Iris Liveness Detection Competition (LivDet-Iris) – The 2025 Edition

Mahsa Mitcheff*[1†] Afzal Hossain*[2†] Samuel Webster*[1†] Siamul Khan[1†] Katarzyna Roszczewska[4†] Juan E. Tapia[6‡], Fabian Stockhardt[6‡] Lázaro J. González-Soler[6‡] Ji-Young Lim[7‡] Mirko Pollok[7‡] Felix Kreuzer[7‡] Caiyong Wang[8‡] Lin Li[8‡] Fukang Guo[8‡] Jiayin Gu[9‡] Debasmita Pal[10‡] Parisa Farmanifard[10‡] Renu Sharma[10‡] Arun Ross[10‡] Geetanjali Sharma[11‡] Shubham Ashwani[11‡] Aditya Nigam[11‡] Raghavendra Ramachandra[12‡] Lambert Igene[2†] Jesse Dykes[2†] Ada Sawilska[4†] Aleksandra Dzieniszewska[4†] Jakub Januszkiewicz[4†] Ewelina Bartuzi-Trokielewicz[5†] Alicja Martinek[5†] Mateusz Trokielewicz[4†] Adrian Kordas[5†] Kevin Bowyer[1†] Stephanie Schuckers[3†] Adam Czajka[1†]

[1]University of Notre Dame, IN, USA; [2]Clarkson University, NY, USA; [3]University of North Carolina-Charlotte, NC, USA; [4]PayEye Poland, Poland; [5]NASK – National Research Institute, Warsaw, Poland; [6]Hochschule Darmstadt (HDA), Germany; [7]DERMALOG Identification Systems GmbH, Germany; [8]School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture (BUCEA), Beijing, China; [9]School of Information and Cyber Security, People's Public Security University of China (PPSUC), Beijing, China; [10]Michigan State University, MI, USA; [11]School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Himachal Pradesh, India; [12]Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway

*Equal contribution †Organizers ‡Competitors Corresponding emails: mmitchef@nd.edu, afhossa@clarkson.edu

Abstract

LivDet-Iris 2025 is the sixth edition of the iris liveness detection competition. Held every two to three years, the competition aims to foster the development of robust algorithms capable of detecting a wide range of physically- and digitally-presented attacks in iris biometrics. The 2025 edition obtained the largest number of submissions in the history of the competition: ten algorithms from five institutions, and one commercial iris recognition system. LivDet-Iris 2025 also introduced new tasks compared to previous editions: (Task 1) a benchmark offered by an industry partner, (Task 2) morphed iris images, in which two different-identity samples were blended into one image, and (Task 3) evaluation of presentation attack detection robustness against advanced manufacturing techniques for textured contact lenses. This edition, for the first time in the series, offers a systematic testing of a commercial iris recognition system (software and hardware) using physical artifacts presented to the sensor. Dermalog-Iris team submitted algorithms that won all tasks, achieving the area under the ROC curve of 90.57%, 68.23% and 99.99% in tasks 1, 2, and 3, respectively. Additionally, we include results for baseline algorithms, based on modern deep convolutional neural networks and trained with all available public datasets of iris images representing bona fide samples and anomalies (physical attacks, eye diseases, post-mortem cases, and synthetically-generated iris images). Test samples created for tasks 2 and 3, and baseline models are made available to offer the state-of-the-art benchmark for iris liveness detection.

1. Introduction

LivDet-Iris 2025 is the sixth edition of the iris liveness detection competition in the LivDet-Iris series, and included in the official IJCB 2025 competition list[1]. Similar to previous competitions, this edition has two parts:

- **Part 1 (Algorithms)**, which involves the evaluation of the software solutions (submitted to the organizers) in three tasks, in which large datasets of iris images representing bona fide samples and various anomalies were used, and
- **Part 2 (Systems)**, which involves the systematic testing of submitted iris recognition systems based on physical artifacts presented to the sensors by laboratory staff.

[1]https://ijcb2025.ieee-biometrics.org/competitions/

The three tasks in **Part 1** introduce novel (compared to past LivDet-Iris editions) ways of testing algorithms:

- **Task 1: Industry Partner's Tests –** The industry partner, PayEye Poland, Poland, evaluated all submissions using a sequestered dataset that reflects the most prevalent physical attacks observed in real-world iris recognition-based payment services. The presentation attack instruments (PAIs) included in this task are paper printouts, irises displayed on an e-book reader, artificial eyes, doll eyes, mannequin eyes, as well as samples synthesized using Generative Adversarial Networks (GANs).
- **Task 2: Deep Learning-Aided Iris Morphing –** In this task, submissions were tested against morphed iris samples, prepared by compositing two iris images representing two identities, with the seams caused by the compositing process "smoothed" by a diffusion model to increase the visual realism.
- **Task 3: Robustness to Advanced Textured Contact Lens (TCL) Manufacturing –** This task focused on robustness of liveness detection methods against modern manufacturing techniques used to produce TCL, including high-resolution printing, multi-layered designs, and improved pigmentation. Such new techniques make TCLs increasingly indistinguishable from bona fide irises. As many existing liveness detection models are trained on older TCL datasets, this task assessed the community's readiness in detecting new TCL brands and production techniques.

