

Fairness in Facial Analysis

A Demographic Performance Evaluation of FaceReader

Amogh Gudi – January 2026

NOLDUS

noldus.com/facereader

Abstract

FaceReader is widely used in academia and industry to automatically classify facial expressions and Action Units (AUs). As such systems increasingly inform research, product decisions, and user experiences, it becomes crucial to understand how reliably they perform across different demographic groups. This whitepaper evaluates FaceReader's facial expression and AU performance across ethnicity, age, and gender using two independent datasets.

FaceReader's expression and AU performance were found to exhibit only modest differences across ethnic groups. These patterns closely resemble human performance/biases. Expression and AU accuracies were similar for male and female subjects. Among age groups, performance is highest for the younger subjects, with a minor drop for older adults. Additionally, results indicate that under good conditions, FaceReader performs on par with human coders overall and shares similar bias patterns across demographic groups.

Our key conclusion is that FaceReader does not exhibit strong systematic demographic bias in this evaluation, although some differences persist particularly for older age groups and specific expressions in some ethnic groups, mirroring patterns also observed among human annotators.

Contents

Abstract 02

Introduction 04

Data and Evaluation Setup 06

- DRAF Dataset 06
- RADIATE Dataset 07

Experiments and Results 08

- Ethnicity-based performance differences 08
- Gender-based performance differences 10
- Age-based performance differences 11
- FaceReader and Human Performance 13

Discussion 14

Recommendations for responsible use 15

Conclusion 16

References 16

Introduction

Automated facial expression analysis has become a foundational technology in affective computing, user experience research, human–computer interaction, and behavioral science. Systems such as FaceReader can estimate emotion categories, continuous affective dimensions, and Facial Action Units (AUs) from videos at scale, which would otherwise require costly and time-consuming human coding.

However, increased adoption has also amplified concerns about fairness and bias in computer vision and affective computing. Studies have shown that face analysis systems can exhibit different error rates across demographic groups, particularly across ethnicity and skin tone, but also age and gender [1,2]. This raises important questions about who is included in the system and who might be disadvantaged when it makes mistakes.

Against this backdrop, international regulations and standards (e.g., the EU AI Act, ISO/IEC AI standards) increasingly emphasize transparent, evidence-based evaluation of AI systems, including disaggregated performance reporting across demographic subgroups. For a system like FaceReader, which is used in diverse applications (from usability studies to psychological research), users need clear answers to questions such as:

- Does the system perform similarly across different ethnic groups?
- Are older adults recognized as accurately as younger adults?
- Is there a systematic performance difference between male and female subjects?
- How do these patterns compare to human observers?

This whitepaper addresses these questions by:

1. Quantitatively evaluating FaceReader's facial expression and AU performance across ethnicity, age, and gender on two independent datasets.
2. Comparing system performance with human annotators on a racially diverse basic emotion dataset.

Our goal is not to claim that FaceReader's estimates are "unbiased," but to provide transparent, data-driven evidence about its behavior across demographic groups[1], enabling more informed deployment and interpretation.

This whitepaper builds upon and extends the work done in Rogers, 2022 [3] and Noldus, 2023 [4] which focusses on studying FaceReader's modelling quality across ethnicities and genders.

[1] Demographic groups are as defined by the datasets (self-report or dataset metadata). We emphasize that these categories are socially constructed and coarsely defined; they are used here solely for fairness evaluation, consistent with how the datasets themselves are described.

Data and Evaluation Setup

We employ two demographically diverse datasets for our evaluation of FaceReader's facial expression classification and Action Unit detection performance.

DRAF: Diverse Real-world Action unit and Facial expression Dataset

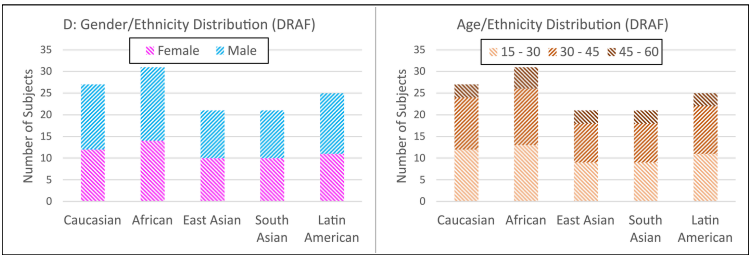


Figure 1: Demographic distribution of subjects in the DRAF dataset. This dataset covers 5 ethnicities with gender and age labels.

