das-Face Performance Report
v3.0

Revisión	Fecha	Descripción	Redactado	Revisado	Aprobado
1	23/02/2021	Update with dasFace v3.0	PZM/MSY	MSY/JGC	MSY

1. Introduction

das-Face is the facial biometric engine designed and developed by Veridas Digital Authentication Solutions S.L. with the goal of performing automatic identity verification (1:1) and identification (1:N) under different scenarios.

In this document, a performance analysis of **das-face** is summarised. Particularly, **verification (1:1)** task will be evaluated against:

- LFW dataset [1], including 13.233 images (from 5749 identities)
- MegaFace evaluation dataset [2], including 106.863 images (from 530 identities) and 1 million distractors (from 690.572 identities).
- IJB-C protocol covariate [6], consisting of 140.739 images (from 3.531 identities), used at Face Recognition Vendor Test organized by NIST.
- RFW dataset [7], including 4 subsets of different races with 6.000 biometric comparisons each.

Additionally, an **identification (1:N)** evaluation has been carried out against the MegaFace evaluation dataset.

These evaluation standards are commonly used in the literature for system evaluation and comparison against the state-of-the-art purposes[1].

das-Face, besides the two operations mentioned above, is capable of performing **liveness detection** using a *passive* procedure based on a selfie image, and a challenge-response *active* methodology based on a selfie and an annotated video. This document shows performance of das-Face for both use cases.

Veridas active liveness detection implemented in Selfie-Alive Pro was tested by iBeta to the **ISO 30107-3 Biometric Presentation Attack Detection Standard** and was found to be in compliance with Level 1.[2]

The document's content is divided as follows: In Section 2, definitions for understanding better analysis and results. In Section 3, analysis of *das-Face* performance in terms of the verification task (1:1), and related to other state-of-the-art proposals. In Section 4, analysis of *das-Face* performance in terms of the identification task (1:N), also in comparison with other

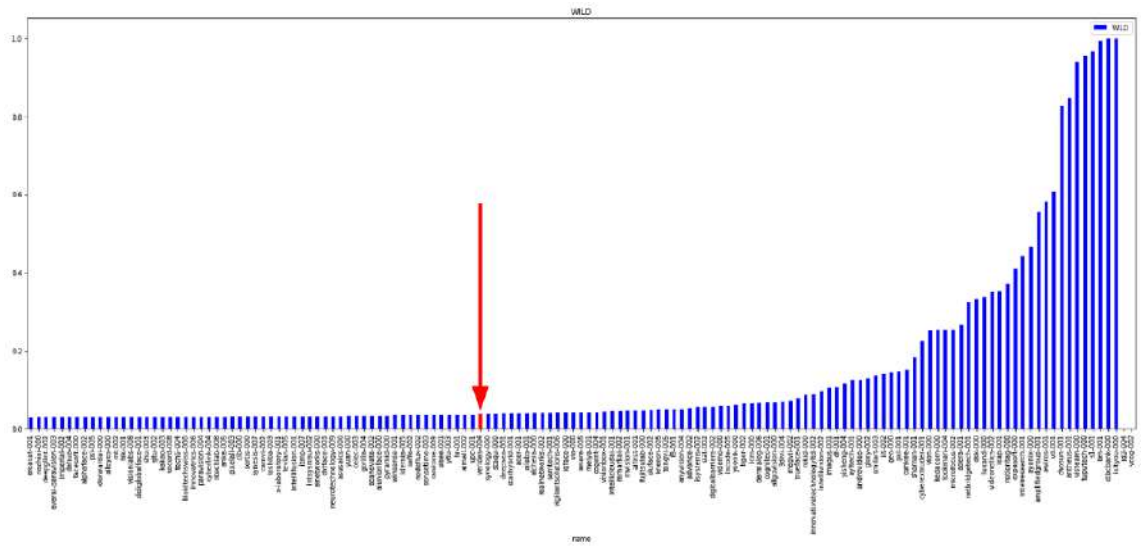
[1] For the MegaFace evaluation standard, performance results for different proposals are periodically updated in their webpage (http://megaface.cs.washington.edu/). However, some concerns have been reported about the validity of such evaluations, as further image processing than the one described in the protocol may have been performed. Under such conditions, it has been determined to compare *das-Face* only with the results shown in MegaFace reference, which do not pose any concerns of the kind.