Competitors had the option of participating in one or both parts of the competition, and any (or all) tasks in Part 1 (Algorithms). In this edition of LivDet-Iris, we received 9 submissions from 5 research teams for Tasks 1, 6 submissions from 3 research teams for Task 2, and 10 submissions from 5 research teams for Task 3. This year also marks the first-ever submission to Part 2 (Systems) challenge. The Area Under the Receiver Operating Curve (AUROC) obtained on the test sets was used as a metric to determine the winner. The *Dermalog-Iris* team's submission *001* wins in all three tasks in Part 1, obtaining AUROC of 90.57%, 68.23% and 99.99% in Tasks 1, 2 and 3, respectively. These results demonstrate (a) a good readiness of the community to react to new types of textured contact lenses, (b) relatively good performance in the task, which used unknown and operational-type spoofs, and (c) challenges with detecting visually-appealing morphed iris images, blending two real identities into a composite iris image. Due to only one submission to Part 2, this part does not have a winner, but the evaluation results are presented in this paper.

Baseline models and codes, as well as instructions on how to request a copy of test datasets used in Tasks 2 and 3 are available at `https://github.com/CVRL/livdet-iris-2025` to offer the latest LivDet-Iris benchmark.

2. Previous LivDet-Iris Competitions

Over the years, the LivDet-Iris competitions have evolved significantly in both scope and complexity. The 2013 edition [35] followed a closed-set scenario, where the types of presentation attacks in the test set were the same as those seen during training, although the images themselves were different. In the 2015 edition [36], the focus shifted to detecting contact lenses, introducing samples from various contact brands. The 2017 competition [33] increased the challenge by incorporating sensor variability, using iris images captured with different devices. The 2020 edition [5] introduced an open-set scenario and several novel attack types, including post-mortem, e-book-displayed, artificial, and printed irises.

Most recently, the 2023 competition [28] featured synthetic iris images generated using a StyleGAN model, including samples representing low and high fidelity of synthesized images, as well as a human subject study, in which humans were asked to detect presentation attacks.

3. Experimental Setup

3.1. Submission Protocol

Participants in Part 1 were asked to submit a Python implementation of their algorithm to the organizers. The program was required to accept two input arguments: the path to the input CSV file (specifying test samples) and the path to the output CSV file (containing the algorithm's results on the test data).

The organizers provided a mock-up Python implementation[2], which standardized the interface between competitors and organizers and simplified the evaluations. Each submission was evaluated by the organizer responsible for the specific task, who locally executed the provided code on the task-specific test dataset.

3.2. Competition Datasets

3.2.1 Train Data and Instructional

As in the previous edition of the LivDet-Iris competitions, the organizers did not provide training data. Participants were allowed to use their own datasets or any publicly available datasets to train their best PAD algorithms with a variety of presentation attack (PA) types and bona fide samples of their choice. The organizers provided only a single sample for each task in Part 1 to inform the participants on the image format of samples used in evaluations.

[2]Available in the LivDet-Iris 2025 repository `https://github.com/CVRL/livdet-iris-2025` in "Submission-Instructions" folder

3.2.2 Part 1 Test Data

All test images are provided as 8-bit grayscale PNG files with a resolution of 640×480 pixels. Each pixel is a single grayscale value, where 0 corresponds to black and 255 to white. Images of authentic irises conform to the IMAGE_TYPE_VGA format, as defined in ISO/IEC 19794-6:2011. However, similar conformance is not guaranteed for spoof samples. Below we discuss datasets used in each task in Part 1 of the competition.

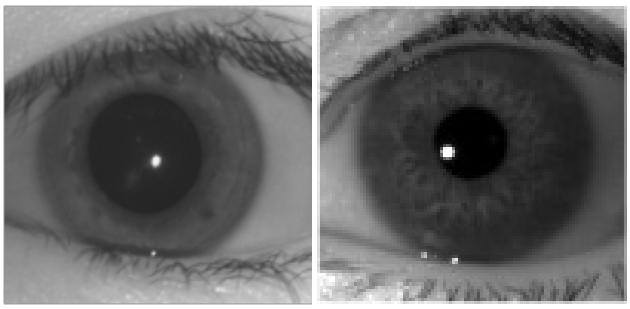
Task 1 Test Data The dataset comprises a total of 37,845 samples, with 21,570 samples ($\sim 56.9\%$) labeled as presentation attack, and 16,275 samples ($\sim 43.0\%$) labeled as bona fide. The dataset includes six types of presentation attacks: synthetic samples (generated using GANs and further enhanced via high-quality restoration techniques in a 50:50 ratio), images displayed on Kindle screens, textured contact lenses, mannequin eyes, artificial plastic eyes, and printed iris images. The bona fide samples exhibit a broad range of natural variation to reflect realistic operational conditions. They were captured under diverse lighting environments, resulting in varying illumination and shadow effects. Some subjects wear eyeglasses, which may introduce reflections partially covering the iris region. Additionally, the presence of makeup, ranging from light to heavy, can, in some cases, obscure parts of the iris texture. However, all bona fide samples are of sufficient quality to be used in iris matching.

Task 2 Test Data The dataset comprises a total of 3,603 samples, with 1,385 ($\sim 38.4\%$) bona fide samples and 2,218 ($\sim 61.6\%$) morphed iris samples. Bona fide samples were selected as all original samples used to produce morph samples. To ensure the generation of novel identities through morphing, we splice iris textures from two distinct subjects by taking the inner iris band (closer to the pupil) from one identity and the outer band (near the limbus) from another one. Then, we evaluate the synthetic sample's identity proximity using HDBIF [3] distance, selecting the sample that minimizes the absolute difference to the two source identities.

