IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/44AI/2024/4430624/
Vol. 2024 (22), Article ID 4430624

# AI Fairness for People with Disabilities: Investigating Facial Recognition System Performance*

Zineb BELKHO[1], Amina ZAHAR[1], Naoual CHAOUNI BENABDELLAH[2]
And Zakariae ABBAD[2]

[1]The National School of Applied Sciences/Cadi AYYAD University, Safi, Morocco

[2]Software and Project Management Research Team
ENSIAS/Mohammed V University, Rabat, Morocco

Correspondence should be addressed to Zineb BELKHO, zinebbelkho1@gmail.com

* Presented at the 44th IBIMA International Conference, 27-28 November 2024 Granada, Spain

## Abstract

Disabled individuals, who comprise approximately 16% of the global population, may be underrepresented in the data used to develop AI systems such as facial recognition technologies. Researchers suggest that such systems may fail to perform effectively for people with distinct facial features, such as those with Down Syndrome or Achondroplasia condition, and could potentially discriminate against them if they were not included during the model training and evaluation process. In this study, we explore the performance of state-of-the-art facial recognition models when applied to individuals with disabilities. Our analysis involves three datasets, the first includes 100 subjects with disabilities, each having 7 to 10 facial images sourced from Creative Commons YouTube videos and Websites, while the other two consist of LFW and CelebA. To ensure a fair comparison, we assessed the quality of each dataset using two distinct algorithms. Following this, we evaluated the performance of three CNN architectures: VggFace, FaceNet, and ArcFace. Our findings reveal that VggFace and FaceNet perform similarly for individuals with disabilities to those in the LFW and CelebA datasets. This indicates that these models generalize well across diverse populations, suggesting no evidence of bias or discrimination against disabled individuals.

**Keywords:** neural network, disability, facial recognition, biometric

## Introduction

Nowadays, systems incorporating biometrics are now widely used in personal, commercial, and government identity management applications. Traditionally, biometrics have been associated with the development of statistical and mathematical methods for analyzing biological data (Chowdhury, 2011). In the international standard ISO/IEC 2382-37 (Rathgeb, 2023), “biometrics” is defined as: “automated recognition of individuals based on their biological and behavioral characteristics.” Facial recognition is a widely used biometric technique with different approaches in continuous development. These include local, holistic, and hybrid methods, each with unique advantages for describing face images. Local approaches focus on specific facial features but can be sensitive to variations in facial expression, occlusions, and pose (Liao, 2012). On the other hand, holistic (or subspace) methods analyze the entire face without needing to extract specific regions or feature points, such as eyes, mouth, or nose. In these methods, the face is represented as a pixel matrix, often converted into feature vectors that are further processed in lower-dimensional spaces. Hybrid methods combine the strengths of both local and subspace approaches to leverage the advantages of each technique.

**Cite this Article as:** Zineb BELKHO, Amina ZAHAR, Naoual CHAOUNI BENABDELLAH And Zakariae ABBAD, Vol. 2024 (22) “AI Fairness for People with Disabilities: Investigating Facial Recognition System Performance " Communications of International Proceedings, Vol. 2024 (22), Article ID 4430624, https://doi.org/10.5171/2024.4430624