[2] https://www.ibeta.com/wp-content/uploads/2020/12/201215-Veridas-PAD-Level-1-Confirmation-Letter.pdf

state-of-the-art proposals. In Section 5, the system calibration for the main use cases is presented. Section 6 presents information on the accuracy of the liveness detection engine.

The face recognition engine developed by Veridas was ranked by NIST as the third best in the world in the WILD category on April 4th, 2019, and it's the subject of continuous development and improvement efforts.

The face recognition engine developed by VERIDAS was ranked by NIST in the top 40% of the systems presented to FRVT 1:1 to the WILD category. The evaluation was performed on 2020 July.[3] Find below a picture of all the competitors in the mentioned WILD category. The VERIDAS system has been marked in red (Results shown from NIST do not constitute an endorsement of any particular system, product, service, or company by NIST.)[4]



VERIDAS achieved a False Non Match Rate (FNMR) of 2.91% for a False Match Rate (FMR) threshold fixed at 0.01%. These figures put VERIDAS a less than 0.2 points from the Top-3 system who achieved a FNMR of 2.73%. Taking into account these results, VERIDAS will comply with the requirements of FIDO for facial biometric verification systems. Specifically, FIDO states that FNMR should be less than 5% for a FMR of 0.01%.[5]

The WILD category is characterized by a non-collaborative subject, so the person whose face is being captured does not have to be facing the camera, and the

[3] https://github.com/usnistgov/frvt/raw/nist-pages/reports/11/frvt_11_report_2020_07_27.pdf
[4] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing
[5] https://fidoalliance.org/specs/biometric/Biometrics-Requirements-v1.0-wd-20190606.html#Performance

picture could show different issues in terms of illumination, contrast, exposure, … Because of the on-boarding pictures nature, the procedure may be similar to WILD category because the person is taking a picture in uncontrolled conditions.

2. Definitions

Definitions to better understand the analysis and results:

- **Negative evaluation**: Evaluation of two images belonging to two different people.
- **Positive evaluation**: Evaluation of two images belonging to the same person.
- **Bonafide**: A presentation attempt that is performed by a trustworthy person.
- **Attack**: A presentation attempt that is performed by an impostor or spoofer.
- **Accuracy**: Percentage of correct answers provided by the system.
- **False Positive Rate (FPR) or False Match Rate (FMR)**: Ratio between the number of negative evaluations wrongly categorized as positive and the total number of actual negative evaluations.
- **True Positive Rate (TPR)**: Ratio between the number of positive evaluations correctly categorized as positive and the total number of actual positive evaluations.
- **False Non Match Rate (FNMR)**: Ratio between the number of positive evaluations rejected by the system and the total number of actual positive evaluations.
- **Identification Rate (IR)**: Ratio between the number of successful identifications and the total number of performed identifications.
- **Attack Percentage Classification Error (APCER)**: Is the ratio between the number of spoof attacks misclassified as authentic over the total number of performed attacks.
- **Bonafide Percentage Classification Error (BPCER)**: Is the ratio between the number of bonafide (actual person's faces) misclassified as attacks over the total number of performed bonafide samples.
- **Verification task (1:1)**: Use case in which two different images containing the face of a person are presented to the system for it to determine if they are (or not) the same person.
- **Identification task (1:N)**: Use case in which an image containing the face of a person is presented to the system, having the system access to a pool of N images each corresponding to an identity, in order for the system to determine to which of the N identities (if any) the presented image belongs to.
- **National Institute of Standards and Technology (NIST)**: Measurement standards laboratory whose mission is to promote innovation and industrial competitiveness.
- **Liveness detection**: An automatic procedure whose purpose is to detect how likely the captured evidence (images, videos, …) belong to an actual person and not to a spoofed sample of a person.

3. *das-Face* performance in identity verification (1:1)

In this section, a performance analysis for *das-Face* will be carried out, using different evaluation standards: LFW dataset, MegaFace dataset, the IJB-C covariate protocol and the RFW dataset.

