Article

Not peer-reviewed version

Mitigating Aging Bias in Facial Expression Recognition through Subgroup-Aware Data Augmentation

Zhengke Gao , Yan Fu * , Bing Ye , Le Chang , Qiyuan Zhu , Yancheng Liu , Alex Mihailidis

Posted Date: 1 June 2026

doi: 10.20944/preprints202605.2112.v1

Keywords: facial expression recognition; aging bias; data augmentation; affective computing

 doi:10.20944/preprints202605.2112.v1

Article

Mitigating Aging Bias in Facial Expression Recognition through Subgroup-Aware Data Augmentation

Zhengke Gao [1], Yan Fu [1,*], Bing Ye [2], Le Chang [3], Qiyuan Zhu [1], Yancheng Liu [1] and Alex Mihailidis [2,4]

[1] School of Mechanical Science and Engineering, Huazhong University of Science and Technology , WuHan, 430074, China

[2] KITE, Toronto Rehabilitation Institute, Toronto, ON, Canada

[3] Institute of Communication, Culture, Information and Technology, University of Toronto, Toronto, ON, Canada

[4] Institute of Biomedical Engineering, University of Toronto, Toronto, ON, Canada

* Correspondence: laura_fy@mail.hust.edu.cn

Abstract

With the accelerating pace of global population aging, emotion-aware technologies have become increasingly important for improving the quality of life and psychological well-being of older adults. However, most facial expression recognition (FER) systems exhibit substantial performance degradation among elderly users due to the lack of age-diverse data and inadequate model adaptation. This study investigates age-related bias in FER and proposes a subgroup-aware data augmentation framework to enhance recognition robustness for older populations. We first retrain a ResNet-50–based age estimation model using the UTK-Face dataset to provide reliable age annotations for three benchmark FER datasets: RAF-DB, AffectNet, and ExpW. Subsequently, we introduce an age-adaptive augmentation strategy that applies stronger transformations to elderly facial images while maintaining moderate augmentation for younger ones. Experimental results demonstrate that the proposed approach significantly improves recognition accuracy and generalization in elderly subgroups without sacrificing performance in younger populations. This work provides a practical and scalable pathway toward age-inclusive affective computing, highlighting the importance of integrating demo-graphic priors into data processing pipelines for fair and trustworthy emotion recognition systems.

Keywords: facial expression recognition; aging bias; data augmentation; affective computing

1. Introduction

The rapid progression of global population aging has brought increasing attention to the emotional and psychological well-being of older adults. By 2030, it's estimated that the global population aged 60 and above will reach 1.4 billion, accounting for 16.7 percent of the total population. By 2050, the number will even increase to 2.1 billion [1]. With the growing prevalence of older adults living alone or in "empty-nest" households [2], the eldercare system faces new challenges: beyond meeting physical and medical needs, it must also address emotional support and psychological well-being [3–5]. As smart home technologies and companion robots [6–10]become increasingly integrated into everyday life, emotion-aware interaction has emerged as a crucial component of high-quality companionship. Intelligent systems must accurately perceive the emotional states of older users in order to trigger appropriate responses—such as providing comfort, issuing reminders, or notifying family members—thereby enhancing both quality of life and perceived security.

 doi:10.20944/preprints202605.2112.v1

Existing emotion recognition techniques in clinical use can be broadly categorized into four modalities: speech, psychological questionnaires, physiological signals, and facial expressions [11]. Speech-based methods rely on verbal behavior and are sensitive to background noise [12][13]. Questionnaires, while accurate, require professional administration and active user participation [14]. Physiological approaches are objective and reliable but often depend on contact-based sensors, limiting their suitability for long-term home use [15]. By contrast, facial expression analysis offers a non-intrusive, real-time, and easily deployable solution, making it the predominant approach in home-based companion systems [16].

Recent advances in deep learning have significantly enhanced the performance of facial expression recognition (FER). Convolutional neural networks (CNNs) and vision transformers (ViTs) achieve accuracies exceeding 90% on widely used benchmark datasets [17–19]. However, these advances have not translated well to older populations. A major limitation lies in dataset composition: most public FER datasets lack explicit age annotations and contain fewer than 10% older adult samples [20]. Prior studies have further demonstrated that commercial emotion recognition systems exhibit substantial performance disparities across age groups [21], underscoring the insufficient adaptability of existing models for older users.

Findings from psychology and cognitive science provide critical insights into this issue. Meta-analyses have shown that older adults are particularly impaired in recognizing negative emotions such as anger, sadness, and fear [22][23]. Moreover, their interpretation of facial expressions relies more heavily on specific facial regions, notably the eyes and mouth corners [24]—regions that current deep learning models often fail to capture robustly. These observations suggest that AI-based emotion recognition systems, if they neglect class imbalance and region-specific cues in older populations, are prone to performance degradation.

To address this challenge, this study proposes a subgroup-aware data augmentation framework that explicitly integrates age as a demographic prior into the facial expression recognition pipeline. Unlike conventional FER approaches that implicitly assume age-invariant facial representations or rely on uniform augmentation strategies, the proposed method directly accounts for age-dependent data scarcity and facial characteristics at the data-processing stage. By systematically differentiating augmentation intensity across age groups, the framework targets the structural causes of age-related performance degradation rather than treating it as a marginal optimization problem. This perspective elevates age bias from a secondary fairness concern to a core reliability issue, enabling emotion recognition systems to function more consistently and safely in real-world eldercare and human–machine interaction scenarios.

The effectiveness of the proposed approach is supported by a large-scale and multi-stage experimental design. This work involves retraining an age estimation model to improve labeling reliability, generating pseudo-age annotations for three widely used FER datasets (RAF-DB, AffectNet, and ExpW), and conducting extensive quantitative analyses of age–emotion distributions. Multiple controlled experiments are performed under identical network architectures and training protocols, including comparisons between age-biased and age-balanced subsets across five age groups. Model performance is evaluated not only in terms of mean accuracy but also stability and consistency through cross-validation, variance analysis, and feature-space visualization. This comprehensive experimental pipeline reflects a substantial empirical effort and enables a rigorous assessment of age-related bias and the proposed mitigation strategy.

2. Related Work

FER aims to automatically identify human emotions from facial cues and has achieved impressive accuracy on standard benchmarks through deep learning. However, despite technical advances, FER systems still exhibit significant performance disparities across demographic groups, particularly for older adults [25–27]. These disparities arise from both data- and model-level factors, resulting in a persistent technological gap between algorithmic progress and inclusive real-world performance [28–31].

 doi:10.20944/preprints202605.2112.v1

At the dataset level, most mainstream FER datasets are strongly biased toward younger adults. AffectNet, one of the largest FER datasets, contains fewer than 10% samples of older adults, and datasets such as FER2013 and The Real-world Affective Faces Database (RAF-DB) provide no explicit age annotations [10,19]. This underrepresentation leads to class imbalance and limits model generalizability to aging faces [20,32]. Additionally, the prevalence of low-quality, web-sourced images further introduces label noise and visual inconsistencies, amplifying the bias. Consequently, models trained on these datasets tend to underperform on older adults, revealing a data insufficiency gap that undermines cross-age adaptability.

Beyond data limitations, existing algorithms also fail to adapt effectively to age-related facial variations. Traditional feature-engineering approaches, such as LBP or HOG combined with SVM, lack robustness to wrinkles, sagging, and texture changes that accompany aging. Deep learning architectures—including ResNet, VGG, and Vision Transformers—achieve high overall accuracy [17][18], yet their performance still degrades substantially on older faces [19]. This decline stems from both biased feature learning and insufficient sensitivity to age-specific cues—for example, older adults rely more on eye and mouth corners when interpreting emotions [24], but current models often fail to capture these localized features robustly.