(Kortli, 2020) In recent years, there have been significant advancements in neural network architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Both CNNs and RNNs have proven to be highly effective in tasks like image detection and recognition. CNNs, in particular, have been widely adopted across various applications due to their success in capturing spatial hierarchies in images. (Kortli, 2020) However, there are growing concerns about bias in facial recognition algorithms, with some being labeled as "racist" or "unfair." (Rathgeb, 2022; Drozdowski, IEEE Transactions on Technology and Society). Efforts to improve fairness have focused on gender (Albiero, 2020), race, and skin tone (Krishnapriya, 2020), but there is a lack of research on the performance of these systems for minority groups, particularly individuals with disabilities who constitute less than 16% of the global population according to World Health Organization (2024). The effectiveness of facial recognition systems depends on the diversity of the training data used to develop the models. Unfortunately, individuals with distinct facial features, such as those with Down syndrome, and achondroplasia, are often underrepresented in training datasets, leading to inaccuracies in facial recognition for these individuals (Guo, 2020). Given this underrepresentation, it is essential to examine whether these systems perform equitably across all population segments. Our study aims to bridge this gap by evaluating the performance of three state-of-the-art facial recognition models using a dataset specifically curated for individuals with disabilities. This dataset was collected from publicly available sources, such as YouTube and Websites. We then compared the results from this dataset against two widely recognized benchmark datasets, LFW and CelebA, which primarily consist of images of individuals without disabilities.

The rest of this work is organized as follows: In Section II, we provide a review of related work. Section III describes the proposed method, including data collection, quality score calculation, and the specific CNN architectures used for face recognition. Section IV presents the experimental results and comparative analysis of the performance across the three datasets. Section V offers a discussion. Finally, Section VI concludes the study and summarizes our findings.

## Related Work

Over the past decade, there has been significant progress in face recognition technology research. Many studies have aimed to improve accuracy, speed, and reliability across different populations. Numerous studies have highlighted the capabilities of state-of-the-art Convolutional Neural Networks (CNNs) for facial recognition tasks.

For instance, Google researchers (Schroff, 2015) proposed a deep convolutional network designed to optimize the embedding (feature vector) directly, rather than relying on an intermediate bottleneck. Their training process involves triplets of roughly aligned matching and non-matching face patches, generated using a novel online triplet mining method. This innovative approach allows the system to achieve state-of-the-art face recognition performance, requiring only 128 bytes per face. On the widely used Labeled Faces in the Wild (LFW) dataset, their system sets a new record accuracy of 99.63%.

Moreover, (Deng, 2019) introduced the Additive Angular Margin Loss (ArcFace) to derive highly discriminative features for face recognition. This model provides a clear geometric interpretation, directly corresponding to the geodesic distance on a hypersphere. Results from the study demonstrated that ArcFace consistently outperforms other state-of-the-art methods, offering high performance with minimal computational overhead.

In another study by (Schroff, 2015), researchers introduced a standard dataset to evaluate the performance of face recognition algorithms in practical applications. They tested popular CNN models on this dataset, and the experimental results revealed that while accuracy remained high as the number of classes increased, the performance of all CNNs significantly declined when tested with samples from a new dataset. Among the models, the VGG-Face model outperformed others, demonstrating superior robustness under various image degradations.

However, researchers have raised concerns about potential bias towards minority groups in these systems. Unfair gender and racial bias have been identified in existing systems. (Angwin, 2016; Bolukbasi, 2016; Buolamwini, 2018) Considerations regarding fairness in AI for people with disabilities have thus far received little attention. (Guo, 2020) identified potential areas of concern regarding fairness in AI for people with disabilities. They hypothesize that these techniques may not work well for individuals with differences in facial features and expressions if not considered during the gathering of training data and evaluation of models. However, in a research study (Trewin, 2018), researchers state that fairness issues for the disabled may be more challenging to address than fairness issues for other groups, given the diverse, nuanced, and dynamic nature of disability itself. Impairments and health conditions that lead to disabilities are not only diverse but also vary in intensity and impact, and often change over time. Even if they are included in training and evaluation data, they may be overlooked as outliers by current AI techniques.

In (Kane, 2020), researchers investigated the experiences of 40 adults with physical disabilities with various sensing technologies, including motion sensors, biometric sensors, speech input, and touch and gesture systems. They conducted an online survey, collecting open-ended responses to understand the challenges these technologies present to individuals

with physical disabilities. The findings reveal that current sensing systems often fail to adequately accommodate the needs of this population, leading to significant difficulties in their use.