3.1. 1:1 task against the LFW evaluation standard

The LFW evaluation dataset includes 13.233 images belonging to 5.749 identities from celebrities and has been widely used by facial biometry systems for comparison against the state-of-the-art purposes.

Results against this dataset for different relevant proposals and in terms of the system accuracy (measured by providing mean and uncertainty values) are shown in **Table I**. *das-Face* achieves a performance level comparable with top-notch face verification systems as FaceNet, developed by Google and DeepFace, developed by Facebook, or the system developed by Baidu.

Face biometry system	Accuracy (%)
das-Face	***99.78 ± 0.24***
Baidu ensemble model [3]	99.77
Baidu single model [3]	99.68
FaceNet (Google) [4]	99.63 ± 0.09
Human (overestimated)	97.53
DeepFace (Facebook) [5]	97.35 ± 0.25

Table I. State-of-the-art face biometric engines performance for identity verification against the LFW evaluation standard.

3.2. 1:1 task against the MegaFace evaluation standard

The MegaFace evaluation standard includes two datasets, one containing genuine identities and the other one containing distractor identities, which will be put in comparison to evaluate both identity verification and face identification. The genuine identities dataset, FaceScrub, includes 106.863 images belonging to 530 different identities of celebrities. Additionally, the distractor dataset includes 1 million images belonging to 690.572 different identities.

This evaluation standard has also been widely used for system performance evaluation in state-of-the-art systems, and their performance results could be used as a standard comparison. Specifically, in Figure 1, results obtained for different facial biometry engines in the identity verification task are shown. In this case, performance is assessed based on different FPR - TPR operating points.

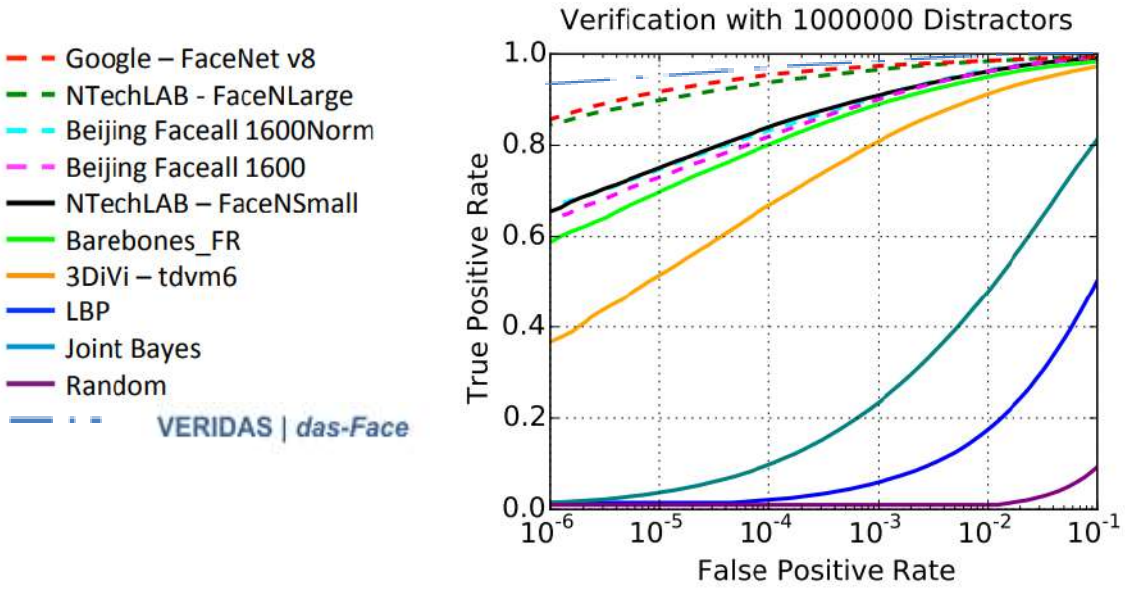


Figure 1. Performance of different state-of-the-art systems evaluated with the MegaFace evaluation standard. Image obtained from [2].

In Table II below, results of *das-Face* face biometric engine under the same conditions are shown.

FPR (%)	TPR (%)
0.1	98.58
0.01	98.03
0.001	97.97
0.0001	97.24
0.00001	96.99
0.000001	96.68

Table II. Performance results of *das-Face* against the MegaFace evaluation standard for the verification task.