In response, researchers have attempted to mitigate demographic bias including gender, age and race, through improved data augmentation, re-weighting, and domain adaptation strategies [33][34]. While these methods modestly enhance fairness, they are typically developed in isolation and not tailored to specific subgroups such as older adults. Moreover, psychological studies suggest that age influences not only expression intensity but also region-specific salience [22][23]. Existing deep learning frameworks rarely integrate these human-factor insights, leaving a model adaptation gap in addressing the unique appearance and perceptual patterns of aging faces.

Collectively, the above limitations highlight two intertwined technological gaps: (1) the uneven distribution of datasets reflecting the insufficient diversity and annotation of elderly samples; (2) the absence of mechanisms that adapt to age-dependent facial features and perceptual cues.

To bridge these gaps, this study introduces a subgroup-aware data augmentation framework that incorporates demographic priors—specifically age—into the FER training process. By adjusting augmentation intensity and feature emphasis according to age group, the proposed method compensates for data scarcity among elderly samples and promotes feature learning aligned with their unique facial characteristics. This approach represents a practical step toward age-inclusive and fair affective computing, ensuring that emotion recognition systems remain robust across the full human lifespan.

3. Materials and Methods

3.1. Datasets and Age Groups

To investigate the effect of age-related bias in FER, this study focuses on three widely used datasets that include elderly samples: RAF-DB, AffectNet, and Expression in the Wild (ExpW).

RAF-DB [35] is a widely adopted benchmark for facial expression research, developed by researchers at the University of British Columbia. It contains 16,339 real-world facial images, each annotated by professional emotion analysts with seven basic emotion categories: happy, surprise, sad, fear, disgust, angry and neutral, covering the primary spectrum of human affective expressions.

AffectNet [36] is currently the largest facial expression dataset available. It comprises over one million images collected via three major search engines, including both manually and automatically annotated samples. Each image is labeled under both discrete emotion categories and the valence–arousal dimensional model. In this study, only the manually annotated subset containing 283,901 images across seven basic emotions was utilized to ensure label reliability and consistency.

ExpW [37] is another large-scale facial expression dataset commonly used in computer vision research. It includes 74,532 real-world facial images collected from Google search results, each

 doi:10.20944/preprints202605.2112.v1

categorized into seven basic emotions. Notably, ExpW exhibits a higher proportion of positive expressions, reflecting the natural distribution of emotional expressions in everyday scenarios.

Together, these datasets not only provide large-scale, in-the-wild facial expression samples but also encompass a wide range of age groups, making them suitable for analyzing age-related disparities in facial expression recognition. To enable quantitative evaluation of such disparities, this study further categorizes individuals into meaningful age groups that reflect both biological and psychosocial developmental stages [38–40].

Specifically, the age spectrum is divided into five life-course periods based on widely accepted gerontological and psychological frameworks: Childhood and Adolescence (0-17 years), Emerging and Early Adulthood (18-39 years), Middle Adulthood (40-64 years), and Late Adulthood, which is further subdivided into Young-old (65-84 years) and Oldest-old (85+ years). The distinction between the Young-old and Oldest-old primarily lies in health status and independence. While the Young-old are generally healthy, productive, and capable of independent living, the Oldest-old often experience chronic health issues requiring long-term or assisted care. These physiological and psychological differences are known to influence emotional expression and perception, underscoring the need to examine model behavior across distinct age cohorts [41][42].

3.2. Age Labeling

After determining the materials, we performed age labeling on the aforementioned three datasets using DeepFace [43], a multi-attribute facial analysis framework that incorporates an age estimation module. The module is originally trained on the IMDB-WIKI dataset [44], which is composed of celebrity images. While IMDB-WIKI provides precise chronological age labels, its reliance on actors' portraits raises concerns of bias: professional actors often appear younger than their actual age, potentially leading to systematic underestimation in age recognition.

To evaluate this limitation, we further validated the DeepFace age module on the UTK Face dataset [45], which is more demographically balanced and includes a wide range of ages (0–116 years) with annotations for gender, race, and facial landmarks. Results revealed that the mean absolute error (MAE) of DeepFace age estimation reached 10.89 years on UTK Face, much higher than the 4.65 years reported on IMDB-WIKI. This discrepancy confirmed that IMDB-WIKI's domain bias significantly affects age prediction.

To address this issue, we retrained the age recognition model on UTK Face using a more robust ResNet-50 backbone [17]. Compared to the VGG-16 architecture adopted in DeepFace, ResNet-50 employs residual learning and batch normalization, which effectively alleviate vanishing gradient problems and improve convergence stability during training [46][47]. Moreover, its deeper yet more efficient design enables stronger feature representation and better generalization to in-the-wild facial variations. These properties make ResNet-50 particularly suitable for age estimation tasks that involve substantial inter- and intra-class variability [48]. We conducted multiple training trials, fine-tuned hyperparameters, and selected the model with the best performance. This retrained model was subsequently used for assigning age labels in the FER experiments, with detailed training configurations described in the following experimental setup section.

3.3. Data Augmentation Method

To mitigate class imbalance across age and emotion groups, we employed a two-level augmentation strategy. For individuals under 65 years, standard augmentation techniques such as random cropping, horizontal flipping, and color jitter were applied. For individuals aged 65 and above, stronger augmentation strategies were used to increase the effective diversity of limited elderly samples, including random rotation (±20°), Gaussian noise, and contrast adjustments. This design aimed to compensate for the underrepresentation of older adults while avoiding overfitting.

 doi:10.20944/preprints202605.2112.v1

4. Results

4.1. Age Prediction Model

The UTK Face dataset, consisting of approximately 24,000 facial images, was di-vided into training and testing subsets. To balance the age distribution, no more than 200 samples were retained for each age in the training set, resulting in 12,000 training images and 3,000 testing images. This sampling strategy ensured representative evaluation while mitigating the inherent skew toward younger age groups.

The original DeepFace age estimation module achieved a mean absolute error (MAE) of 10.89 years on the UTK Face dataset. A more detailed analysis, as illustrated in Figure 1, reveals distinct variations in prediction accuracy across age intervals (grouped in 10-year bins). The model performed best in the 30–39 age range, with an MAE of ±4.3 years, fol-lowed by the 20–29 range at ±4.5 years. However, the prediction error increased sub-stantially toward both younger and older ends of the spectrum: MAE reached ±22.2 years for the 0–9 group, ±27.0 years for the 70–79 group, and peaked at ±40.2 years for the 90–99 group.

A comparison between mean predicted and mean actual ages further revealed a consistent trend of bias. For samples under 30 years old, the model tended to overestimate age: in the 20–29 age group, the mean predicted age was 29.3 years, slightly higher than the true mean of 25.4 years. In contrast, for individuals aged 30 and above, the model systematically underestimated age, with the deviation widening as age increased—from 1.7 years lower in the 30–39 group to over 40 years lower in the 90–99 group.

These results indicate that the DeepFace model trained on the IMDB-WIKI dataset exhibits optimal performance for adults aged 20–39 years, but its accuracy declines sharply for both younger children and older adults. The trend highlights the domain bias of the IMDB-WIKI dataset, where training samples are dominated by celebrity portraits with limited age diversity, leading to systematic underestimation of advanced-age faces and overestimation of youthful ones.

We retrained a new age estimation model based on ResNet-50 using the UTK Face dataset to improve robustness across diverse age groups. The model was fine-tuned by unfreezing the last convolutional block (Layer 4) and the fully connected layer. Learning rates of 1×10^{-4} and 1×10^{-3} were assigned to these layers, respectively. To better capture the ordinal nature of age, the regression task was reformulated into 99 binary classification subtasks, following an ordinal regression framework.

The results showed a substantial improvement in age estimation performance under this training configuration, as illustrated in Figure 2. The final model achieved a MAE of 7.00 years on the test set, representing a significant enhancement compared to the un-modified DeepFace baseline. Across the 0–89-year range, the MAE remained consistently below ±10 years, with only the 90–99-year interval exceeding this threshold at ±16.4 years.