The DRAF (Diverse Real-world Action unit and Facial expression) dataset was designed to capture real-world variation in facial behavior for AU detection and facial expression classification:

- This dataset contains ≈3,100 facial images from ≈125 subjects.
- Subjects were recruited via paid crowdsourcing. Participants recorded themselves while producing both posed and natural facial expressions. Real-world conditions: heterogeneous lighting, head pose, camera types, and image quality.
- Demographics (see Figure 1):
  - Ethnicity/region: Caucasian, African, East Asian, South Asian, Latin American.
  - Age: 18–60 years, categorized into three groups: 15–30, 30–45, and 45–60.
  - Gender: Both male and female subjects, evenly distributed.
- Each image is annotated for 10 commonly occurring AUs (1, 2, 4, 5, 6, 9, 12, 17, 25, 26) by certified FACS coders.

This dataset thus provides a challenging, diverse, real-world benchmark for AU detection and expression classification.

RADIATE: Racially Diverse Basic Emotion Stimuli Dataset

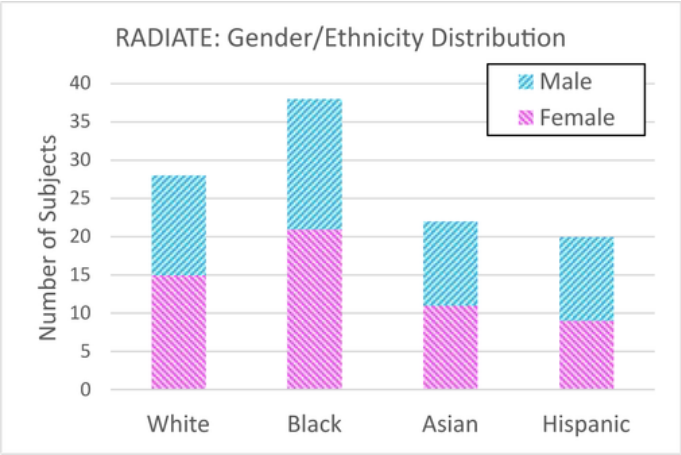


Figure 2: Demographic distribution of subjects in the RADIATE dataset. This dataset spans of 4 ethnicities with an even distribution of gender.

The RADIATE (Racially Diverse Affective Expression) dataset by Conley, et al., 2018 [5] is a widely used face stimulus set explicitly designed to increase racial and ethnic diversity in emotion research.

- ≈800 images from ≈110 subjects were selected from this dataset.
- Subjects include (see Figure 2) adult models (56 female, 53 male) aged 18–30, from four ethnicity groups: White, Black, Hispanic, and Asian.
- The selection includes each model posing each of the 7 basic facial expressions: Neutral, Happy, Sad, Angry, Surprised, Scared, and Disgust.
- Images were acquired under controlled lab conditions with consistent lighting and camera setup.
- In addition, this dataset includes human judgments from crowd-sourced participants, who classified the emotion in each image using a forced-choice task, providing psychometric measures of validity and reliability. We use these published human accuracies as our human baseline for each emotion and demographic group.

RADIATE thus offers a good test set to compare FaceReader's performance across groups with the context of human performance.

Experiments and Results

We present the experimental setup and report FaceReader's performance across demographic groups, followed by a comparison with human observers where available.

Ethnicity-based performance differences

Facial expression classification across Ethnicities

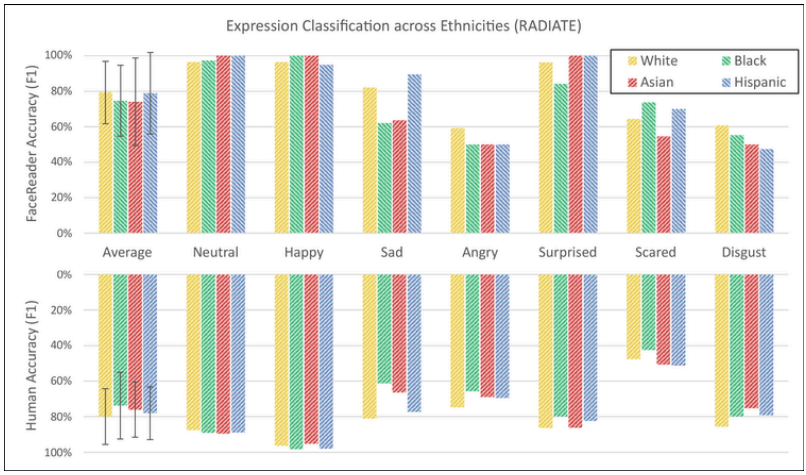


Figure 3: Expression classification accuracy across ethnicities in RADIATE. Top chart shows FaceReader accuracy (F1), while the bottom shows crowd-sourced human labeller accuracy. On average, FaceReader shows slightly higher accuracy for Caucasian and Hispanic ethnicities compared to Black and Asian; however, humans exhibit a very similar pattern.