In a research study by (Rathgeb, 2024), the researchers evaluated the performance of face recognition systems specifically for individuals with Down syndrome, marking the only known study of facial recognition focused on a disabled group. The findings reveal that there is a significant decrease in face recognition performance for individuals with Down syndrome, primarily due to an increased likelihood of false matches.

## Methods

### *Data collection*

There is a critical shortage of datasets that include individuals with disabilities, presenting a significant challenge for the evaluation and development of equitable facial recognition systems. To bridge this gap, we embarked on the creation of a diverse dataset featuring 100 individuals with disabilities, encompassing conditions such as Down syndrome and treacher collins syndrome (Table 1 summarizes all the disabilities represented in our dataset). This dataset was carefully assembled from YouTube videos available under Creative Commons licenses, ensuring an equitable distribution of male and female participants, with 55% of the subjects being male and 45% female. Each individual is represented by at least seven distinct images, resulting in a comprehensive collection of 977 images.

**Table 1 Disabilities representend in our collected dataset.**

| Type | Diasability |
|---|---|
| **Genetic Disorders** | Down Syndrome |
| | Treacher Collins Syndrome |
| | Apert Syndrome |
| | Achondroplasia |
| | Spondyloepiphyseal Dysplasia Congenita |
| | Hypohidrotic Ectodermal Dysplasia |
| | Marfanoid-Progeroid-Lipodystrophy Syndrome |
| | Little Johnston Syndrome |
| | Progeria |
| **Neurological and Developmental Disabilities** | Mental Illness |
| | Intellectual Disability |
| | Cerebral Palsy |
| | Developmental Disabilities |
| | Amyotrophic Lateral Sclerosis (ALS) |
| **Physical Disabilities and Conditions** | Brittle Bone Disease (Osteogenesis Imperfecta) |
| | Dwarfism ( pecific conditions include Achondroplasia) |
| | Blindness |
| | Craniofacial Disability |

The images were captured at intervals of 3 to 8 seconds, allowing us to capture a range of facial expressions while maintaining low intraclass variations, given that all images were extracted from the same videos. Our newly collected dataset fulfills several key requirements for evaluating facial recognition systems. The subject's face in each image is predominantly frontal, with slight variations in pose and facial expressions, reflecting common real-world conditions. Additionally, unique identity labels have been assigned to each subject to facilitate the investigation of recognition performance. For this study, the focus is on young adults and adults, as they represent the primary users of facial recognition technologies, ensuring that our findings are relevant to the most common applications of these systems. To ensure a comprehensive evaluation, we selected two benchmark datasets, LFW and CelebA, for comparison. These datasets are commonly used in facial recognition studies and allow for a direct comparison of the models' performance on different datasets.

During the preprocessing stage, we utilized the Multi-Task Cascaded Convolutional Neural Network (MTCNN) (Zhang el al. 2016) to detect and crop faces from all three datasets as illustrated in Figure 1.