A comparison between the mean predicted and mean actual ages revealed that the new model demonstrated markedly higher sensitivity to age variations. Despite re-training the overall pattern of slight overestimation at younger ages and underestimation at older ages, the deviations were significantly reduced. Specifically, the model tended to overpredict for individuals aged 0–49 years, whereas for those aged 50 years and above, the predictions became slightly lower than the ground-truth values.

 doi:10.20944/preprints202605.2112.v1

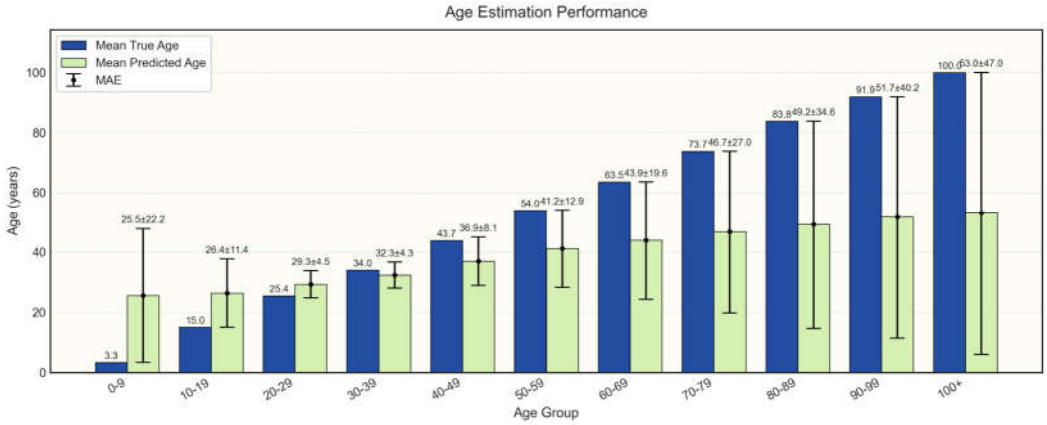

Figure 1. Performance of the DeepFace Age Estimation Module on the UTK Face Dataset.

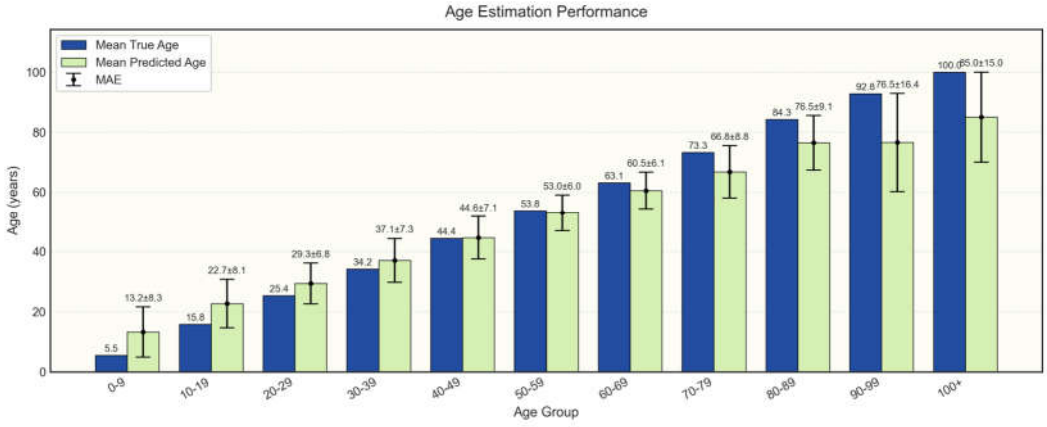

Figure 2. Performance of the re-trained ResNet-50 age estimation model.

4.2. Age–Emotion Distribution of FER Datasets

Based on this improved age recognition model, we subsequently generated pseudo-age labels for the three facial expression datasets—RAF-DB, AffectNet, and ExpW—and analyzed their emotion-age distributions.

The statistical analysis was conducted using age pseudo labels predicted by the re-trained ResNet-50 age recognition model. The results in Figure 3 indicate that across all three datasets, the vast majority of samples fall within the 18 and 64 year range, with emotions predominantly classified as happy and neutral. In the RAF-DB dataset (16k images in total), no samples above 85 years were found, and only fewer than 100 samples appeared in the 65 and 85 year range. Similarly, the emotion categories were dominated by happiness and neutrality.

Although AffectNet and ExpW contain substantially larger total image counts and exhibit broader age coverage, the representation of elderly individuals remains limited. The emotion distributions in both datasets are likewise skewed toward positive and neutral expressions, reflecting the intrinsic bias of web-collected facial datasets and further underscoring the necessity of age-balanced data for fair emotion recognition.

 doi:10.20944/preprints202605.2112.v1

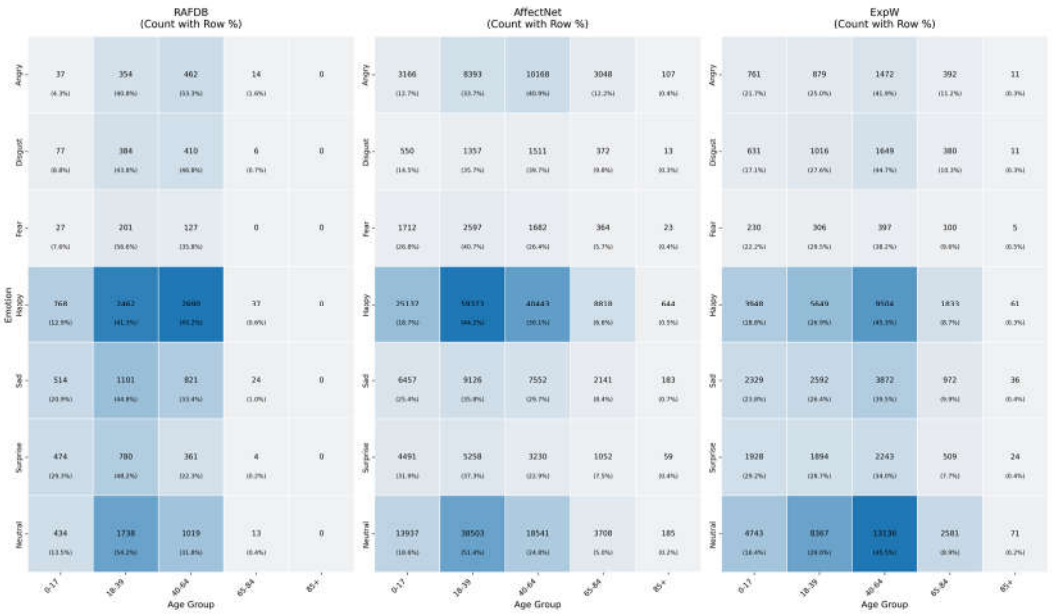

Figure 3. Age-emotion distribution across datasets: (a) RAFDB, (b)AffectNet and (c)ExpW. Across all datasets, samples are predominantly concentrated in the 18–64 years age range, while elderly samples (65+ years) account for less than 10% of the total. Emotion distributions exhibit a consistent bias toward positive (happy) and neutral expressions, with negative emotions (e.g., fear, disgust) being significantly underrepresented—especially in the Oldest-old group. Notably, RAF-DB lacks samples aged 85+ years, and both AffectNet and ExpW show limited coverage of elderly age groups despite their larger overall sample sizes.