Additionally, performance results of *das-Face* versus other state-of-the-art systems are shown in Table III, highlighting True Negative results for a predefined level of False Positive results of 10^{-4}:

Face biometry system	TPR (%)
das-Face	***97.24***
Google - FaceNet v8	95.0
NTechLAB - FaceNLarge	93.9
NTechLAB - FaceNSmall	83.9
Beijing FaceAll 1600Norm	83.7
Beijing FaceAll 1600	81.7

Table III. System performance for verification task (1:1) under MegaFace evaluation standard in terms of TPR @ FPR=10^{-4}.

It can be seen from results provided in Tables II and III that *das-Face* performance is located above Google's FaceNet v8.

3.3. 1:1 task against the IJB-C evaluation standard

The IJB-C evaluation standard consists of different evaluation protocols. Covariate protocol was used at FRVT NIST challenge until 2018 June, replaced afterwards by a larger dataset. Therefore, this dataset is a good evaluation benchmark, allowing VERIDAS to compare results with NIST competitors. Covariate 1:1 verification protocol uses single image templates to allow further analysis of an algorithm's performance on individual covariates. All images were obtained from creative videos and photographs. The protocol includes 140.739 templates taken from 3.531 identities. **Table IV** shows comparison in terms of TPR @ FPR=10^{-4}, taking into account a subset of the 62 participants in the FRVT NIST challenge at date 2018/04/03.

Face biometry system	TPR (%)
das-Face	***84.5***
NTechLab (1st)	72.9
NTechLab (2nd)	67.6
FDU (3rd)	65.6

VisionLabs (4th)	63.1
VisionLabs (5th)	60.3
NeuroTechnology (6th)	57.9
NeuroTechnology (7th)	57.3
Itmo (10th)	55.0
Tiger (20th)	40.7
AnyVision (30th)	31.6
Innovatrics (40th)	23.2
Itmo (50th)	10.8
IsItYou (60th)	0.0

Table IV. System performance for verification task (1:1) under IJB-C covariate evaluation standard in terms of TPR @ FPR=10^{-4}.

3.4. 1:1 task against the RFW evaluation standard

The dataset RFW (Racial Faces in-the-Wild) [7] proposes to analyze racial bias of facial biometric systems. The dataset is divided into different subsets: Caucasian, Indian, Asian and African. Each subset contains 6.000 biometric pairs to compute the accuracy of the system. **Table V** below shows the results taken from referenced report attached (Reference [7]) and das-Face performance on the same set of data

Model	**Caucasian**	**Indian**	**Asian**	**African**
das-Face	***98.72%***	***97.62%***	***97.39%***	***97.69%***
Microsoft	87.60%	82.83%	79.67%	75.83%
Face++	93.90%	88.55%	92.47%	87.50%
Baidu	89.13%	86.53%	90.27%	77.97%
Amazon	90.45%	87.20%	84.87%	86.27%

Table V System performance for verification task (1:1) under RFW evaluation standard in terms accuracy.

4. *das-Face* performance in identification (1:N)

The MegaFace evaluation standard allows to assess a face biometric engine under the identification task in which a particular identity is searched within a pool of an N number of distractors. Under this use case, reported state-of-the-art results are shown in Figure 2.