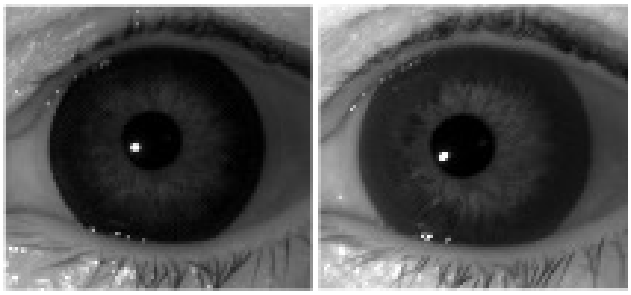
The above splicing process introduces unnatural seams that must be corrected. We trained a diffusion model to "inpaint" regions of iris images located at the seams. Specifically, to train the diffusion model random concentric band-shaped regions are replaced with noise and the network is trained to replace these regions following a diffusion "inpainting" framework similar to that proposed in Palette by Saharia et al. [25]. Unlike traditional diffusion pipelines (and Palette) that only utilize the Mean Squared Error (MSE) loss, our loss function combines a pixel-wise MSE term with perceptual similarity components (Learned Perceptual Image Patch Similarity – LPIPS, and Multi-Scale Structural Similarity – MS-SSIM) to improve high-frequency texture fidelity. Finally, to eliminate boundary artifacts from splicing, we replace the seam with a band of noise and "inpaint" it using our trained diffusion model, creating a perceptually coherent iris image of a "mixed" identity.

Task 3 Test Data The dataset comprises 564 bona fide iris images and 788 samples captured from subjects wearing textured contact lenses. The presentation attack samples represent nine different contact lens, and are categorized into two quality classes: High Quality and Pixelated. 644 high-quality lenses originate from six different contact lens, while 144 pixelated samples originate from three other contact lens. For algorithm testing, we used at least 500 bona fide and 500 presentation attack images to ensure balanced evaluation.

Fig.1 shows sample images from Task 2 and Task 3 that were used by the organizers to evaluate the submitted algorithms. Task 1 samples, originating from a commercial co-organizer, are not being released.



(a) Task 2: Morphed sample (b) Task 3: Bona fide



(c) Task 3: Pixelated (d) Task 3: High quality

Figure 1: Example of cropped images from Task 2 (iris morphing) and Task 3 (bona fide, pixelated and high quality textured contact lens) used in the evaluation of the submitted algorithms.

3.2.3 Part 2 Test Data

In **Part 2**, system-level PAD performance is evaluated through laboratory-conducted spoofing attempts, with vendors indicating whether they support PAD and optionally providing liveness scores. The evaluation includes 500 bona fide and 500 spoof attempts, with bona fide samples acquired according to NIST IREX-V [24] and ISO/IEC 19794-6 [13] guidelines. Spoof attempts are designed to trigger iris image capture using 166 printed iris images, 166 iris images displayed on a Kindle device, and 168 attempts involving a live person wearing various types of textured contact lenses. Non-responses to spoof attempts are considered correct rejections, whereas failures to capture bona fide iris samples are treated as failures to acquire and are factored into the final system ranking.

3.3. Performance Evaluation Metrics

In **Part 1**, algorithm performance is measured using the Attack Presentation Classification Error Rate **(APCER)** and the Bonafide Presentation Classification Error Rate **(BPCER)**, as recommended by ISO/IEC 30107-1:2017 [14]. Both APCER and BPCER are calculated at a fixed acceptance threshold of 0.5 to evaluate the generalization capability of the implementation without threshold tuning. Additionally, the Area Under the Receiver Operating Characteristic curve **(AUROC)** is computed using APCER and $1 -$ BPCER scores. The closer the AUROC is to 1.0, the better the algorithm.

In **Part 2**, system performance is evaluated using APCER and BPCER. Both metrics are calculated based on a fixed acceptance threshold provided by the team that submitted the system. A lower combined value of APCER and BPCER indicates better overall performance. All samples that are not processed (crashed, skipped, null response) contribute towards a Non-Response Rate (NRR):

$$\text{NRR} = \frac{\text{Total Number of Non-Responses}}{\text{Total Number of Samples Evaluated}}$$

3.4. Winner Selection

Since the three tasks in **Part 1 (Algorithms)** are distinct, we allowed each task to have its own winner based on the highest AUROC score in a given task.

The winner of **Part 2 (Systems)** will be the one with the lowest combined BPCER and APCER.

4. Submitted Algorithms

4.1. Team: BUCEA

Method The BUCEA team submitted two algorithms: AMF-IPAD for both Task 1 and Task 3, and SSDG exclusively for Task 3. AMF-IPAD is an attention-assisted multilevel fusion framework for iris liveness detection. It uses an iris-mask-guided attention module to separately extract ocular features for global attacks and iris features for local attacks. The model integrates complementary information at the image, feature, and score levels using a multilevel fusion strategy. Input images are formed by concatenating CLAHE-enhanced, HOG-extracted, and raw grayscale versions into a three-channel input, which is processed through a shared ResNet18 backbone split into ocular and iris branches. Both branches share architecture and weights, but focus on different regions, with the iris branch capturing finegrained features crucial for liveness detection. A multihead attention module adaptively fuses features from both branches. The model was trained using data augmentation, Adam optimization, and a weighted joint loss combining pixel-level supervision and self-distillation. Final predictions were made by aggregating weighted outputs from the ocular, iris, and fused branches. Single-Side Domain Generalization (SSDG) framework was initially developed as an end-to-end approach for face anti-spoofing [16] and has been adapted to enhance the cross-domain generalization ability of textured contact lens detection.