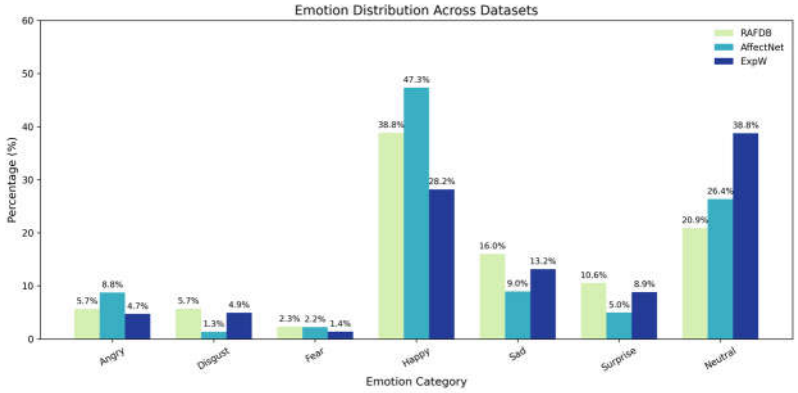

Figure 4. Emotion distribution across datasets: RAFDB, AffectNet and ExpW.

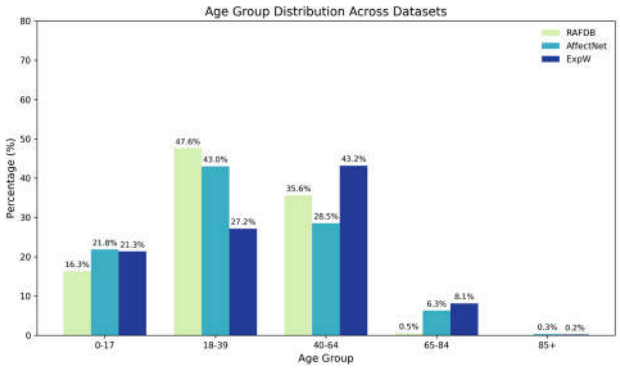

Figure 5. Age distribution across datasets: RAFDB, AffectNet and ExpW.

 doi:10.20944/preprints202605.2112.v1

4.3. Age–Emotion Distribution of FER Datasets

This section aims to validate the core hypothesis: imbalanced data distribution leads to insufficient reliability of model accuracy, where majority demographic groups are over-represented in model learning while minority groups are neglected. To this end, we design a controlled experiment on the AffectNet dataset, using two differently sampled subsets to fine-tune the same ImageNet-pretrained ResNet-50 model, and compare the accuracy reliability of the model under the two data settings.

We construct two subsets from the AffectNet dataset following strict sampling rules, ensuring that only the age distribution differs while other variables (emotion category ratio, image quality, annotation standard) remain consistent:

- Original Age Distribution Subset (OADS)

This subset retains the inherent age distribution bias of AffectNet. We perform random stratified sampling according to the original age group proportion, with the total sample size set to 15,000 to ensure sufficient data volume for ResNet-50 fine-tuning (thus avoiding overfitting), control computational costs, and maintain identical scales between the two differently distributed subsets to eliminate experimental biases caused by sample size discrepancies. Specifically, the 18–39 years old group (majority group) accounts for 43% of the samples, while the 65–84 years old (young-old) and 85+ years old (oldest-old) groups (minority groups) account for 6% and 0.3% respectively, consistent with the age distribution characteristics of the full AffectNet dataset.

- Balanced Age Distribution Subset (BADS)

To eliminate the impact of age imbalance, we adopt a relatively balanced sampling strategy for each age group, in which 3,000 images were sampled from each of the first four age groups, while all 973 images from the fifth age group were included. Thus, the subset also has a total sample size of 12,973, with the first four age groups adjusted to 23%, and the 85+ years old groups accounting for 7.5%. Stratified sampling is performed within each age group to maintain the original emotion category ratio, ensuring that the model's learning of emotional features is not affected by changes in emotion distribution.

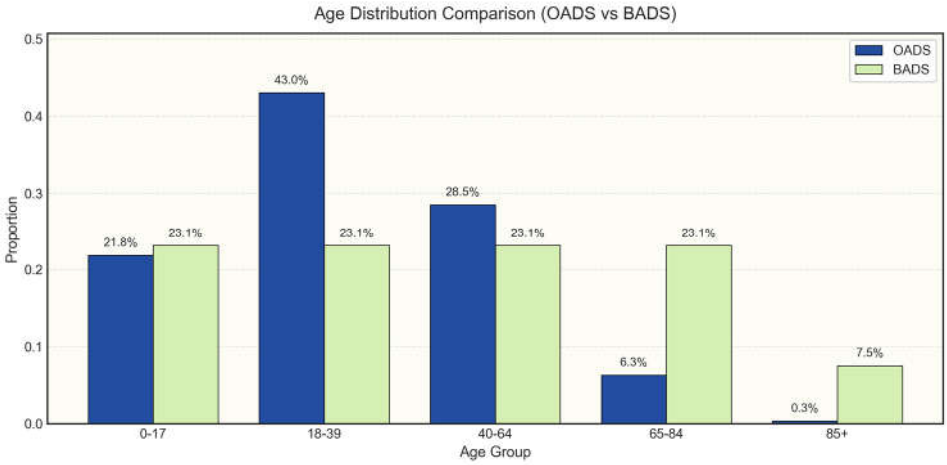

Figure 6. Age distribution Comparison of OADS and BADS.

A 5-fold cross-validation was conducted on the subset, using an ImageNet-pretrained ResNet-50 model with a layered fine-tuning strategy: all layers were initially frozen, followed by unfreezing of layer4 and the replaced fully connected layer (7 output classes for emotions), with distinct learning rates of 1e-4 and 1e-3 applied respectively. Training images underwent resizing to 224×224, random horizontal flipping, and ImageNet-standard normalization, while validation images used the same resizing and normalization without augmentation; both processes adopted a batch size of 64 and the

 doi:10.20944/preprints202605.2112.v1

Adam optimizer. The model was trained for a maximum of 50 epochs with early stopping (patience=5) triggered by validation accuracy, with cross-entropy loss as the optimization objective. For each fold, we recorded training loss and validation accuracy, calculated age-wise accuracy across the five age groups, and saved training logs, loss/accuracy curves, age-wise accuracy heatmaps, and the best-performing model weights for further analysis.

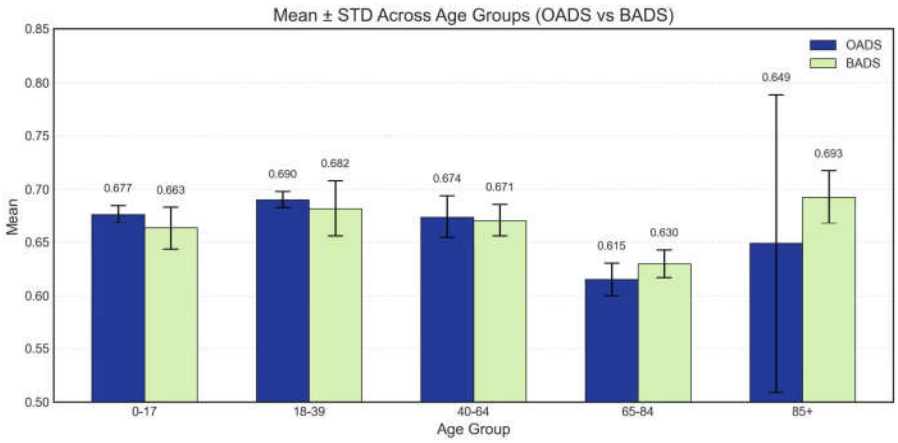

Figure 7. Age-wise Recognition Accuracy and Stability Comparison Between OADS and BADS. For majority age groups (0–17, 18–39, and 40–64), both datasets achieve comparable mean accuracies, indicating that age balancing does not compromise performance on well-represented groups. In contrast, BADS yields clear improvements for elderly groups, particularly for ages 85+, where higher mean accuracy and substantially reduced variance are observed.

