. Ma, “Spatio-Temporal Relation and Attention Learning for Facial Action Unit Detection,” *arXiv preprint arXiv:2001.01168*, 2020.

[61] C. D. Kim, J. Jeong, and G. Kim, “Imbalanced continual learning with partitioning reservoir sampling,” in *16th European Conference on Computer Vision (ECCV)*. Springer-Verlag, 2020, pp. 411—428.

[62] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, “Lifelong Learning of Spatiotemporal Representations With Dual-Memory Recurrent Self-Organization,” *Frontiers in Neurorobotics*, vol. 12, p. 78, 2018.

[63] M. K. Titsias, J. Schwarz, A. G. de G. Matthews, R. Pascanu, and Y. W. Teh, “Functional Regularisation for Continual Learning with Gaussian Processes,” in *International Conference on Learning Representations (ICLR)*, 2020.

[64] R. Kemker and C. Kanan, “FearNet: Brain-Inspired Model for Incremental Learning,” *CoRR*, vol. abs/1711.10563, 2018.

[65] N. Kamra, U. Gupta, and Y. Liu, “Deep generative dual memory network for continual learning,” *CoRR*, vol. abs/1710.10368, 2017.

Nikhil Churamani received his MSc degree with distinction in Intelligent and Adaptive Systems from Universität Hamburg in 2018. He is currently a doctoral student at the Affective Robotics and Intelligence (AFAR) Lab at the Department of Computer Science and Technology, University of Cambridge, UK. His research interests include Affective Computing, Continual Learning, Computer Vision, Deep Learning and Human-Robot Interaction. His current research focuses on Continual Learning for Affective Robotics investigating lifelong and continual learning of affect in social robots focused on Human-Robot Interaction and affect-driven learning.

Ozgur Kara is an undergraduate student in Electrical-Electronics Engineering Department at the Bogazici University and will graduate in 2022. He will start his doctoral studies at Machine Learning Ph.D program, Georgia Institute of Technology, Atlanta, USA, in 2022. He has a strong interest primarily in the fields of Computer Vision, Machine Learning, Continual Learning, and Controllable & Explainable ML systems.

Hatice Gunes (Senior Member, IEEE) received the Ph.D degree in computer science from the University of Technology Sydney, NSW, Australia. She is a Professor with the Department of Computer Science and Technology, University of Cambridge, UK, leading the Affective Intelligence and Robotics Lab. Her expertise is in the areas of affective computing and social signal processing cross-fertilising research in human behaviour understanding, computer vision, signal processing, machine learning, and social robotics. She has published over 125 papers in the above areas. Prof Gunes is the former President (2017-2019) of the Association for the Advancement of Affective Computing, was the General Co-Chair of ACII 2019, and Program Co-Chair of ACM/IEEE HRI 2020 and IEEE FG 2017. Her research has been supported by various competitive grants, with funding from the Engineering and Physical Sciences Research Council, UK (EPSRC), Innovate UK, British Council, Alan Turing Institute and EU Horizon 2020. She is a Fellow of the EPSRC, a Staff Fellow of Trinity Hall Cambridge, and was a Faculty Fellow of the Alan Turing Institute.