Training Data For Task 1, the dataset used for training and validation comprises 132,944 images (56,436 bonafide and 76,508 attack samples), and combines open-sourced datasets, such as: LivDet-Iris 2017, NDCLD'15, IIITD Contact Lens Iris, ND CrossSensor-Iris-2013, and BUCEA team's self-collected print dataset and self-generated synthetic images based on StyleGAN2 and StyleGAN3. The dataset contains many types of attacks, such as contact lenses, paper printouts, synthetic samples, and doll eyes. For Task 3, the dataset used for training and validation consists of 84,876 images (47,495 bonafide and 37,381 attack samples), and combines open-sourced datasets, such as: LivDet-Iris 2017, NDCLD'13, IIITD Contact Lens Iris, IFVEAI1000, CASIAH100. The dataset contains various textured contact lens patterns.

4.2. Team: Dermalog-Iris

Method The proposed algorithm employs three separate neural networks to detect paper printouts, patterned contact lenses, and morphing attacks. Each network was trained independently using overlapping image patches. While many existing approaches use a single classifier to detect multiple types of presentation attacks, this method combines the outputs of three independently trained classifiers. A key aspect of this approach is the decomposition of iris images into overlapping patches, which helps mitigate the limitations posed by small training datasets. For each image, the patch-level predictions were aggregated to produce a final classification. These results from the three networks were then fused to compute a final liveness score. The contact lens detection network takes nine patches derived from the normalized (rubbersheet) iris image, whereas the print-out

and morph detection networks each process nine patches extracted from the full eye image.

Training Data The dataset used for training and validation comprises both publicly available datasets and proprietary Dermalog datasets, which include paper printouts, patterned contact lenses, and morphed iris images. Specifically, the LivDet-Iris 2023 competition database, along with its complementary images, was employed. In addition, three internal Dermalog databases containing diverse iris image samples were utilized to enhance variability and robustness. In total, approximately 130,000 near-infrared iris images were used for training purposes.

4.3. Team: EyeFortress (EF)

Method The EF developed four PAD algorithms based on advanced image classification architectures. The first model used ResNet50 backbone with a simple two-class linear classifier. The second model leveraged the DINOv2 framework with a ViT-14 backbone and a non-linear classifier. The third model used a learned ensemble of four pre-trained networks (ResNet50, DenseNet121, DINO-ResNet50, and DINOv2-ViT14), feeding normalized spoof probabilities into a lightweight MLP for final prediction. The fourth model enhanced the latter one with a dynamic ensemble using attention to weight each network's output. For all models, the images were augmented with random flips, $\pm 10^{\circ}$ rotations, and affine shifts, then resized and normalized using ImageNet statistics. The models were fine-tuned from ImageNet-1K weights for 50 epochs, with softmax score used to predict the liveness.

Training Data The data set used for training and validation comprised samples from Clarkson databases from 2013, 2015, 2017, and 2023, as well as the NDCLD databases from 2013, 2015, and 2017, and the IIIT_WVU database, as well as a high-quality dataset collected by the EF team, consisting of 336 bona fide iris images and 1,072 E-display iris images captured using mobile cameras.

4.4. Team: HDA

Method The HDA team developed an algorithm based on LoRA fine-tuning of a CLIP base model from [27]. On top of the CLIP model, a single neuron layer was also optimized for binary classification. The input images were transformed according to CLIP transformations of RGB images (in this case, the grayscale NIR image was repeated into each RGB channel) and resized to 224×224 pixels. The model was optimized using LoRA for 50 epochs and the results were computed with the model that obtained the best validation results (after 38 epochs).

Training Data The dataset used for training and validation comprises five PAs such as: paper printouts, patterned contact lenses, cadaver irises, synthetic irises, and morphed samples, plus bona fide iris samples. Some of these images are from LivDet-Iris 2020 benchmark. In total, 70,000 images were used, evenly distributed across classes.

4.5. Team: MSU

Method The MSU team submitted two algorithms, D-NetPAD [26] and SPAD [23], both of which use DenseNet-121 [11] as the backbone. In both approaches, segmented iris images are resized to 224×224 pixels before being fed into the network. For D-NetPAD, the iris region was extracted using the VeriEye detector[3] during training, while the Iris-SAM segmenter [7] was used for the evaluation tasks conducted by the organizers. Iris-SAM was guided by the bounding boxes generated by YOLOv8[4], finetuned using the Clarkson dataset [34]. D-NetPAD model was trained using stochastic gradient descent (SGD) with a momentum of 0.9, a learning rate of 0.005, a batch size of 20, and cross-entropy loss over 50 epochs. SPAD followed a similar preprocessing pipeline, using Iris-SAM for segmentation, followed by cropping and resizing. It was adapted for binary classification by replacing the final DenseNet layer with a sigmoid output. SPAD was fine-tuned using the Adam optimizer with a learning rate of 0.0001 and binary cross-entropy loss for 50 epochs. The dataset was split 80:20 for training and validation, and the best-performing model was selected based on the lowest validation loss.