Table 1 compares the age-wise emotion recognition performance of models trained on the OADS and the BADS. For majority age groups (0–17, 18–39, and 40–64), both datasets achieve comparable mean accuracies, with only minor differences within one standard deviation. This indicates that balancing the age distribution does not degrade the model's ability to learn discriminative features for well-represented age groups. For elderly age groups, BADS shows clear advantages. The mean accuracy for the 65–84 group increases from 0.615 to 0.630, accompanied by a slightly lower standard deviation. A more significant improvement is observed for the 85+ group, where BADS raises the mean accuracy from 0.649 to 0.693 while dramatically reducing the standard deviation from 0.140 to 0.025. The large standard deviation observed in OADS for elderly groups suggests unstable feature learning caused by severe data scarcity, where model performance varies strongly across folds. In contrast, the reduced variance in BADS indicates more consistent exposure to elderly samples during training, leading to more stable gradient updates and more robust age-specific feature representations.

Table 1. Results of the comparative experiment.

	OADS		BADS	
Age groups	**mean**	**std**	**mean**	**std**
0-17	0.677	0.008	0.663	0.020
18-39	0.690	0.008	0.682	0.026
40-64	0.674	0.020	0.671	0.015
65-84	0.615	0.015	0.630	0.013
85+	0.649	0.140	0.693	0.025

To further analyze how age distribution affects the learned feature representations, Figure 8 presents t-SNE visualizations of the penultimate-layer features for representative folds whose overall performance is closest to the cross-validation average. When trained on OADS, emotion features

 doi:10.20944/preprints202605.2112.v1

exhibit substantial overlap across categories, with several minority emotions being absorbed into dense regions dominated by majority classes. This phenomenon suggests that severe age imbalance encourages the model to learn overly averaged representations, resulting in unstable and less discriminative features for un-der-represented age groups. In contrast, models trained on BADS demonstrate a more structured feature space, where emotion clusters are more compact and locally separable. Although complete separation is not achieved—reflecting the inherent difficulty of facial expression recognition across age groups—the improved organization and reduced entanglement indicate more consistent feature learning. These qualitative observations are consistent with the reduced performance variance and improved mean accuracy for elderly age groups reported in Table 2, confirming that age-balanced training facilitates more robust and stable emotion representations.

Overall, these results verify that age distribution bias in FER datasets can undermine the reliability of model accuracy, particularly for under-represented age groups, thereby motivating further investigations into dataset processing strategies to alleviate age-related bias.

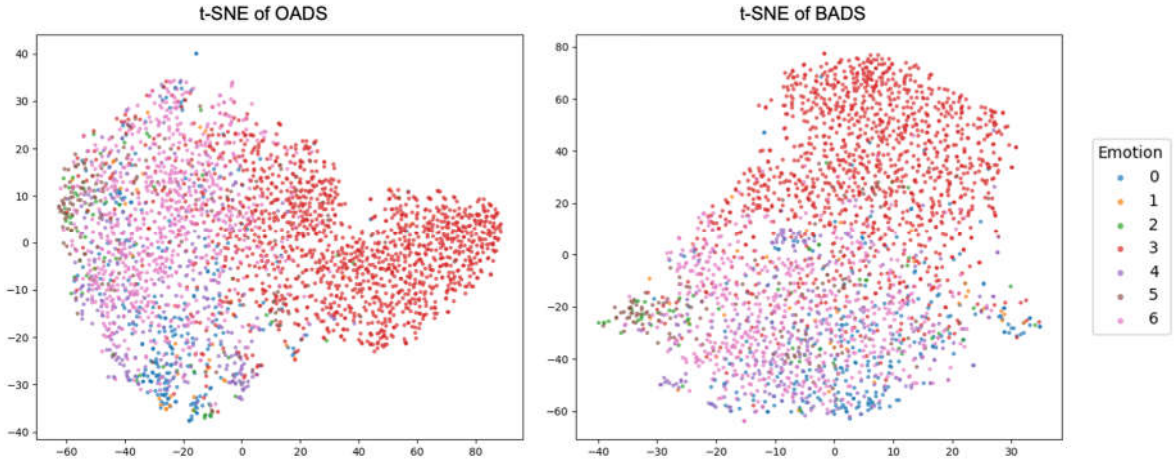

Figure 8. t-SNE visualization of emotion feature representations learned from the OADS and the BADS. The plots show representative folds whose overall performance is closest to the cross-validation average. Each point corresponds to one facial image, colored by emotion category. Compared with OADS (left), BADS (right) exhibits more compact and structured emotion clusters, indicating more stable and consistent feature learning under an age-balanced training scheme.

4.4. Baseline Models for Emotion Recognition

The baseline model serves as the benchmark for comparison to evaluate the effectiveness of the proposed method. Therefore, it must be reasonable, convincing, and represent a conventionally acceptable model. To establish a benchmark for emotion recognition performance across different age groups, we selected ResNet-50 as the baseline architecture due to its stable performance and relative fairness among various widely used deep learning architectures.

This model was first pre-trained on the ImageNet dataset and then fine-tuned on the RAF-DB, AffectNet, and ExpW datasets to ensure comparability with prior research. ImageNet is a seminal large-scale benchmark dataset in computer vision, consisting of over 14 million annotated images covering 1,000 core object categories (e.g., animals, daily objects, natural scenes), with labels structured based on the WordNet lexical semantic network. Its images span diverse scenarios, illumination conditions, and viewing angles, making it an ideal resource for learning universal visual features. The adoption of an ImageNet-pretrained model is justified by three key advantages. First, it enables generic visual feature transfer: during pre-training, the model acquires low-to-middle-level features (e.g., edges, textures, local contours) that are fundamental to all visual tasks, including FER. These features serve as a robust foundation for identifying emotion-related facial details such as eye wrinkles and mouth curvature, eliminating the need to learn such basic patterns from scratch on FER datasets. Second, it alleviates overfitting caused by limited and imbalanced FER data: datasets like

 doi:10.20944/preprints202605.2112.v1

RAF-DB and ExpW suffer from scarce samples in underrepresented age-emotion subgroups (e.g., 85+ years old with negative emotions), and pre-training leverages ImageNet's massive data scale to reduce the model's reliance on small FER subgroups. Third, it ensures fair baseline comparability: using an ImageNet-pretrained ResNet-50 aligns with standard practices in FER research, allowing direct and reliable performance comparisons with state-of-the-art methods.

The baseline model utilized the training set from RAF-DB and was improved using its test set. The original RAF-DB test set comprised 3,068 images but lacked images from the 85+ age group. Therefore, we supplemented it with the elderly segment from AffectNet and ExpW datasets to obtain a balanced test set. Specifically, we adopted a hierarchical unfrozen strategy: layer 4 unfrozen at the 10th epoch, layer 3 at the 15th epoch, and layer 2 at the 20th epoch. Concurrently, we employed a hierarchical learning rate, with 1e-4 for fully connected layers and 1e-5 for convolutional layers. Additionally, we set the batch size to 64 and the total number of epochs to 40.

4.5. Baseline Models for Emotion Recognition

To address the discrepancies observed, we have designed a subgroup-aware augmentation strategy that is tailored to the unique challenges posed by elderly facial expressions. Unlike traditional augmentation methods, which apply uniform transformations to all samples, this approach adjusts the intensity of the augmentation based on the age group of the sample.

First, we constructed a new balanced dataset for fine-tuning. This comprises 500 images selected from each age-emotion subgroup in AffectNet, supplemented with an additional 1,000 images from ExpW. We then employed a five-fold cross-validation approach, randomly partitioning the images into training, validation and test sets.

Specific settings included:

- Freezing scope: Unfreeze all layers in Layer 4; unfreeze the last two bottleneck layers in Layer 3; unfreeze all fully connected (FC) layers; freeze all other layers.
- Optimizer: AdamW;
- Heterogeneous learning rates: FC: 1e-3; Layer 4 and Layer 3: 1e-4; Weight decay: 1e-4; Epochs: 40. Batch size: 32.
- Early stopping was employed, with training halted if no improvement occurred over 5 epochs.