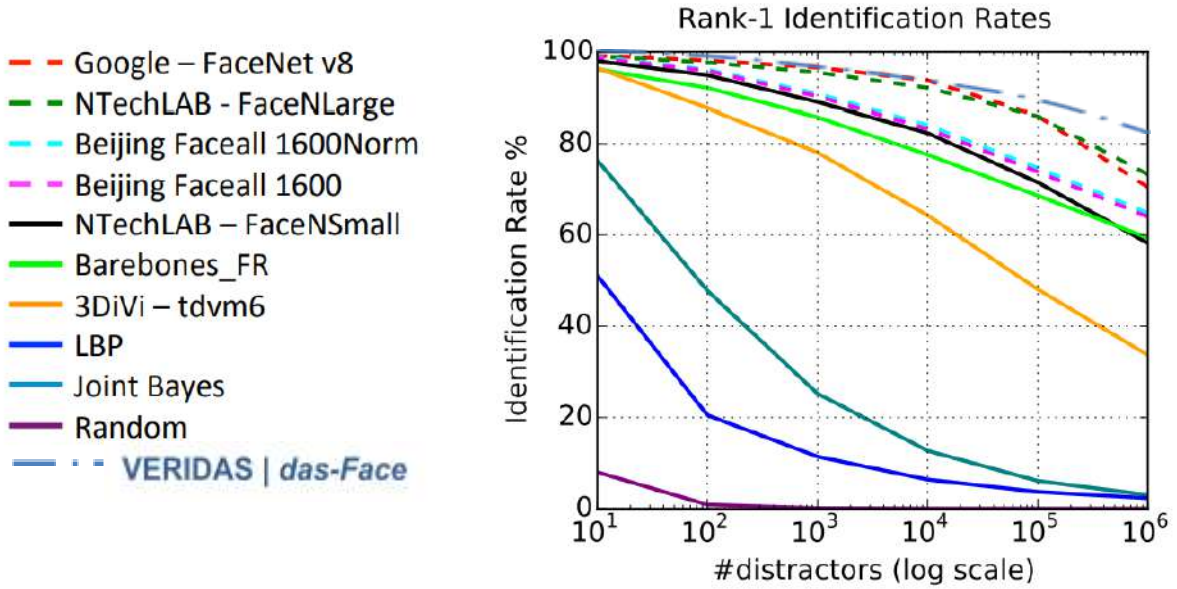


Figure 2. Performance of different state-of-the-art face biometric engines against the MegaFace evaluation standard for the identification task. Image obtained from [2].

Detail of das-Face performance under the same conditions is summarised in Table VI.

# Distractors	IR (%)
10^1	98.70
10^2	98.25
10^3	98.03
10^4	97.79
10^5	94.29
10^6	80.61

Table VI. Performance results of *das-Face* against the MegaFace evaluation standard for the identification task.

Additionally, summary of performance in terms of TPR at 10.000 distractors for same participants and *das-Face* is shown in Table VII:

Face biometry system	IR (%)
das-Face	***97.8***
Google - FaceNet v8	93.2
NTechLAB - FaceNLarge	91.7
Beijing FaceAll 1600Norm	84.2
Beijing FaceAll 1600	83.0
NTechLAB - FaceNSmall	82.2

Table VII. System performance for the identification task under MegaFace evaluation standard in terms of TPR @ 10k distractors.

Based on the results above, das-face performs better than all **2017 MegaFace Recognition Challenge participants.**

5. Face verification technologies

5.1. Selfie vs Selfie

When using the system with selfie photos, the response may change because of the characteristics of this particular use case. Results of the system for the case of selfie-vs-selfie are presented in **Table VIII**, evaluated using an internal database created for this purpose.

Similarity Threshold	FPR (%)	FNR (%)
0.50	0.061	0.065
0.55	0.039	0.080
0.60	0.026	0.091
0.65	0.019	0.100
0.70	0.013	0.118

0.75	0.010	0.147
0.80	0.008	0.203
0.85	0.006	0.259
0.90	0.004	0.406
0.95	0.002	1.069

Table VIII System performance for selfie vs selfie

For instance, choosing a 0.80 as threshold, all biometric comparisons with score above 0.80 will be considered as the same person, and all comparisons with a score below 0.80 will be considered as different persons. For 0.80, in the case of selfie vs selfie, a 0.203% of the comparisons of a person's selfies will be rejected (false negative), and only a 0.006% of the cases will be incorrectly classified as authentic (false positive).

5.2. Selfie vs ID Document

When using the system to compare a selfie photo and an identity document photograph crop, the response may change again because of the characteristics of this particular use case.

The influence of the ID document manufacturing process, the effect of environmental conditions during the capture process of both the document and the selfie, the presence of visual artifacts in the document image, the effect of the capture technology and the lens used, the possible facial complements a person may wear, as well as the time difference between the two photos, make the biometric comparison process in a digital onboarding process extremely variable.