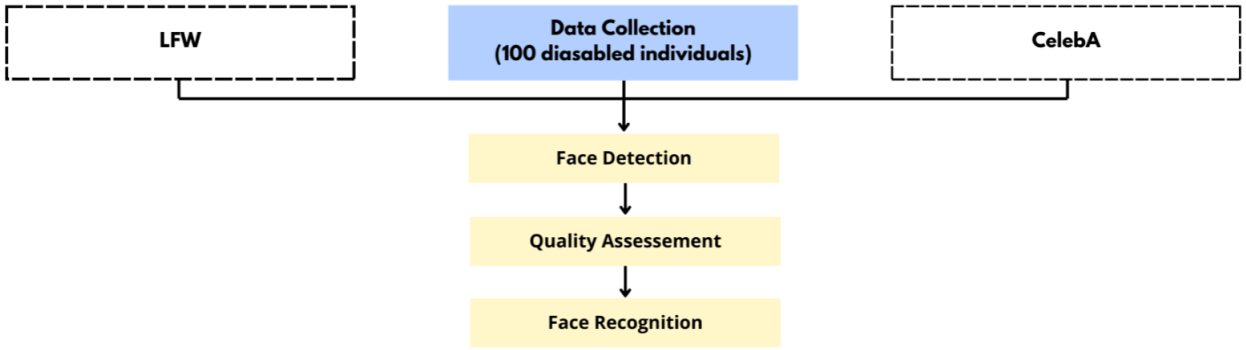


**Fig 1: Model development flowchart.**

## *Quality assessment*

In the context of biometrics, a quality metric is a function that takes a biometric sample as its input and returns an estimation of its quality level. This quality level is typically associated with the utility of the sample, meaning the expected recognition accuracy when using the sample in a biometric system (Hernandez et al. 2019). By assessing the data quality, we ensure that the samples are of a sufficient standard to be reliably used in face recognition models, thereby enabling a fair comparison across the datasets.

Two open-source deep learning-based algorithms **FaceQnet** (Hernandez-Ortega, 2019) and **MagFace** (Meng, 2021) are used to predict the quality of the three datasets, each with its unique approach. FaceQnet returns a quality score as an estimator of biometric performance, ranging between 0 and 1, where a higher score indicates better performance potential of the face image in recognition tasks. On the other hand, MagFace evaluates the quality of face images by computing the magnitude (or norm) of the face embeddings. A higher magnitude in MagFace corresponds to a higher-quality image, indicating that the facial features are more distinguishable and likely to lead to more accurate recognition results.

## *Models*

In our recognition experiments, we used three advanced models, VGGFace (Parkhi, 2015), ArcFace (Deng, 2019) , and FaceNet (Schroff, 2015), to evaluate their effectiveness on a dataset focusing on disabled individuals, in comparison to benchmark datasets. The feature extraction process involved RGB color space (3 channels), and the models generated specific dimensional embedding vectors based on their architecture.

To assess the performance of these models, we employed the Equal Error Rate (EER) metric, which is particularly important as it represents the point where the False Acceptance Rate and False Rejection Rate intersect, providing insight into the model's ability to correctly authenticate authorized individuals while minimizing the likelihood of incorrectly accepting unauthorized individuals. In addition, we also considered mated and non-mated similarity scores by calculating the Euclidean distance, to understand how well the system differentiates between matching and non-matching pairs of images.

## *Experimental results*

The quality assessment results are summarized in Table 2. Both algorithms provided comparable results across the Handicap, LFW, and CelebA datasets, indicating that the images in all datasets are of similar quality. These findings confirm that the datasets are suitable for facial recognition tasks, ensuring a fair comparison of the models' performance.

**Table 2: Sample Quality Statistics Distributions (average score and Std. DEV.) across datasets.**

| Algorithm | Database | $\mu$ | $\sigma$ |
|---|---|---|---|
| FaceQnet | Handicap | 0.5601 | 0.0595 |
| | LFW | 0.6143 | 0.0592 |
| | CelebA | 0.6 | 0.0578 |
| MagFace | Handicap | 26.0274 | 3.6511 |
| | LFW | 24.5854 | 2.9293 |
| | CelebA | 29.3393 | 3.1498 |

On the other hand, Table 3 illustrates the performance results of three state-of-the-art CNN architectures.

**Table 3: EER and Recognition Statistics for Mated and Non-Mated Distributions (average score and Std. DEV.) across datasets.**