For the under-65 age group, we applied standard augmentation techniques, such as horizontal flipping, random cropping, and mild brightness/contrast adjustments, to avoid over-regularization. For the Young-old and the Oldest-old groups, we used enhanced augmentation techniques to more accurately capture variability in facial features. These specific measures included random rotation (±25°), elastic deformation to simulate differences in facial muscles, Gaussian blur to reproduce variability in image acquisition, and targeted contrast enhancement to highlight texture around the eyes and mouth.

This design stems from two key considerations. First, datasets containing elderly individuals are relatively scarce and often exhibit low intra-group diversity, limiting the ability of deep learning models to generalize across facial variations associated with aging. By applying stronger and more diverse augmentation strategies, data variability can be effectively expanded, thereby improving model robustness. Second, tailoring augmentation schemes to specific subgroup characteristics helps the model learn discriminative yet generalizable features, making it more adaptive to both young and elderly populations.

Empirical evaluations confirm that the proposed subgroup-aware augmentation substantially enhances recognition accuracy for elderly users while maintaining stable performance for younger individuals. These results demonstrate the efficacy of incorporating demographic prior information into the data processing pipeline and suggest that age-sensitive augmentation provides a practical path toward fair and inclusive emotion recognition systems.

As illustrated in Figure 9, the model trained on the hierarchically augmented dataset exhibits a marked improvement in facial expression recognition performance. The base-line model tended to

 doi:10.20944/preprints202605.2112.v1

classify most facial expressions as neutral, achieving an accuracy of 82% for the happy class, 56% for neutral, and 52% for surprise, while almost entirely failing to correctly identify negative emotions such as anger, disgust, fear, and sadness.

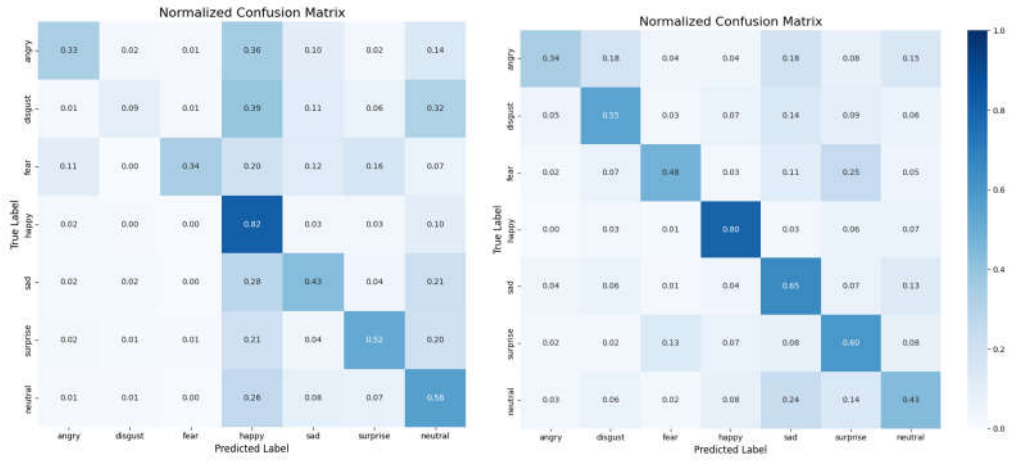

Figure 9. Normalized confusion matrix of baseline model (left) and promoted model (right). The baseline model shows a strong bias toward predicting neutral, achieving relatively high accuracy for majority emotions (e.g., happy, neutral, and surprise) while largely failing to recognize negative emotions such as anger, disgust, fear, and sadness. In contrast, the promoted model trained with subgroup-aware augmentation demonstrates substantially improved and more balanced recognition performance across all emotion categories, with notable gains for previously under-recognized negative emotions.

In contrast, the enhanced model trained with the subgroup-aware augmentation demonstrated a significant performance gain across all emotion categories. Specifically, happiness achieved the highest recognition accuracy of 80%, followed by sadness (65%) and surprise (60%). Other emotions also showed notable improvements compared to the baseline, indicating that the proposed approach effectively balanced recognition capabilities across both positive and negative emotions.

Figure 10 compares the emotion recognition accuracy of the baseline and the pro-moted models across joint emotion–age subgroups. For the baseline model, performance is highly uneven across both emotion categories and age ranges. While majority emotions such as happy and neutral achieve relatively high accuracy in younger and middle-aged groups, recognition performance degrades substantially for minority emotions, particularly anger, disgust, and fear. This degradation is most pronounced in elderly subgroups, where several emotion–age combinations exhibit near-zero accuracy, indicating severe instability and poor generalization caused by extreme data scarcity and biased feature learning. Although the highest accuracy is observed in the happy emotion for the 65–84 age group, this result is not statistically reliable, as the test set contains only a very small number of samples in this subgroup. In fact, for ages above 64, only the happy (65–84) and sad (65–84) categories include any test samples, while all other emotion–age combinations have zero samples. Therefore, the reported 100% accuracy is achieved on an extremely limited sample size and does not reflect the true recognition capability of the model. Consequently, the analysis in this figure primarily focuses on emotion–age groups below 64 years old, where sufficient test data are available for meaningful evaluation.

 doi:10.20944/preprints202605.2112.v1

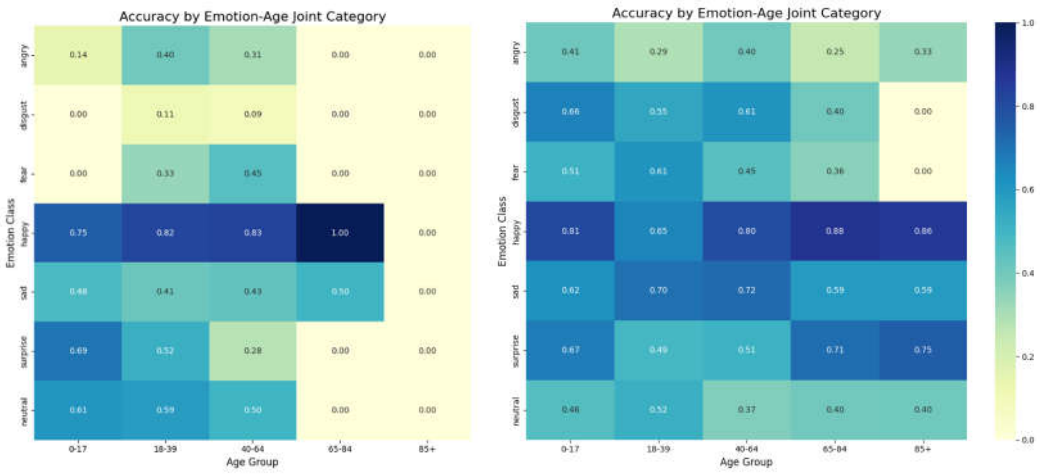

Figure 10. Age-emotion distribution across datasets of baseline model (left) and promoted model (right). The baseline model exhibits pronounced performance imbalance across both emotion categories and age ranges, with severe degradation for minority emotions (e.g., anger, disgust, and fear), particularly in elderly groups, where near-zero accuracy is observed due to extreme data scarcity. Although a peak accuracy appears in the happy category for ages 65–84, this result is statistically unreliable, as only a very limited number of test samples are available for this subgroup, while most other emotion–age combinations above 64 contain no samples. After applying subgroup-aware augmentation, the promoted model achieves consistent performance improvements across nearly all age ranges, notably enhancing recognition accuracy for elderly groups and negative emotions, while reducing variance across age subgroups. Overall, the results indicate improved robustness and fairness under balanced training.

After applying the subgroup-aware augmentation strategy, the promoted model demonstrates marked and consistent improvements across nearly all age ranges and emotion categories. Notably, recognition accuracy for elderly groups (65–84 and 85+) increases substantially, with previously under-represented emotions achieving non-trivial accuracy levels. The most significant gains are observed for negative emotions such as disgust and fear, whose accuracy improves across all age groups, indicating enhanced discriminative capability under balanced training. In addition, performance variance across age subgroups is visibly reduced, suggesting more stable and robust feature learning. The best overall performance is achieved in the happy category, where high accuracy is maintained consistently across all age groups. These results further confirm that mitigating age distribution bias effectively enhances both fairness and reliability in facial expression recognition models.

