das-Face Performance Report
v3.2

Revisión	Fecha	Descripción	Redactado	Revisado	Aprobado
1	23/02/2021	Update with dasFace v3.0	PZM/MSY	MSY/JGC	MSY
2	23/04/2021	Update with NIST 1:N results	PZM/MSY	MSY/JGC	MSY
3	06/05/2021	Add NIST 1:1 results	JGC	MSY/PZM	MSY
4	22/06/2021	Update to dasFace v3.2	JGC/DGS	MSY/PZM	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)** is evaluated against multiple datasets internally

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

Additionally, an **identification (1:N)** evaluation of our model has been carried out by NIST as part of the 2021 March FRVT report.[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.

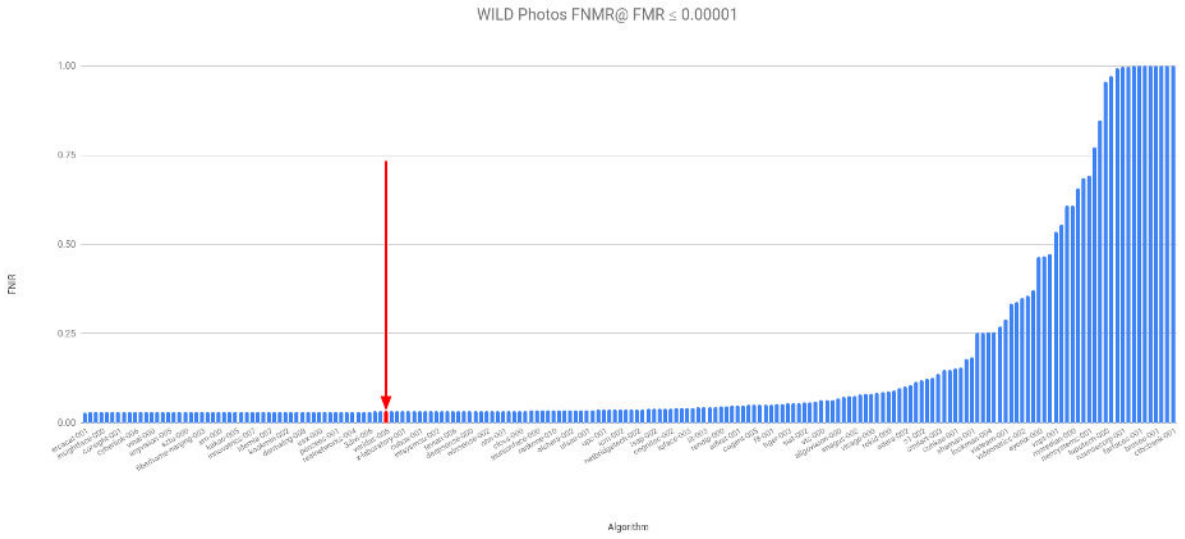
Veridas active liveness detection implemented in Selfie-Alive Pro was tested by iBeta to the **ISO 30107-3 Biometric Presentation Attack Detection Standard** and was found to be in compliance with Level 1.[2]

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). In Section 4, analysis of *das-Face* performance in terms of the identification task (1:N. 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.

[1] https://pages.nist.gov/frvt/html/frvt1N.html
[2] https://www.ibeta.com/wp-content/uploads/2020/12/201215-Veridas-PAD-Level-1-Confirmation-Letter.pdf

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 25% of the systems presented to FRVT 1:1 to the WILD category. The evaluation was performed on 2021 April.[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 2.86% for a False Match Rate (FMR) threshold fixed at 0.01%. These figures put VERIDAS a less than 0.3 points from the Top-3 system. 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 picture could show different issues in terms of illumination, contrast, exposure, …