Images printed on European identity documents

Selfie images from onboarding processes

In this case, the **Table IX** is more suitable to state the behavior of the system. The system has been trained with document and selfie pairs, in order to adapt the system to this use case. The table has been computed by using an internal testing dataset, created for this purpose, with 3.416 real cases of selfie and document images.

Similarity Threshold	FPR (%)	FNR (%)
0.50	2.51	1.78
0.55	1.84	2.01
0.60	1.29	2.12
0.65	0.95	2.26
0.70	0.62	2.43
0.75	0.39	2.73
0.80	0.23	2.96
0.85	0.13	3.31
0.90	0.05	4.04
0.95	<0.01	5.52

Table IX System performance for selfie vs ID Document

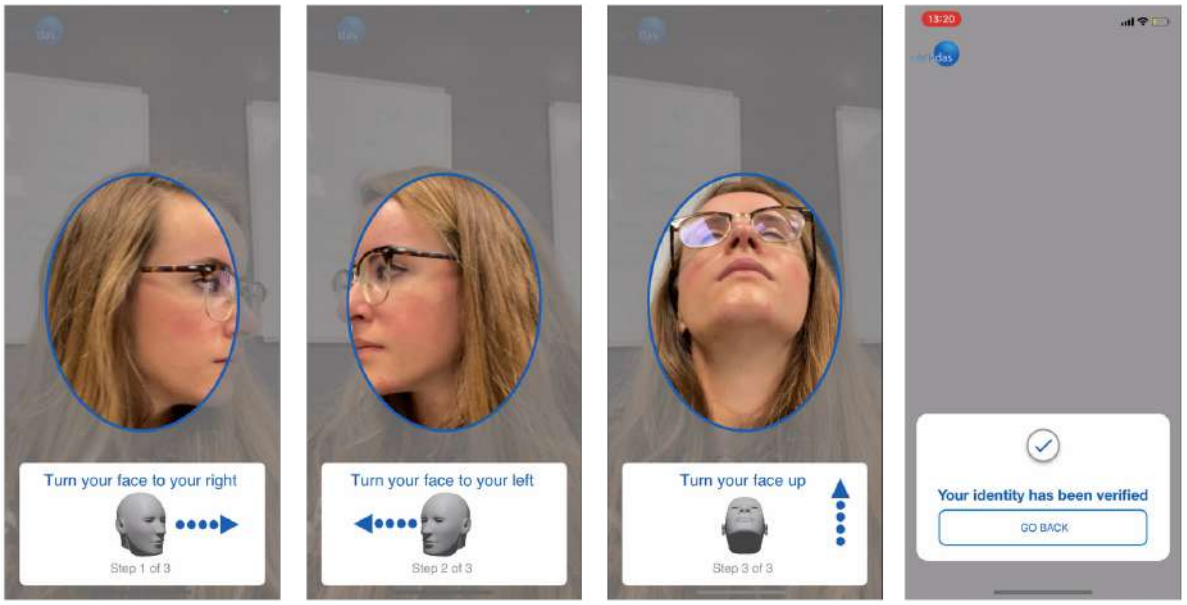
Based on our experience, the operation point is usually at least 0.70. For instance, choosing a 0.70 as threshold, all biometric comparisons with score above 0.70 will be considered as the same person, and all comparisons with a score below 0.70 will be considered as different persons. For 0.70, in the case of selfie vs document, a 2.43% of the comparisons of a person selfie and its corresponding legit ID card will be rejected (false

negative), and only a 0.62% of the cases comparing a selfie and a ID card corresponding to different persons will be incorrectly classified as authentic (false positive).

6. Liveness detection technologies

6.1. SDK Selfie Alive Pro

das-Face incorporates also an active liveness detection procedure based on a challenge-response method. das-Face generates a challenge that is consumed by Selfie-Alive Pro (SAP) SDK, and the device will start the interaction with the user. During the interaction, the user is asked to capture a selfie photograph and to record a small video of



his face performing a few random head movements. The number of random movements is configurable by the integrator, we recommend 2 movements as standard, and 6 movements for maximum security. Once everything is recorded, the SDK will delegate all the captured evidence, and the device must send all the data back to the das-Face server for its processing. das-Face will analyze the video and selfie data looking for liveness evidence.