5. Discussion

This study examined the age-related bias in facial expression recognition and proposed a subgroup-aware data augmentation strategy to improve model performance, particularly for elderly populations. The findings highlight several key aspects of how age affects emotion recognition systems and demonstrate that integrating demographic priors into data preprocessing can effectively mitigate bias across age groups.

5.1. Influence of Age on FER Performance

The experimental findings presented in this study empirically validate the data gap and algorithmic gap identified in prior research on facial expression recognition (FER). Although modern architectures such as ResNet-50 are often regarded as relatively fair and balanced across demographics, our baseline results reveal a clear performance degradation among older adults. This disparity echoes previous reports that emotion interpretation varies systematically with age [21][22] and highlights how conventional FER systems, optimized on youth-dominated datasets, fail to generalize effectively to aging populations.

 doi:10.20944/preprints202605.2112.v1

The observed data gap stems from the chronic underrepresentation of elderly individuals in most public FER datasets. As demonstrated in our dataset analysis, older faces constitute less than 10% of available samples, and age annotations are often missing or inaccurate. Such imbalance constrains feature diversity and forces models to rely on biased distributions dominated by younger faces. Consequently, deep learning models exhibit reduced robustness when encountering the morphological and textural variations characteristic of aging—such as wrinkles, drooping muscles, and reduced expression in-tensity.

In parallel, an algorithmic gap persists due to inadequate sensitivity to age-dependent cues. Physiological and psychological studies suggest that older adults depend more heavily on specific facial regions—particularly the eyes and mouth corners—for emotion decoding [24]. However, existing FER models, which typically learn spatially global features, seldom emphasize these localized regions. This mismatch leads to diminished feature salience and misclassification of subtle emotional expressions among older individuals.

Together, these findings confirm that both data- and model-level deficiencies jointly undermine cross-age generalization in FER. Bridging these gaps requires an integrated approach that not only increases data diversity but also adapts the learning process to the perceptual and morphological characteristics of different age groups—a motivation that directly informs the subgroup-aware augmentation strategy proposed in this study.

5.2. Effectiveness of Subgroup-Aware Augmentation

The proposed subgroup-aware augmentation method effectively reduced these discrepancies by introducing adaptive augmentation intensity based on the age of each sample. For individuals aged 65 years and above, augmentations such as random rotation, elastic deformation, and targeted contrast enhancement improved both feature diversity and the salience of discriminative facial regions. These operations align with psychological evidence that older adults rely more heavily on the eye and mouth regions for emotion interpretation.

Empirical evaluations demonstrated that models trained under this strategy achieved notably higher recognition accuracy for elderly faces while maintaining stable performance for younger groups. This outcome indicates that age-aware augmentation not only compensates for data imbalance but also encourages the network to learn age-specific cues that are often neglected in standard training pipelines.

Consistent with findings from other bias-mitigation studies, the incorporation of data augmentation techniques enables the model to better internalize target features and improves performance on underrepresented demographic groups, thereby enhancing the fairness and inclusivity of facial expression recognition systems [49–52].

5.3. Implications for Fair Affective Computing

The results have broader implications for fairness in affective computing. Among all potential sources of bias, datasets remain the most significant contributor. The present study reaffirms that most existing FER datasets are heavily skewed toward younger populations and positive emotions. Prior research has also revealed demographic im-balances related to race, gender, and other factors, reflecting systematic biases inherent in data collection processes [53–57]. These issues largely stem from the origins of the datasets—images collected through web searches or social media platforms naturally reflect user behavior, where younger individuals constitute the majority and positive expressions are more frequently shared [58,59]. Consequently, these datasets unintentionally encode social and demographic asymmetries into the training data.

Creating fairly distributed and demographically balanced datasets is therefore a critical step toward achieving equitable emotion recognition. Until such resources become widely available, researchers must remain aware of these biases and employ corresponding mitigation strategies—such as targeted augmentation, re-weighting, or domain adaptation—to alleviate the problem as much as possible [49,61].

 doi:10.20944/preprints202605.2112.v1

At the model level, current FER systems often claim generalizability across demographics, yet they overlook the structural imbalance embedded in their training data. By demonstrating that even modest modifications to the augmentation design can substantially improve inclusivity, this study provides a practical framework for developing fairer affective technologies. Such approaches are particularly vital for eldercare and home-companion applications, where accurate emotional understanding plays a key role in fostering trust, comfort, and user acceptance among older adults [42].

5.4. Limitations and Future Work

Despite its contributions, this study has several limitations that warrant further investigation. First, the re-trained age recognition model, although substantially improved, still exhibited a mean absolute error (MAE) of approximately seven years, which may introduce uncertainty in age labeling near category boundaries. Future research could address this issue by incorporating multimodal validation methods—for example, combining predicted age with survey-based metadata or by developing new facial expression datasets that include richer demographic and contextual annotations to reduce labeling errors.

Second, the current augmentation strategy was limited to visual-level transformations. Incorporating semantic-level augmentation, such as synthetic expression generation using diffusion models or GAN-based architectures, could further enhance robustness and compensate for underrepresented emotional or demographic groups.

Third, while this study focused primarily on age-related disparities, other demo-graphic biases—particularly racial and ethnic biases—remain an open challenge in FER systems. Numerous studies have demonstrated that models trained on imbalanced datasets often exhibit lower accuracy for darker-skinned or minority ethnic groups, reflecting underlying representation gaps and potential algorithmic discrimination [54,61]. Future work could integrate race-balanced sampling, fairness-aware loss functions, or domain adaptation strategies to jointly mitigate both age and racial biases and promote more inclusive emotion recognition.

Lastly, this study primarily focused on CNN-based architectures. Extending the subgroup-aware training framework to transformer-based or hybrid architectures (e.g., ViT, ConvNeXt) would provide deeper insights into how model architecture interacts with demographic fairness and representation learning.

6. Conclusions

This study addressed a critical gap in facial expression recognition by focusing on age-related disparities and proposing a subgroup-aware data augmentation approach. Through retraining of the age estimation model and adaptive augmentation tailored to elderly facial characteristics, the framework achieved improved robustness and fairness across age groups. The results demonstrate that integrating demographic priors—specifically age-sensitive augmentation—effectively enhances recognition performance for older adults without compromising accuracy for younger users.

Overall, this work provides a practical and scalable pathway toward age-inclusive affective computing, offering methodological insights for developing emotionally intelligent systems that can serve aging populations more equitably. Future research will ex-tend these findings to multimodal emotion recognition and real-world deployment in home-based assistive environments.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to the protection of the ongoing study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