Training Data D-NetPAD was trained on a proprietary dataset comprising 13,851 iris images [26], including 9,660 bona fide samples and 4,291 presentation attacks such as printed images, artificial eyes, textured contact lenses, Kindle-displayed samples, and transparent domes placed over printed eyes. SPAD was trained on the publicly available LivDet-Iris 2017 (excluding the unknown test subset) [34] and LivDet-Iris 2020 [4] datasets, using both training and test partitions.

5. Baseline Algorithms

5.1. Architectures

To train baseline algorithms, we selected ResNet101 [10], DenseNet121 [11], and Visual Transformer (ViT-B/16) [6] architectures, as they demonstrated the highest performance in the LivDet-Iris 2023 competition [29].

5.2. Training

The baselines were trained under two different scenarios: (1) using only authentic bona fide and attack samples, along with inpainting-based synthetic images, and (2) additionally incorporating synthetically generated attack samples produced using StyleGAN2-ADA [18] and the Improved Denoising Diffusion Probabilistic Model (IDDPM) [22]. Fig.

[3] https://www.neurotechnology.com/verieye.html
[4] https://github.com/ultralytics/ultralytics

2 presents example images that were used for training the baseline models.

Table 1 summarizes the datasets used to train the baseline models. We allocated 80% of each dataset for training and the remaining 20% for validation during model development. Each model was trained for 50 epochs with a batch size of 32. Cross-entropy loss and the SGD optimizer were applied consistently across all models. The training was conducted using cropped iris images, with a 16-pixel padding around the iris region. We applied a range of standard augmentation techniques adopted from [17], including affine transformations, random horizontal flipping, sharpening, blurring, contrast and brightness adjustments, and the addition of Laplace-distributed noise.

Image Type	Contributing Dataset	# of Samples	Total # of Samples
Bona fide	ATVS-FIr [8]	800	405,167
	BERC_IRIS_FAKE [21]	2,776	
	CASIA-Iris-Thousand [12]	19,952	
	CASIA-Iris-Twins [12]	3,181	
	Disease-Iris v2.1 [30] [30]	1,438	
	IIITD Combined Spoofing Database [20]	4,531	
	IIITD Contact Lens Iris [19]	13	
	LivDet-Iris Clarkson 2015 [36]	813	
	LivDet-Iris IIITD-WVU 2017 [33]	2,944	
	LivDet-Iris Warsaw 2017 [33]	5,167	
	LivDet-Iris Warsaw 2015 [36]	36	
	LivDet-Iris Clarkson 2017 [33]	3,949	
	LivDet-Iris 2020 [4]	5,331	
	Notre Dame (internal)	354,236	
Artificial	BERC_IRIS_FAKE [21]	80	803
	LivDet-Iris 2020 [4]	526	
	Notre Dame (internal)	197	
Textured contact lenses	BERC_IRIS_FAKE [21]	140	31,708
	IIITD Contact Lens Iris [19]	3,420	
	LivDet-Iris Clarkson 2015 [36]	1,107	
	LivDet-Iris Clarkson 2017 [33]	1,881	
	LivDet-Iris IIITD-WVU 2017 [33]	1,700	
	Notre Dame (internal)	19,124	
Post-mortem	Post-Mortem-Iris v3.0 [31]	4,866	10,636
	NIJ [2]	5,770	
Paper printouts	ATVS-FIr [8]	800	19,593
	BERC_IRIS_FAKE [21]	1,600	
	IIITD Combined Spoofing Database [20]	1,371	
	LivDet-Iris Clarkson 2015 [36]	1,745	
	LivDet-Iris IIITD-WVU 2017 [33]	1,766	
	LivDet-Iris Warsaw 2017 [33]	6,841	
	LivDet-Iris Warsaw 2015 [36]	20	
	LivDet-Iris Clarkson 2017 [33]	2,250	
	LivDet-Iris 2020 [4]	1,049	
	Notre Dame (internal)	2,151	
Displayed on e-ink device	LivDet-Iris 2020 [4]	81	2,147
	Notre Dame	2,066	
Synthetic	CASIA-Iris-Syn V4 [32]	10,000	10,000
Synthetic GenAI	IDDPM (Notre Dame internal)	9,638	19,643
	StyleGAN (Notre Dame internal)	10,005	
All combined			**471,677 (491,320)***

*including samples generated by StyleGAN and IDDPM

Table 1: Numbers and sources of iris images used to train baseline models.

6. Results and Competition Winners

Figure 3 shows the ROC curves for Part 1, Tasks 1–3, highlighting the best-performing algorithms for both the baseline and the submitted algorithms. Detailed results for these tasks are provided in Tables 2, 3, and 5, respectively. The results for Part 2 are summarized in Table 4.

6.1. Part 1 (Algorithm)

The **Dermalog-Iris** team delivered the top-performing algorithm across all three tasks in Part 1 of LivDet-Iris 2025. In **Task 1** (Industry Partner's Tests), their algorithm reached an AUROC of **0.9057**, in **Task 2** (Deep Learning-Aided Iris Morphing), an AUROC of **0.6823**, and in **Task 3** (Advanced Textured Contact Lens), an AUROC of **0.9999**. Below we offer more detailed comments related to the observed performance in all three tasks.