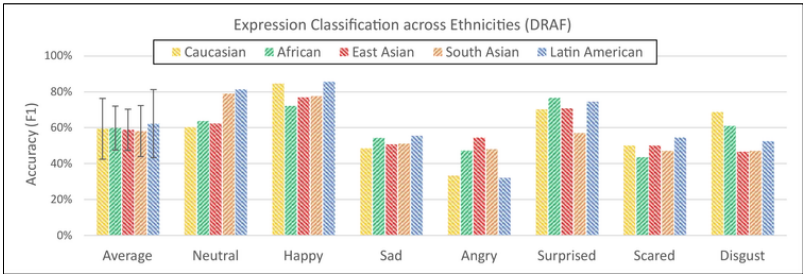


Figure 4: FaceReader's expression classification accuracy across ethnicities in DRAF. On average, FaceReader shows uniform performance across all ethnicities, with Latin American subjects showing a slightly higher accuracy.

On RADIATE, FaceReader's average expression classification accuracy shows moderate variation across ethnic groups (see Figure 3): FaceReader obtains an accuracy of ≈80% on Caucasian and Hispanic subjects, while Black and Asian subjects give ≈75%. Human observers exhibit very similar patterns. *This suggests that part of the performance gap reflects broader perceptual biases and cross-group recognition difficulties that also affect humans, rather than being introduced solely by the algorithm. For example, a potential reason for the bias among humans could be that majority of them were likely US-based, and therefore more familiar with Caucasian facial expressions.*

On DRAF, where images are more varied and challenging, FaceReader's expression-related performance is more uniform across ethnicity/region (see Figure 4): accuracies across Caucasian, African, East Asian, South Asian, and Latin American subjects are broadly similar; the largest deviation observed is a ≈4 percentage-point higher accuracy for Latin American faces compared to other groups. *Given the relatively small absolute differences and the inherent noise in real-world data, we interpret this as no substantial ethnicity-driven degradation in expression-related performance.*

AU detection across Ethnicities

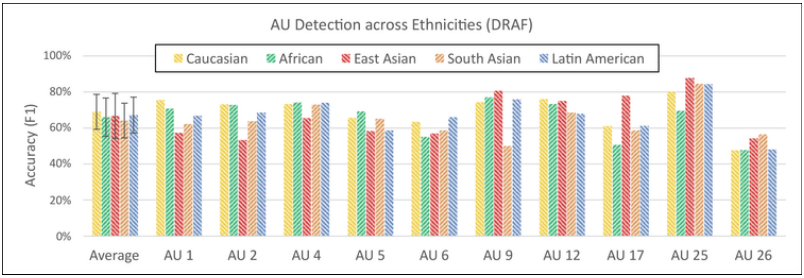


Figure 5: FaceReader's AU detection accuracy across ethnicities in DRAF. Caucasian and South Asian groups achieve the highest and the lowest accuracy respectively, but the differences between ethnic groups are modest on average.

For AU detection on DRAF (see Figure 5): average AU detection accuracy across ethnicity groups lies in a narrow band (≈64–69%), where the lowest accuracy is obtained for the South Asian group (≈64%) and the highest accuracy for the Caucasian group (≈69%). *While these differences are measurable, they are modest in magnitude, especially considering the challenging in-the-wild recording conditions. The absence of large outliers suggests that the AU models are reasonably robust across the sampled ethnic groups, though there is still room for improvement, particularly for South Asian faces.*