| System | Database | Mated | | Non-Mated | | EER |
|---|---|---|---|---|---|---|
| | | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | |
| VggFace | Handicap | 0.8495 | 0.2303 | 1.3239 | 0.0581 | 5.3 |
| | LFW | 0.9814 | 0.1519 | 1.3354 | 0.0557 | 5.26 |
| | CelebA | 1.0265 | 0.1787 | 1.3576 | 0.0394 | 5 |
| FaceNet | Handicap | 0.6731 | 0.2457 | 1.3112 | 0.122 | 5.87 |
| | LFW | 0.7111 | 0.1705 | 1.3603 | 0.1043 | 5.33 |
| | CelebA | 0.7690 | 0.2014 | 1.3765 | 0.1088 | 4 |
| ArcFace | Handicap | 0.8729 | 0.2565 | 1.3371 | 0.1115 | 16.2 |
| | LFW | 0.9074 | 0.1429 | 1.2814 | 0.1043 | 14.96 |
| | CelebA | 0.9269 | 0.1979 | 1.3823 | 0.0687 | 8.5 |

The VggFace model achieved EERs of 5.3%, 5.26%, and 5% for the Handicap, LFW, and CelebA datasets, respectively. FaceNet also performed well with EERs of 5.87%, 5.33%, and 4% for the same datasets. However, ArcFace demonstrated higher EERs of 16.2%, 14.96%, and 8.5% for the Handicap, LFW, and CelebA datasets, respectively. The performance gap among these models can be attributed to the lower efficacy of ArcFace, as supported by multiple studies. A study by Firmansyah et al. (2023) assessing the performance of ArcFace, FaceNet, and FaceNet512 within the DeepFace framework revealed that FaceNet is more accurate than ArcFace. In a research study by Iype et al. (2023), further reinforces this conclusion by comparing ArcFace, FaceNet, and OpenFace, finding that FaceNet consistently outperforms ArcFace in accuracy. Additionally, CelebA achieved the best EER across all three models, likely due to its superior overall image quality. While the general quality of the Handicap, LFW, and CelebA datasets is comparable, CelebA stands out for providing more favorable conditions for face recognition, contributing to the lower EERs observed across the models.

As Table 3 demonstrates, the mean-mated Euclidean distance for the Handicap dataset is lower across all models, indicating effective recognition of matching individuals. Additionally, the non-mated Euclidean distances remain relatively similar to those for other datasets. These results collectively suggest that VGGFace and Facenet systems perform well across various datasets, with no substantial evidence of discrimination against individuals with disabilities.

## Discussion

Many individuals with disabilities have concerns about sharing their disability status due to potential discrimination and exclusion. For example, a recent study by (Ameri, 2018) showed that disclosing a disability in a job application cover letter led to a 26% decrease in positive responses from employers, even when the disability is unlikely to affect job performance. Another study by (Kruse, 2018) emphasizes that employed individuals with disabilities earn, on average, less than their non-disabled counterparts. This pay gap is not due to differences in productivity but rather reflects potential discrimination, supporting findings from prior literatures by (Baldwin, 2014) and (DeLeire, 2001). As a result, there are few resources available under Creative Commons licenses. This limitation impacted the amount of data gathered, which means that the data does not cover all disabilities, affecting the generalizability of results. Additionally, even if some disabilities are covered, they may be treated as outliers due to insufficient data. Future research should focus on expanding datasets and addressing the consequences of this limitation.

## Conclusion

Despite the advancements in biometric systems, especially facial recognition technologies, research on their performance for minority groups is limited. To our knowledge, this work presents a proof of concept study evaluating facial recognition technologies for individuals with different disabilities using state-of-the-art models. The results showed that VGGFace and FaceNet effectively recognized individuals with disabilities, with no significant performance differences between disabled and non-disabled populations.