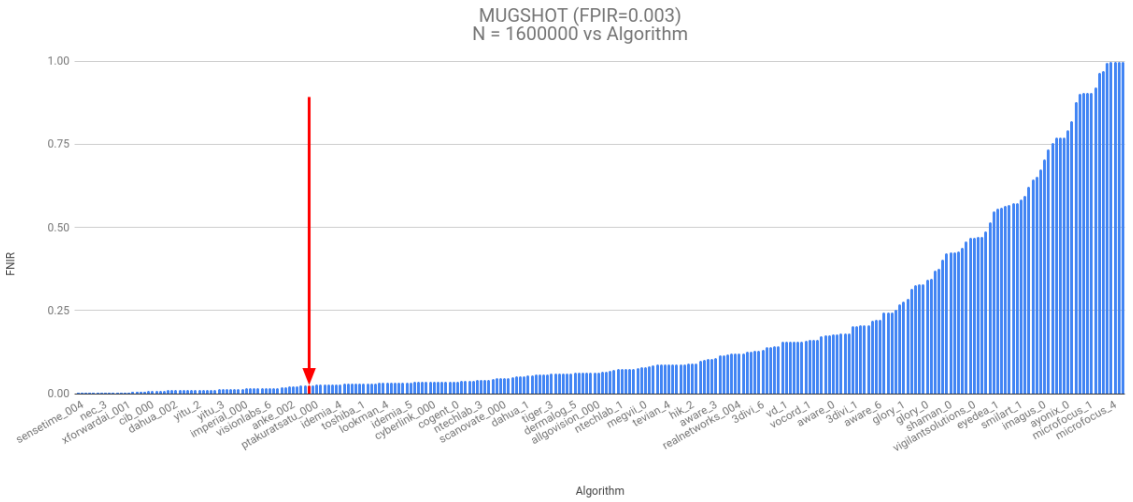
[3] https://github.com/usnistgov/frvt/blob/nist-pages/reports/11/frvt_11_report_2021_04_16.pdf
[4] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing
[5] https://fidoalliance.org/specs/biometric/Biometrics-Requirements-v1.0-wd-20190606.html#Performance

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.

The face recognition engine developed by Veridas was ranked by NIST in the top 25% best systems in the world in the MUGSHOT category on April 16th, 2021, and it's the subject of continuous development and improvement efforts.

The face recognition engine developed by VERIDAS was ranked by NIST in the 63 of 271 systems presented to FRVT 1:N to the MUGSHOT category. The evaluation was performed on 2021 April.[6] Find below a picture of all the competitors in the mentioned MUGSHOT 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.)[7]

VERIDAS achieved a False Negative Identification Rate (FNIR) of 2.44% for a False Positive Identification Rate (FPIR) threshold fixed at 0.3%, and with a gallery with N=1.6M.



The MUGSHOT category is characterized by a collaborative subject almost following ISO 19794-5, so the person whose face is being captured is in good acquisition conditions.

2. Definitions

[6] https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf
[7] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

Definitions to better understand the analysis and results:

- **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.
- **National Institute of Standards and Technology (NIST)**: Measurement standards laboratory whose mission is to promote innovation and industrial competitiveness.
- **Negative evaluation**: Evaluation of two images belonging to two different people.
- **Positive evaluation**: Evaluation of two images belonging to the same person.
- **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 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.
- **Identification Rate (IR)**: Ratio between the number of successful identifications and the total number of performed identifications.
- **Identification Rank (R)**: Upper bound of the position where the match should be in the list of candidates returned by the system in order to consider it a successful match.
- **False Negative Identification Rate (FNIR):** Ratio between the number of positive identifications rejected because they do not achieved the threshold, over the total number of actual positive identifications.
- **False Positive Identification Rate (FPIR):** Ratio between the number of negative identifications accepted because they do achieved the threshold, over the total number of negative identifications.
- **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.
- **Bonafide**: A presentation attempt that is performed by a trustworthy person.
- **Attack**: A presentation attempt that is performed by an impostor or spoofer.
- **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.

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

In this section, the latest version of das-Face performance is carried out, using NIST evaluations as a reference.