Gender-based performance differences

Facial expression classification across Gender

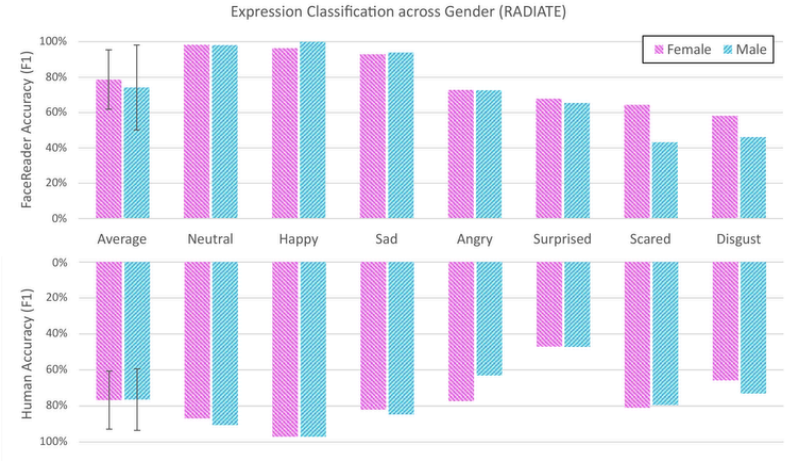


Figure 6: Expression classification accuracy across gender in RADIATE. Top chart shows FaceReader accuracy (F1), while the bottom shows crowd-sourced human labeller accuracy. FaceReader's accuracy is in the same approximate range for both genders with a slightly higher score for female over male subjects; humans also show a similar pattern.

On RADIATE, FaceReader's expression accuracy (see Figure 6) is in the same approximate range for both male (≈74%) and female subjects (≈78%). This difference was largely driven by differences in Scared and Disgust expressions. Human observers show a similar pattern with slightly higher accuracy for females in comparison to males. *In other words, both FaceReader and human raters do not show a strong or systematic gender bias on this dataset; any differences are small relative to overall accuracy.*

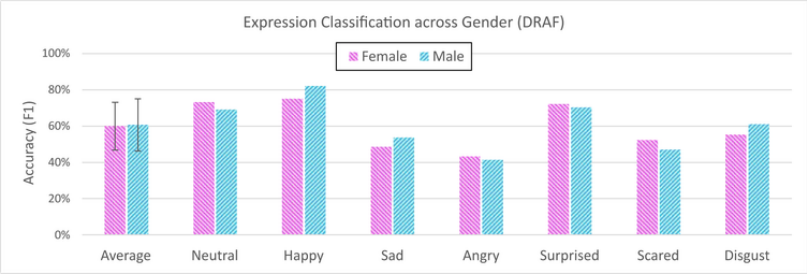


Figure 7: FaceReader's expression classification accuracy across gender in DRAF. There is only a marginal difference in accuracies for both genders.

On DRAF (see Figure 7), expression-related performance for male and female subjects are similar, with average differences < 1% point across emotions. *This indicates no meaningful gender-driven discrepancy in expression recognition under real-world conditions for this dataset.*

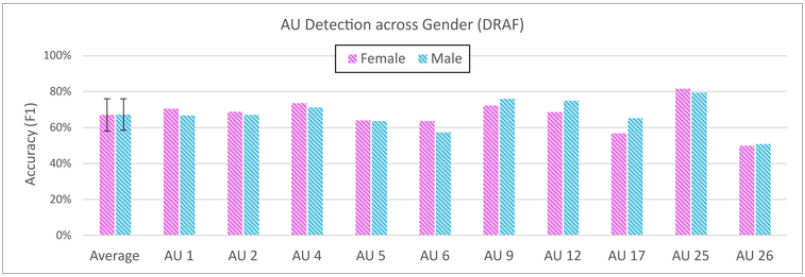


Figure 8: FaceReader's AU detection accuracy across gender in DRAF. There is no significant difference in accuracies between the two groups.

For AU detection on DRAF (see Figure 8), average AU accuracy is nearly identical for male and female subjects (within ≈0.2%). *Such a small difference is within the range one would expect due to dataset sampling and noise, again pointing to no strong gender bias in AU detection performance.*

Age-based performance differences

Age effects are evaluated using the DRAF dataset, which contains a broader age range.

Facial expression classification across Age Groups

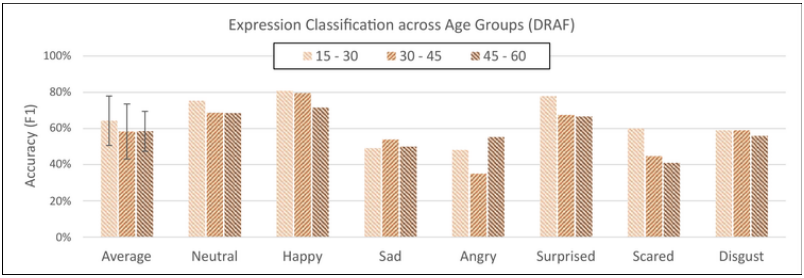


Figure 9: FaceReader's expression classification accuracy across age groups in DRAF. Best performance is obtained for the youngest age group, and performance on older age groups is slightly and progressively lower.