Veridas active liveness detection implemented in Selfie-Alive Pro was tested by iBeta to the **ISO 30107-3 Biometric Presentation Attack Detection Standard** and was found to be in compliance with Level 1.[6]

[6] https://www.ibeta.com/wp-content/uploads/2020/12/201215-Veridas-PAD-Level-1-Confirmation-Letter.pdf

Having achieved this result, and because the ISO 30107-3 testing was performed with 2D printouts, paper masks, 3D layered photos, and replayed photos in screens, the Selfie-Alive Pro solution can be found into the levels A and B indicated by FIDO recommendations.[7]

The performance of the system has been measured for different scenarios:

- **Replay-attacks with different screens**: This kind of attack consists of using a screen to reproduce the picture of another person's face and capturing it via the spoofer device camera.
- **Video-replay-attacks with different screens**: This one are video compositions showing the person face executing the challenge, it requires full collaboration of the subject to be spoofed because we need to gather videos of his head performing the movement.
- **3D animation and deep fake attacks**: These ones are attacks where the face of this person is artificially generated and animated using a software.
- **Print-attacks in different qualities**: This kind of attack involves one or more printouts of a human face.
- **Print-mask-attack in different qualities**: This time the attack consists of printing the face into paper and cutting it to build a sort of 2D mask.
- **3D-layered-photo-mask**: This mask is a composition of several printed photographs glued together to imitate some 3D effect.

The system is trained in several databases combining different kinds of attacks. The system's performance has been evaluated over an internal database composed of 695 authentic samples and 1104 attacks. The numbers with this dataset are depicted in **Table X**:

[7] https://fidoalliance.org/specs/biometric/requirements/#TriagePAD

Liveness Threshold	APCER (%)	BPCER (%)
0.50	10.2	0.8
0.55	8.4	1.3
0.60	6.5	1.7
0.65	5.4	1.8
0.70	3.0	3.0
0.75	1.9	3.1
0.80	1.2	4.1
0.85	0.7	6.0
0.90	0.23	10.7
0.95	<0.23	20.6

Table X System performance for SAP

Based on **Table X**, using an operation point at 0.70, the 3% of authentic cases will be rejected, the 3% of spoofing attempts will be misclassified as authentic.

Optimal performance requires following constraints:

- All evidence must be kept as returned by the SDK, any additional compression may lead to accuracy problems.
- Face must be of 150px width to ensure it can be processed by the anti-spoofing system. To ensure accuracy, we recommended faces with more than 320px width.
- Face is expected to be frontal with the camera in the selfie.
- Face movements should be smooth during the video record.

Based on **Table X**, the following thresholding criteria are recommended:

- When the threshold > 0.7 the attempt is classified as “bona fide”.
- When the threshold < 0.5 the attempt is classified as “attack”.
- When the threshold is in between 0.5 and 0.7 the attempt is doubtful and should be reviewed by a human operator.

6.2. Passive Liveness Detection Engine

das-Face also includes a passive liveness detector designed to avoid fraudulent access. Given a selfie photo, the detector estimates a score of the photo being captured from an actual person's face. The performance of the system has been measured in replay-attack scenarios:

- Replay-attacks with different screens: This kind of attack consists of using a screen to reproduce the picture of another person's face and capturing it via the spoofer device camera.

The system is trained in several databases combining different kinds of attacks. The evaluation of the system has been performed on an internal database with 1000 authentic selfies and 1104 presentation attacks, achieving a 95.0% overall accuracy.

The anti-spoofing performance is shown in **Table XI**. Notice the performance of the system is shown for different authenticity thresholds, i.e., 1.00 means authentic and 0.00 means spoof attempt.