 doi:10.20944/preprints202605.2112.v1

FER	Facial Expression Recognition
MAE	Mean Absolute Error

References

1. United Nations. Department Of Economic and Social Affairs, Population Division. World Population Prospects 2022. Available: https://population.un.org/wpp/. 2024/03/30.
2. Khatir M A, Modanloo M, Dadgari A, et al. Empty nest syndrome: a concept analysis[J]. Journal of Education and Health Promotion, 2024, 13(1): 269.
3. Xi J Y, Liang B H, Zhang W J, et al. Effects of population aging on quality of life and disease burden: a population-based study[J]. Global Health Research and Policy, 2025, 10(1): 2.
4. Fong J H. Disability incidence and functional decline among older adults with major chronic diseases[J]. BMC geriatrics, 2019, 19(1): 323.
5. World Health Organization. Ageing and health. Available: https://www.who.int/zh/news-room/fact-sheets/detail/ageing-and-health. 2024/03/30
6. Yuan X, Guo Q, Bennett Gayle D D. Personal intelligent agents: empowering or marginalizing older adults?[J]. Aslib Journal of Information Management, 2025: 1-25.
7. Marziali R A, Franceschetti C, Dinculescu A, et al. Reducing loneliness and social isolation of older adults through voice assistants: literature review and bibliometric analysis[J]. Journal of medical Internet research, 2024, 26: e50534.
8. Balakrishnan J, Dwivedi Y K, Hughes L, et al. Enablers and inhibitors of AI-powered voice assistants: a dual-factor approach by integrating the status quo bias and technology acceptance model[J]. Information Systems Frontiers, 2024, 26(3): 921-942.
9. Chou Y H, Lin C, Lee S H, et al. User-friendly Chatbot to mitigate the psychological stress of older adults during the COVID-19 pandemic: Development and usability study[J]. JMIR Formative Research, 2024, 8: e49462.
10. Moussawi S, Koufaris M, Benbunan-Fich R. The role of user perceptions of intelligence, anthropomorphism, and self-extension on continuance of use of personal intelligent agents[J]. European Journal of Information Systems, 2023, 32(3): 601-622.
11. Dzedzickis A, Kaklauskas A, Bucinskas V. Human emotion recognition: Review of sensors and methods[J]. Sensors, 2020, 20(3): 592.
12. Abbaschian B J, Sierra-Sosa D, Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models[J]. Sensors, 2021, 21(4): 1249.
13. Lugović S, Dunđer I, Horvat M. Techniques and applications of emotion recognition in speech[C]//2016 39th international convention on information and communication technology, electronics and microelectronics (mipro). IEEE, 2016: 1278-1283.
14. Klonsky E D, Victor S E, Hibbert A S, et al. The multidimensional emotion questionnaire (MEQ): Rationale and initial psychometric properties[J]. Journal of Psychopathology and Behavioral Assessment, 2019, 41(3): 409-424.
15. Samal P, Hashmi M F. Role of machine learning and deep learning techniques in EEG-based BCI emotion recognition system: a review[J]. Artificial Intelligence Review, 2024, 57(3): 50.
16. Sivapriyan R, Kumar N P, Suresh H L. Analysis of facial recognition techniques[J]. Materials Today: Proceedings, 2022, 57: 2350-2354.
17. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
18. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
19. Park H, Shin Y, Song K, et al. Facial emotion recognition analysis based on age-biased data[J]. Applied Sciences, 2022, 12(16): 7992.
20. Chen Y, Joo J. Understanding and mitigating annotation bias in facial expression recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14980-14991.

 doi:10.20944/preprints202605.2112.v1

21. Ruffman T, Henry J D, Livingstone V, et al. A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging[J]. Neuroscience & Biobehavioral Reviews, 2008, 32(4): 863-881.
22. Low A C Y, Oh V Y S, Tong E M W, et al. Older adults have difficulty decoding emotions from the eyes, whereas easterners have difficulty decoding emotion from the mouth[J]. Scientific reports, 2022, 12(1): 7408.
23. Faustmann L L, Eckhardt L, Hamann P S, et al. The effects of separate facial areas on emotion recognition in different adult age groups: A laboratory and a naturalistic study[J]. Frontiers in psychology, 2022, 13: 859464.
24. Fabbrizzi S, Papadopoulos S, Ntoutsi E, et al. A survey on bias in visual datasets[J]. Computer Vision and Image Understanding, 2022, 223: 103552.
25. Kopalidis T, Solachidis V, Vretos N, et al. Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets[J]. Information, 2024, 15(3): 135.
26. González-Gualda L M, Vicente-Querol M A, García A S, et al. An exploratory study of the effect of age and gender on face scanning during affect recognition in immersive virtual reality[J]. Scientific Reports, 2024, 14(1): 5553.
27. Wasi A T, Rafi T H, Islam R, et al. GReFEL: Geometry-Aware Reliable Facial Expression Learning under Bias and Imbalanced Data Distribution[C]//Proceedings of the Asian Conference on Computer Vision. 2024: 4368-4384.
28. Park J, Cho S B. Age-Unbiased Facial Emotion Recognition with Regularizing Self-Attention Value Vector[C]//International Conference on Intelligent Data Engineering and Automated Learning. Cham: Springer Nature Switzerland, 2024: 472-480.
29. Ma J, Liu X, Li Y. A Comparative Study Recognizing the Expression of Information Between Elderly Individuals and Young Individuals[J]. Psychology Research and Behavior Management, 2024: 3111-3120.
30. Chhua K, Wen Z, Hathalia V, et al. From Bias to Balance: Detecting Facial Expression Recognition Biases in Large Multimodal Foundation Models[J]. arXiv preprint arXiv:2408.14842, 2024.
31. Xu T, White J, Kalkan S, et al. Investigating bias and fairness in facial expression recognition[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 506-523.
32. Lv J J, Shao X H, Huang J S, et al. Data augmentation for face recognition[J]. Neurocomputing, 2017, 230: 184-196.
33. Huang Y, Peng J, Cai Z, et al. Facial expression recognition with age-group expression feature learning[C]//2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024: 1-8.
34. Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.
35. Mollahosseini A, Hasani B, Mahoor M H. Affectnet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.
36. Mahmoudi M A, Chetouani A, Boufera F, et al. Kernelized dense layers for facial expression recognition[C]//2020 IEEE international conference on image processing (ICIP). IEEE, 2020: 2226-2230.
37. World Health Organization. (2022). Ageing and health: Life-course perspective on ageing. Geneva: WHO.
38. Erikson E H, Erikson J M. The life cycle completed (extended version)[M]. WW Norton & Company, 1998.
39. Carstensen L L, Fung H H, Charles S T. Socioemotional selectivity theory and the regulation of emotion in the second half of life[J]. Motivation and emotion, 2003, 27(2): 103-123.
40. Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1701-1708.
41. Rothe R, Timofte R, Van Gool L. Deep expectation of real and apparent age from a single image without facial landmarks[J]. International Journal of Computer Vision, 2018, 126(2): 144-157.
42. Abdolrashidi A, Minaei M, Azimi E, et al. Age and gender prediction from face images using attentional convolutional network[J]. arXiv preprint arXiv:2010.03791, 2020.
43. Levi G, Hassner T. Age and gender classification using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 34-42.

 doi:10.20944/preprints202605.2112.v1

44. Zhang F, He Q, Kuang K, et al. Distributionally generative augmentation for fair facial attribute classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 22797-22808.
45. Dehdashtian S, He R, Li Y, et al. Fairness and Bias Mitigation in Computer Vision: A Survey[J]. arXiv preprint arXiv:2408.02464, 2024.
46. Wang J, Chung Y, Ding Z, et al. From Majority to Minority: A Diffusion-based Augmentation for Underrepresented Groups in Skin Lesion Analysis[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024: 14-23.
47. Kawai H, Ito K, Chen H T, et al. FSErasing: Improving Face Recognition with Data Augmentation Using Face Parsing[J]. IET Biometrics, 2024, 2024(1): 6663315.
48. Hosseini M M, Fard A P, Mahoor M H. Faces of fairness: Examining bias in facial expression recognition datasets and models[J]. arXiv preprint arXiv:2502.11049, 2025.
49. Khalil A, Ahmed S G, Khattak A M, et al. Investigating bias in facial analysis systems: A systematic review[J]. IEEE Access, 2020, 8: 130751-130761.
50. Udefi A M, Aina S, Lawal A R, et al. An analysis of bias in facial image processing: A review of datasets[J]. International Journal of Advanced Computer Science and Applications, 2023, 14(5).
51. Fabbrizzi S, Papadopoulos S, Ntoutsi E, et al. A survey on bias in visual datasets[J]. Computer Vision and Image Understanding, 2022, 223: 103552.
52. Alvi M, Zisserman A, Nellåker C. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings[C]//Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.