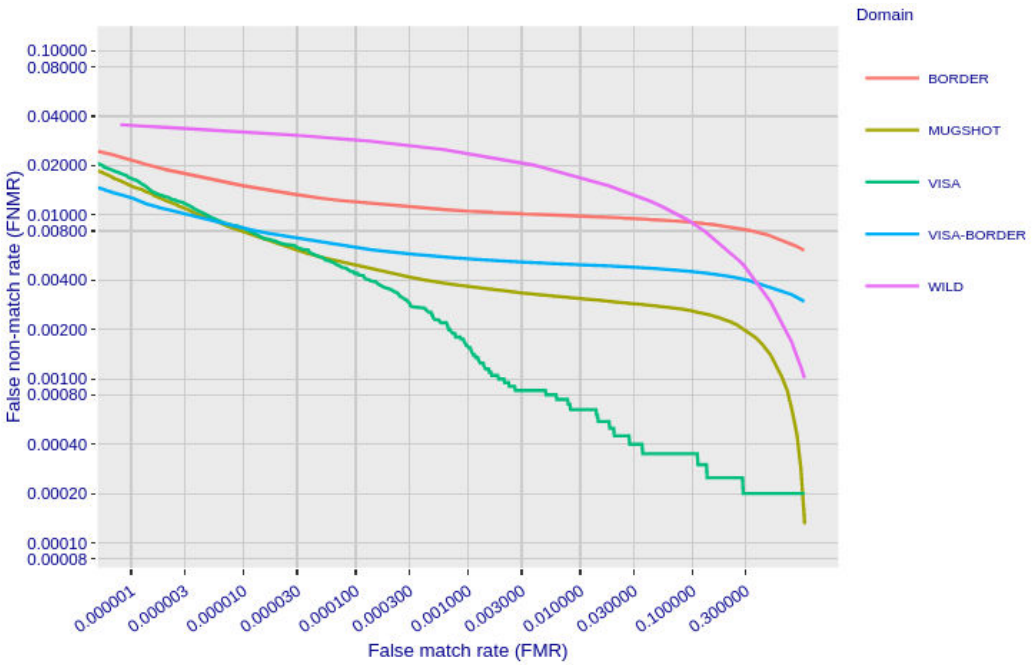
3.1. 1:1 Ongoing Face Recognition Vendor Test (FRVT) - Verification

This section describes the most relevant results on our model in the NIST 1:1.

On verification, multiple datasets are used when the system is tested. Currently the next types of datasets are used on NIST benchmarks:

- Wild images: Typical social-media images, with many photojournalism-style images. Images are given to the algorithm. This category is the most difficult one, as it is not collaborative.
- Mugshot images: Photographic portrait of a person from the waist up, typically taken after a person is arrested. These photos have a reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type.
- VISA images: Photos collected in immigration offices,
- Border images: Border crossing images collected in primary immigration lanes.

The next graph report shows the system performance on different domains.



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

This section describes the most relevant results of our biometric model in the NIST FRVT 1:N.[8] In identification, a particular identity (probe) is searched within a pool of N number of known identities (gallery). The next Table reports how the system FNIR is affected by the size N of the gallery, fixed an operational point for FPIR=0.1% and using the dataset FRVT'2018 MUGSHOT, and when the identification rank is R=1.

N	VERIDAS-001 FNIR (%)
640K	2.78%
1.6M	3.73%
3M	4.91%
6M	7.53%
12M	15.40%

Table VI System performance for identification task (1:N) under FRVT 1:N evaluation for FRVT'2018 MUGSHOT dataset and different gallery sizes (N).

From this table it can be deduced that, if you have a database with 1.6 million people, the FPIR of 0.1% means that 1 out of every 1000 identifications where the probe is not in the gallery the system finds a match with a wrong person, and the shown FNIR means the system incorrectly rejects a 3.73% of the time a person that actually is in the gallery.

In the next Table, it is reported the FNIR of the system for different FPIR operational thresholds, when the gallery size is N=3k and N=640k, and when the identification rank is R=1.

Gallery size	FPIR (%)	FNIR(%)	Threshold
3.000	0.001%	3.0%	>0.99

[8] Results shown from NIST do not constitute an endorsement of any particular system, product, service, or company by NIST.
https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

	0.01%	1.3%	>0.98
	0.1%	0.4%	>0.96
	1%	<0.1%	>0.90
640.000	0.03%	8%	>0.98
	0.1%	3.7%	>0.97
	1%	1.6%	>0.95

Table VII Calibration curve for identification task (1:N)

The previous table shows that for a confidence over the threshold of 0.95, if the gallery database contains N=640K persons, the system finds false matches 1% of the time, and incorrectly rejects 1.6% of the people that actually are in the gallery.

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 IX**, evaluated using an internal database created for this purpose. This table is the same for Veridas Native and HTML SDKs (mobile & desktop).

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 IXI 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, 0.203% of the comparisons of a person's selfies will be rejected (false negative), and only a 0.008% 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

The facial biometrics engine is specifically trained for the selfie vs. document comparison use case, allowing for optimized performance. The Veridas biometric engine is robust in the following situations.