As shown in Figure 9, FaceReader achieves the highest expression classification accuracy in the youngest age group, while performance of older subjects is ≈6 percentage points lower on average compared to the youngest group. *This pattern is consistent with the broader literature: many face analysis models are trained on data skewed toward younger adults, which can result in slightly degraded performance on older faces if not explicitly corrected.*

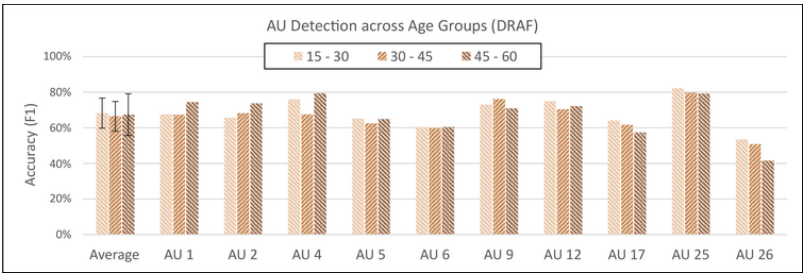


Figure 10: FaceReader's AU detection accuracy across age groups in DRAF. Average accuracies are very similar across all age groups.

For AU detection across the same three age groups (see Figure 10): average accuracies are very similar across groups; the youngest group (15–30 years) outperforms the oldest group (45–59 years) by only about 2 percentage points. *This indicates that AU detection is less sensitive to age than full expression classification in the DRAF data, though the youngest group still enjoys a small advantage.*

FaceReader and Human Performance

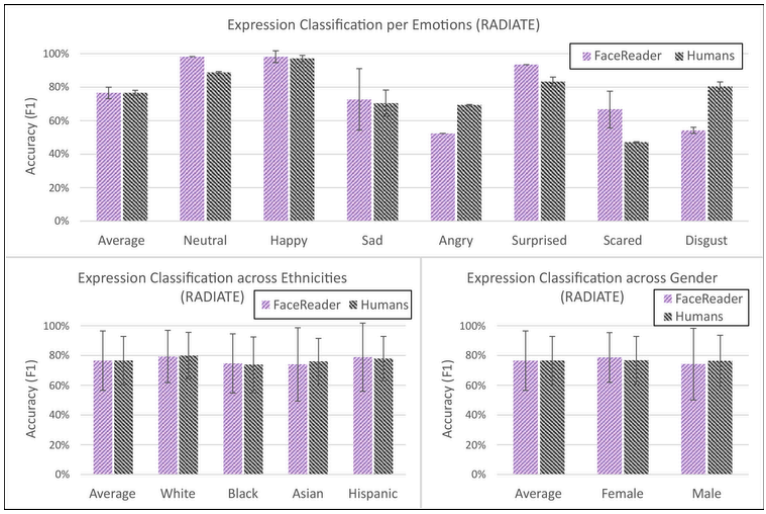


Figure 11: Comparative expression classification of FaceReader vs crowd-sourced human coders in RADIATE. On average, FaceReader and humans obtain the same accuracy on this dataset, with FaceReader comparatively higher accuracy for some emotions, and lower on others. The pattern of performance differences across demographic groups is also very similar between FaceReader and humans.

A particularly important question is whether FaceReader's demographic and overall performance patterns differ substantially from those of humans. To answer this, we can explicitly compare FaceReader's accuracy with that of (crowd-sourced) human raters (see Figure 11).