Task 1 (Industry Partner's Tests) Although the Dermalog-Iris team achieved the highest AUROC for this task, the MSU team's D-NetPAD algorithm demonstrated the lowest (best) APCER of 0.0644 at a threshold value of 0.5 among all submissions. Meanwhile, the HDA team achieved the lowest (best) BPCER of 0.0474 among all submissions. These results suggest that MSU's D-NetPAD performed particularly well against the attack types introduced by the industry, while the HDA team's algorithm demonstrated higher accuracy in identifying only bona fide irises (see Table 2). As illustrated in Figure 3, several algorithms still struggled with a task involving bona fide iris samples and attacks collected in a real-world setup including diverse and unknown acquisition conditions.

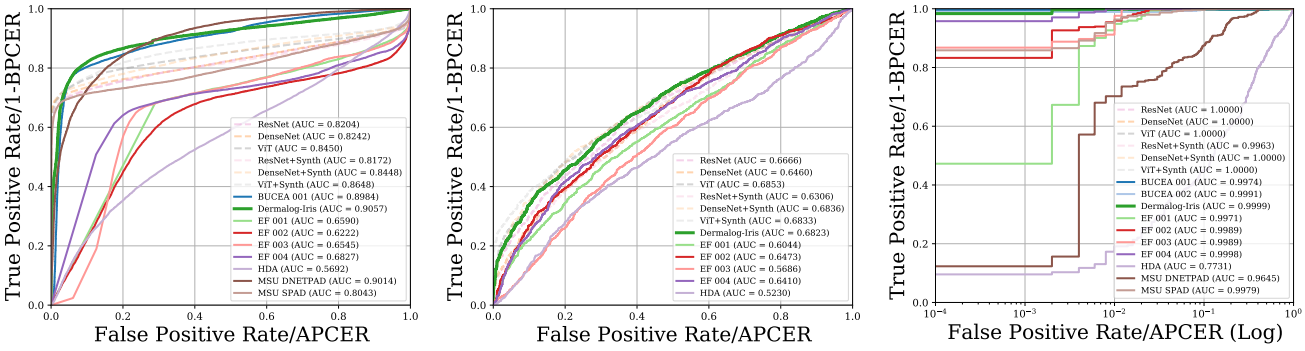
Team	Algorithm	AUROC ↑	APCER ↓ @0.5 thr.	BPCER ↓ @0.5 thr.
Baseline	ResNet	0.8242	0.3003	0.0224
	DenseNet	0.8204	0.3082	0.0344
	ViT	0.8450	**0.2894**	0.0245
Baseline + *synthetic training samples*	ResNet	0.8448	0.2910	**0.0214**
	DenseNet	0.8172	0.3066	0.0305
	ViT	**0.8648**	0.2947	0.0262
BUCEA	001	0.8984	0.1129	0.3294
Dermalog-Iris	001	**0.9057**	0.1069	0.2826
EyeFortress (EF)	001	0.6590	0.1519	0.8490
	002	0.6222	0.1887	0.8928
	003	0.6545	0.1202	0.9332
	004	0.6827	0.1662	0.8706
hda_team	hda_teamV1	0.5692	0.8829	**0.0474**
MSU	D-NetPAD	0.9014	**0.0644**	0.4033
	SPAD	0.8043	0.2538	0.2762

Table 2: Results for Part 1, Task 1 (Industry Partner's Tests). For each metric, the best-scoring competitor algorithm and best-scoring baseline algorithm are **bolded**. The algorithm Dermalog-Iris 001 wins Task 1 for achieving the greatest AUROC among competitors.

Task 2 (Deep Learning-Aided Iris Morphing) Only three of the five teams chose to participate in this task. While the Dermalog-Iris team achieved the highest AUROC and was declared the winner, the HDA team's algorithm

(a) Bona fide (b) Printout (c) Artificial (d) Kindle (e) Post-mortem (f) TCL (g) Synthetic

Figure 2: Example of a bona fide iris image (a) and all presentation attacks used in the training of baseline models (b-g).



(a) Task 1 (Industry Partner's Tests) (b) Task 2 (Iris Morphing) (c) Task 3 (Advanced Contact Lenses)

Figure 3: ROC curves for Part 1, Tasks 1-3. For each task, the winning algorithm (Dermalog-Iris) is plotted with a bold line. All baseline algorithms are plotted with transparent dashed lines. For Task 3, APCER is plotted at log-scale to better illustrate differences between algorithms.

recorded the lowest (best) APCER of 0.4292 at a threshold of 0.5, and Dermalog-Iris achieved the lowest (best) BPCER at 0.0051 (see Table 3). The lower AUROC values as shown in Figure 3 show that submitted PAD algorithms had difficulty in accurately identifying morphed samples as presentation attacks, what highlighting a significant challenge posed by this type of spoofing.

Team	Algorithm	AUROC ↑	APCER ↓ @0.5 thr.	BPCER ↓ @0.5 thr.
Baseline	ResNet	0.6666	0.9946	0.0000
	DenseNet	0.6460	0.9950	0.0000
	ViT	**0.6853**	0.9775	0.0000
Baseline	ResNet	0.6306	0.9959	0.0000
+ synthetic	DenseNet	0.6836	0.9959	0.0000
training samples	ViT	0.6833	**0.9648**	0.0000
Dermalog-Iris	001	**0.6823**	0.9058	**0.0051**
EyeFortress (EF)	001	0.6044	0.7759	0.1040
	002	0.6473	0.7606	0.0866
	003	0.5686	0.9125	0.0729
	004	0.6410	0.8300	0.0556
hda_team	hda_teamV1	0.5230	**0.4292**	0.5227

Table 3: Results for Part 1, Task 2 (Deep Learning-Aided Iris Morphing). For each metric, the best-scoring competitor algorithm and best-scoring baseline algorithm are **bolded**. The algorithm Dermalog-Iris 001 wins Task 2 for achieving the greatest AUROC among competitors.