- Presence of glare in the printed photo area.
- Presence of the kinegram and other visual artifacts on the printed photo.
- Temporal difference between the selfie photo captured by the user and the printed photo.
- Changes in the face: presence of beard, mustache, hair changes, glasses, make-up, etc.
- Presence of face-mask.

In this case, the **Table X** 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. This table is the same for Veridas Native and HTML SDKs (mobile & desktop).

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

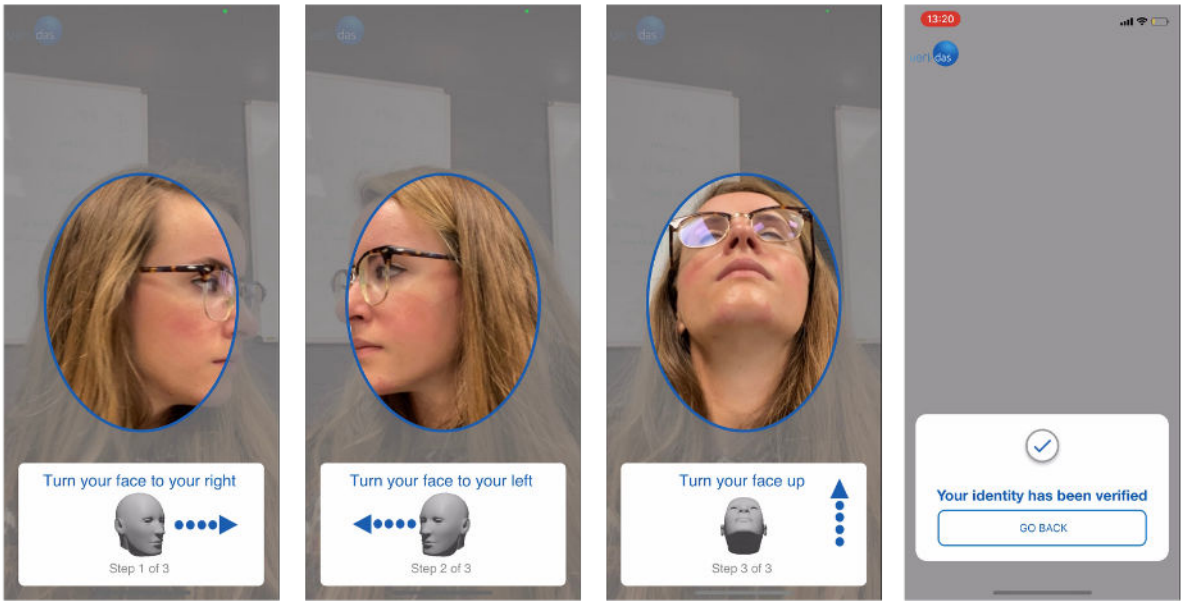
Table X System performance for selfie vs ID Document

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.[9]

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.[10]

[9] https://www.ibeta.com/wp-content/uploads/2020/12/201215-Veridas-PAD-Level-1-Confirmation-Letter.pdf
[10] https://fidoalliance.org/specs/biometric/requirements/#TriagePAD

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 XI**:

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 XI System performance for SAP

Based on **Table XI**, using an operation point at 0.70, the 3% of authentic cases will be rejected and 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.
    - This size allows processing of images taken with Native and HTML SDKs provided by Veridas. Using other capture procedures may harm the correct operation of the system.
- Face is expected to be frontal with the camera in the selfie.
- Face movements should be smooth during the video record.

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.

6.2. 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 and print-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.
- **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 evaluation of the system has been performed on an internal database with 1000 authentic selfies and 1104 presentation attacks, achieving a 95.5% overall accuracy.

The anti-spoofing performance is shown in **Table XII**. 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	22.77	0.7
0.55	15.94	1.1
0.60	12.73	1.7
0.65	8.93	2.3
0.70	5.98	3.0
0.75	4.22	4.9
0.80	3.12	8
0.85	1.85	10.9
0.90	1.51	14.8
0.95	0.42	23.1

Table XII System performance for passive liveness detection

Based on **Table XII**, using an operation point at 0.70, the 3.0% of authentic cases will be rejected, the 5.98% of spoofing attempts will be misclassified as authentic.

Based on **Table XII**, 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.

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

[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.