Liveness Threshold	REPLAY APCER (%)	BPCER (%)
0.50	23.5	1.0
0.55	17.4	1.8
0.60	13.0	1.9
0.65	11.0	2.4
0.70	7.8	3.2
0.75	4.0	4.8
0.80	1.1	7.5
0.85	0.2	9.0
0.90	<0.2	10.7
0.95	<0.2	14.1

Table XI System performance for passive liveness detection

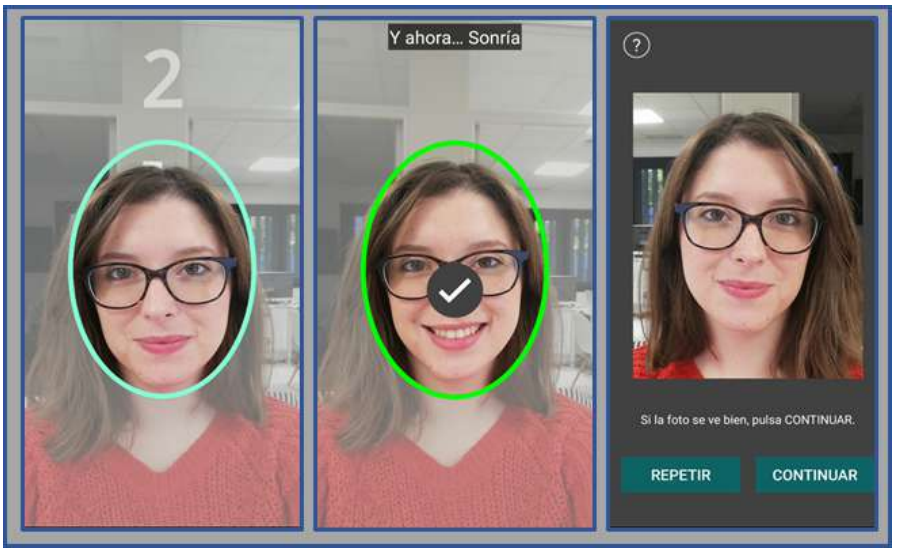
Based on **Table XI**, using an operation point at 0.70, the 3.2% of authentic cases will be rejected, the 7.8% of replay-attack spoofing attempts will be misclassified as authentic.

Based on **Table XI**, the following thresholding criteria are recommended:

- When the threshold >= 0.7 the attempt is classified as “bona fide”.
- When the threshold < 0.5 the attempt is classified as “attack”.
- When the threshold is in between 0.5 and 0.7 the attempt is doubtful and should be reviewed by a human operator.

6.3. SDK Selfie Alive

This system performance shown in **Table XI** is dependent on the input image quality. In order to ensure the indicated performance, it is mandatory to use the Veridas Selfie Alive SDK for iOS, Android, or mobile HTML. Using other capture procedures may harm the correct operation of the system.



Selfie Alive SDK example for iOS and Android

The Selfie Alive SDK asks the user to perform an action to capture the Selfie photo. This functionality prevents the use of static photos represented on a screen or printed on paper. Once the photo has been captured, das-Face processes the image using the anti-spoofing techniques described in the previous section (see **Table XI**).

Optimal performance requires following constraints in given images:

- Images must be kept as returned by the SDK, any additional compression may lead to accuracy problems.
- Face must be of 150px width to ensure it can be processed by the anti-spoofing system, which allows to process images captured via Veridas HTML SDKs. To ensure accuracy, we recommended faces with more than 320px width, which is feasible when using native SDKs.

- Face is expected to be frontal with the camera.

7. References

[1] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *"Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments"*. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

[2] Ira Kemelmacher-Shlizerman, Steve Seitz, Daniel Miller, Evan Brossard. *"The MegaFace Benchmark: 1 Million Faces for Recognition at Scale".* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] J. Liu, Y. Deng, and C. Huang. *"Targeting ultimate accuracy: Face recognition via deep embedding"*. arXiv:1506.07310, 2015.

[4] F. Schroff, D. Kalenichenko, and J. Philbin. *"Facenet: A unified embedding for face recognition and clustering"*. CVPR, 2015.

[5] Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. *"Deepface: closing the gap to human-level performance in face verification"*. Proc. Conference on Computer Vision and Pattern Recognition 1701–1708 (2014).

[6] Maze, B., et al. "IARPA Janus Benchmark – C: Face Dataset and Protocol". 11th IAPR International Conference on Biometrics (2018).

[7] Wang, M., et al. "Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network". arXiv:1812.00194, 2019.

[8] FIDO Alliance. Biometric Requirements v1.0 (PAD criteria). 2019. url: https://fidoalliance.org/specs/biometric/requirements/ (visited on 2020-07-10).