Task 3 (Robustness to Advanced Textured Contact Lens Manufacturing) In this task, again the Dermalog-Iris algorithm obtained the lowest AUROC and was declared the winner. With the exception of the HDA team (0.7731) and MSU's D-NetPAD algorithm (0.9645), all submissions achieved an AUROC above 0.99. The highest (worst) APCER was detected for HDA algorithm and the highest (worst) BPCER was found for the EF team's second algorithm (see Table 5). Despite advances in technology and texture realism by contact lens manufacturers, these results indicate that textured contact lenses are an easily detectable presentation attack by modern iris PAD algorithms. Seeing the results in Task 3, we may conclude that detecting textured contact lenses, even those manufactured recently, appears to be a solved problem in iris PAD.

6.2. Part 2 (Systems)

Dermalog-Iris was the only team that participated in Part 2 (Systems). Here, they achieved an APCER of 0.0000 for both printed and kindle irises, and an APCER of 0.1310 for irises with TCL with a total BPCER of 0.0000. Combined APCER + BPCER of 0.1310 at the acceptance threshold suggested by the manufacturer (80). Table 4 presents the results of Task 3.

For attack samples, approximately 66% could not be processed by the device and were automatically considered attacks. In contrast, the NRR for bona fide samples was 0.00,

APCER↓ @80 thr. (Printed Image)	APCER↓ @80 thr. (Kindle Display)	APCER↓ @80 thr. (Textured Lens)	BPCER↓ @80 thr.	APCER + BPCER↓ @80 thr.	NRR attacks only	NRR bona fide only
0.0000	0.0000	0.1310	0.0000	0.1310	0.6640	0.0000

Table 4: Results for Part 2 (Systems) for the only system submitted to this part (Dermalog-Iris).

Team	Algorithm	AUROC ↑	APCER ↓ @0.5 thr.	BPCER ↓ @0.5 thr.
Baseline	ResNet	**1.000**	0.0110	0.0000
	DenseNet	**1.000**	0.0165	0.0000
	ViT	**1.000**	0.0147	0.0000
Baseline + synthetic training samples	ResNet	0.9963	0.0147	0.0000
	DenseNet	0.9999	**0.0000**	0.0000
	ViT	0.9999	**0.0000**	0.0000
BUCEA	001	0.9974	0.0110	**0.0000**
	002	0.9991	0.1121	**0.0000**
Dermalog-Iris	001	**0.9999**	0.0110	0.0020
EyeFortress (EF)	001	0.9971	**0.0000**	0.0620
	002	0.9989	**0.0000**	0.1040
	003	0.9989	**0.0000**	0.0400
	004	0.9998	**0.0000**	0.0140
hda_team	hda_teamV1	0.7731	0.5551	0.0920
MSU	D-NetPAD	0.9645	0.1618	0.0760
	SPAD	0.9979	0.1599	**0.0000**

Table 5: Results for Part 1, Task 3 (Robustness of PAD to Advanced Manufacturing Methods of Textured Contact Lens Patterns). For each metric, the best-scoring competitor algorithm and best-scoring baseline algorithm are **bolded**. The algorithm Dermalog-Iris 001 wins Task 3 for achieving the greatest AUROC among competitors.

indicating that all bona fide samples were successfully processed.

7. Conclusions and Discussion

This edition of the LivDet-Iris competition received submissions from five teams. Out of these, one algorithm chose not to participate in Task 1 of Part 1, which was evaluated by the industry partner PayEye Poland and two teams (4 algorithms) chose not to participate in Task 2 of Part 1, which focused on iris morphing. The Dermalog-Iris team was identified as the winner of Part 1, achieving the highest AUROC across all three tasks. However, relying only on AUROC as a metric in biometric tasks—such as PAD—has certain limitations [15].

Competitors' performance across the three tasks of Part 1 suggest strengths and weaknesses of modern iris PAD algorithms. Task 1 results indicate that submitted PAD models still struggle under real-world iris PA conditions. The samples, collected by the industry partner PayEye Poland, reflect realistic operational environments, with natural variations in lighting, shadows, eyeglasses, and makeup that can partially obscure the iris region, making classification and PAD tasks more challenging. The morphed attacks of Task 2 were consistently difficult, and high APCER values show that most algorithms failed to detect morphed iris images. The low performance indicates weakness in recognizing digitally-presented iris attacks, which may be attributed to both the novelty of the PA as well as traditional training data bias towards physically-presented PAs. Conversely, Task 3 results suggest that PAD models are effective at detecting textured contact lenses, despite advanced manufacturing techniques mimicking human iris texture. Additionally, comparing baselines results for Part 1 with the results of submitted algorithms indicated that some of these algorithms outperformed the baselines. Also, adding synthetically generated samples using StyleGAN and diffusion models did not appear to improve baseline performance (see ROC curves shown in Figure 3). These observations suggest that, when dealing with unseen attacks or real-world iris presentation attack samples, PAD algorithms struggle to accurately label samples, regardless of their type or the amount of training data used. Dermalog-Iris submitted the only algorithm for Part 2, Systems, which involved testing models using physical artifacts presented to the sensors, demonstrated strong performance on both bona fide and presentation attack samples.