On RADIATE, overall performance of FaceReader as well as humans is ≈76.5% on average. For some emotions (e.g., Scared, Surprised, Neutral), FaceReader significantly outperforms average human accuracy. For others (e.g., Disgust, Angry), FaceReader's performance is somewhat lower than human performance. *Overall, this indicates that FaceReader can perform on par with human observers when recording conditions are good.*

In addition, FaceReader and human raters show similar relative performance trends across ethnic groups: higher accuracies on White/Hispanic faces and slightly lower on Black/Asian faces. The magnitude of these differences is comparable between humans and FaceReader. *This alignment suggests that FaceReader broadly inherits human-like difficulty patterns, rather than introducing entirely new or amplified demographic biases. That said, this does not absolve the system of bias, it simply indicates that bias patterns are shared rather than uniquely algorithmic.*

Discussion

Across two complementary datasets (DRAF and RADIATE) and three demographic axes (ethnicity, age, gender), we find no strong, systematic demographic bias: facial expression classification performance differences across ethnicity, age, and gender are generally small to moderate (2–6 percentage points) rather than large or catastrophic. AU detection performance was found to be particularly stable across all groups, with small differences typically in the low single digits. This greater stability is likely because AU detection targets relatively objective, localized patterns of muscle movement in specific parts of the face. This is likely more agnostic to demographic variations as compared to classifying basic emotions, which relies on a more holistic interpretation of the entire face and mapping onto relatively subjective categories.

FaceReader shows somewhat higher accuracy for White and Hispanic faces and slightly lower for Black and Asian faces in RADIATE, a pattern nearly identical to that of human raters. On DRAF, ethnicity/region differences are even smaller, with only minor advantages for Latin American faces.

On gender, expression and AU performance for male and female faces is essentially the same, with differences much smaller than overall accuracy.

In terms of age, younger adults (15–30) are recognized most accurately. Older age groups (30–60) show a moderate drop (≈2–6 points) in accuracy, but not a collapse in performance. This reflects both dataset imbalances and well-known difficulties in modeling facial changes with age. However, these accuracies indicate the performance across age groups remains good and usable.

FaceReader's overall accuracy matches human performance on RADIATE. Demographic performance patterns (which groups are easier/harder) closely mirror those seen in human observers.

Taken together, these results indicate that FaceReader is reasonably balanced across demographic groups in the evaluated settings, with no evidence of extreme bias in the sense of very high performance for one group and very poor performance for another.

Recommendations for responsible use

Given these findings, the following recommendations may be considered by users and researchers using FaceReader:

1. Disaggregated reporting: When publishing results, report disaggregated results for each demographic subgroup wherever possible. This can help detect context-specific issues.
2. Human oversight: For sensitive or critical applications, FaceReader outputs should be treated as decision support, not final decisions. Human experts can review borderline cases, especially for demographics or conditions known to be challenging.
3. Accounting for group differences: If a particular demographic or user group is especially important for your application, consider steps to account for group differences. For example, collecting a larger/more balanced dataset for that group, and/or applying post-hoc calibration or group-wise thresholding to reduce performance gaps.
4. Understanding dataset limitations: When designing studies using FaceReader, it is useful to clearly document factors like the demographic composition of your sample, the context (lab vs. real-world), and any known variables relevant to your population.

Conclusion

This whitepaper presents a demographically disaggregated evaluation of FaceReader's facial expression and AU recognition performance. The results show that FaceReader's performance is broadly consistent across ethnicity, age, and gender, with only modest differences that largely mirror patterns also seen in human observers.

These results can be interpreted as evidence that FaceReader, in its current form, is suitable for a broad range of research and applied contexts. FaceReader is not "bias-free", but its behavior across demographic groups is quantified and well-enough understood to support responsible and informed use, especially when combined with good study design, human oversight, and continued monitoring.

References

[1] Cheong, J., Kalkan, S., & Gunes, H. (2021). The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing: Overview and Techniques. *IEEE Signal Processing Magazine*, **38(6)**, 39–49.

[2] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR **81**, 77–91.

[3] Rogers, J. (2022). Ethnicity & FaceReader 9 – A FairFace Case Study. *Measuring Behavior: 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods*, (pp. 117 - 122).

[4] Noldus, I. (2023). FaceReader 9 - FairFace validation. Whitepaper.

[5] Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. (2018). The racially diverse affective expression (RADIATE) face stimulus set. Psychiatry Research, 1059-1067.

We are represented by a worldwide network of distributors and regional offices. Visit our website for contact information.

noldus.com

International headquarters
Noldus Information Technology bv
Wageningen, The Netherlands
E-mail: contact@noldus.com

North American headquarters
Noldus Information Technology Inc.
Leesburg, VA, USA
E-mail: info@noldus.com

Due to our policy of continuous product improvement, information in this document is subject to change without notice. FaceReader is a trademark of Vicarious Perception Technologies BV.
© 2026 Noldus Information Technology bv. All rights reserved.