## References


- Albiero, V. a. K. K. a. V. K. a. Z. K. a. K. M. C. a. B. K. W., 2020. Analysis of gender inequality in face recognition accuracy. Proceedings of the ieee/cvf winter conference on applications of computer vision workshops, pp. 81--89.
- Ameri, M. a. S. L. a. A. M. a. B. F. S. a. M. P. a. K. D., 2018. The disability employment puzzle: A field experiment on employer hiring behavior. ILR Review, 71(2), pp. 329--364.
- Angwin, J. a. L. J. a. M. S. a. K. L., 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, Volume 23, pp. 77--91.
- Baldwin, M. L. a. C. C., 2014. Re-examining the models used to estimate disability-related wage discriminatio. Applied Economics, 46(12), pp. 1393--1408.
- Bolukbasi, T. a. C. K.-W. a. Z. J. Y. a. S. V. a. K. A. T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, Volume 29.
- Buolamwini, J. a. G. T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency, pp. 77--91.
- Chowdhury, A., 2011. Revolution in authentication process by using biometrics. International Conference on Recent Trends in Information Systems, pp. 36--41.
- DeLeire, T., 2001. Changes in wage discrimination against people with disabilities: 1984-93. Journal of Human Resources, pp. 144--158.
- Deng, J. a. G. J. a. X. N. a. Z. S., 2019. Arcface: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4690--4699.
- Drozdowski, P. a. R. C. a. D. A. a. D. N. a. B. C., IEEE Transactions on Technology and Society. Demographic bias in biometrics: A survey on an emerging challenge. 89--103, 1(2), pp. 89--103.
- Guo, A. a. K. E. a. V. J. W. a. W. H. a. M. M. R., 2020. Toward fairness in AI for people with disabilities SBG@ a research roadmap. ACM SIGACCESS accessibility and computing, Issue 125, pp. 1--1.
- Hernandez-Ortega, J. a. G. J. a. F. J. a. H. R. a. B. L., 2019. Faceqnet: Quality assessment for face recognition based on deep learning. 2019 International Conference on Biometrics (ICB), pp. 1--8.
- Kane, S. K. a. G. A. a. M. M. R., 2020. Sense and accessibility: Understanding people with physical disabilities' experiences with sensing systems. Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility, pp. 1--14.
- Kortli, Y. a. J. M. a. A. F. A. a. A. M., 2020. Face recognition systems: A survey. Sensors, 20(2), p. 342.
- Krishnapriya, K. a. A. V. a. V. K. a. K. M. C. a. B. K. W., 2020. Issues related to face recognition accuracy varying based on race and skin tone. IEEE Transactions on Technology and Society, 1(1), pp. 8--20.
- Kruse, D. a. S. L. a. R. S. a. A. M., 2018. Why do workers with disabilities earn less? Occupational job requirements and disability discrimination. British Journal of Industrial Relations, 56(4), pp. 798--834.
- Liao, S. a. J. A. K. a. L. S. Z., 2012. Partial face recognition: Alignment-free approach. IEEE Transactions on pattern analysis and machine intelligence, 35(5), pp. 1193--1205.
- Meng, Q. a. Z. S. a. H. Z. a. Z. F., 2021. Magface: A universal representation for face recognition and quality assessment. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14225--14234.
- Organization, W. H., 2024. World Health Organization Disability. [Online]
- Available at: https://www.who.int/health-topics/disability#tab=tab_1
- [Accessed 19 Aug 2024].
- Parkhi, O. a. V. A. a. Z. A., 2015. Deep face recognition. BMVC 2015-Proceedings of the British Machine Vision Conference 2015.
- Rathgeb, C. a. D. P. a. F. D. C. a. D. N. a. B. C., 2022. Demographic fairness in biometric systems: What do the experts say?. IEEE Technology and Society Magazine, pp. 71--82.
- Rathgeb, C. a. I. M. a. H. D. a. H. S. a. {. B., 2024. Testing the Performance of Face Recognition for People with Down Syndrome. arXiv preprint arXiv:2405.11240.
- Rathgeb, C. a. K. J. a. U. A. a. B. C., 2023. Deep learning in the field of biometric template protection: An overview. arXiv preprint arXiv:2303.02715.
- Schroff, F. a. K. D. a. P. J., 2015. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815--823.
- Trewin, S., 2018. AI fairness for people with disabilities: Point of view. arXiv preprint arXiv:1811.10670.