LivDet-Iris 2025, in addition to the results compiled in this paper, open-sources all baseline algorithms, and releases Task 2 and Task 3 test datasets to serve as the most recent benchmark in iris PAD evaluation.

Note about using AUROC as a comparison metric AUROC measures the ability of the system to distinguish between bonafides and PAs across all thresholds. But PAD systems, depending on the setup, may operate at either low false detection rates (yielding low APCER) due to strict security requirements, or low false alarm rate (yielding low BPCER) due to the goal of not increasing the system's overall false rejection rate. AUROC does not emphasize performance in these low-APCER or low-BPCER regions, potentially obscuring vulnerabilities to sophisticated attacks or method's over-sensitiveness to potentially anomalous features [15]. Furthermore, AUROC does not reflect the actual distribution of attack types or their varying levels of difficulty, treating all errors equally regardless of the real-

world impact of different attacks. In highly imbalanced PAD datasets, where attack samples may far outnumber genuine ones or vice versa, AUROC can give a misleading impression of overall performance. This concern has been expressed in other fields as well [9, 1].

However, while AUROC does not necessarily articulate performance at extreme operating points, high AUROC scores generally indicate implementation adaptability. Further, our tasks explore relatively balanced class distributions, avoiding pitfalls associated with class imbalance. Considering the pros and cons of AUROC in assessing biometric PAD performance, as well as the metric's familiarity in the biometric community, we opted to use it as our winner-selection metric.

Acknowledgements

This material is based upon work supported by the U.S. National Science Foundation under grants No. 2237880 and 1650503. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation. Caiyong Wang was funded by the Beijing Natural Science Foundation (4242018).

References

[1] E. Bahn and M. Alber. On the limitations of the area under the roc curve for ntcp modelling. *Radiotherapy and Oncology*, 144:148–151, 2020.

[2] A. Czajka, D. J. Chute, A. Ross, P. J. Flynn, and K. W. Bowyer. Software tool and methodology for enhancement of unidentified decedent systems with post-mortem automatic iris recognition. In *New York, 2019–2021. Inter-university Consortium for Political and Social Research [distributor]*. 2023.

[3] A. Czajka, D. Moreira, K. Bowyer, and P. Flynn. Domain-specific human-inspired binarized statistical image features for iris recognition. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 959–967. IEEE, 2019.

[4] P. Das, J. McFiratht, Z. Fang, A. Boyd, G. Jang, A. Mohammadi, S. Purnapatra, D. Yambay, S. Marcel, M. Trokielewicz, et al. Iris liveness detection competition (livdet-iris)-the 2020 edition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.

[5] P. Das, J. McGrath, Z. Fang, A. Boyd, G. Jang, A. Mohammadi, S. Purnapatra, D. Yambay, S. Marcel, M. Trokielewicz, P. Maciejewicz, K. Bowyer, A. Czajka, S. Schuckers, J. Tapia, S. Gonzalez, M. Fang, N. Damer, F. Boutros, A. Kuijper, R. Sharma, C. Chen, and A. Ross. Iris Liveness Detection Competition (LivDet-Iris) - The 2020 Edition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2020.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] P. Farmanifard and A. Ross. Iris-sam: Iris segmentation using a foundation model. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 394–409. Springer, 2024.

[8] J. Galbally, J. Ortiz-Lopez, J. Fierrez, and J. Ortega-Garcia. Iris liveness detection based on quality related features. In *5th IAPR Int. Conf. on Biometrics (ICB)*, pages 271–276, New Delhi, India, March 2012.

[9] S. Halligan, D. G. Altman, and S. Mallett. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European Radiology*, 25:932–939, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning. *Image Recognition*, 7, 2015.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[12] Institute of Automation, Chinese Academy of Sciences. CASIA Biometrics Database Collections. http://biometrics.idealtest.org. Accessed: June 09, 2023.

[13] International Organization for Standardization. Information technology — Biometric data interchange formats — Part 6: Iris image data. ISO/IEC 19794-6:2011, 2011. `https://www.iso.org/standard/50869.html`.

[14] ISO/IEC 30107-3. Information technology – Biometric presentation attack detection – Part 3: Testing and reporting, 2016.

[15] A. K. Jain, B. Klare, and A. Ross. Guidelines for best practices in biometrics research. In *Proceedings of the 8th IAPR International Conference on Biometrics (ICB)*, pages 541–545, Phuket, Thailand, May 2015.

[16] Y. Jia, J. Zhang, S. Shan, and X. Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8484–8493, 2020.

[17] A. B. Jung, K. Wada, J. Crall, S. Tanaka, J. Graving, C. Reinders, S. Yadav, J. Banerjee, G. Vecsei, A. Kraft, Z. Rui, J. Borovec, C. Vallentin, S. Zhydenko, K. Pfeiffer, B. Cook, I. Fernández, F.-M. De Rainville, C.-H. Weng, A. Ayala-Acevedo, R. Meudec, M. Laporte, et al. imgaug. `https://github.com/aleju/imgaug`, 2020. Online; accessed 01-Feb-2020.

[18] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[19] N. Kohli, D. Yadav, M. Vatsa, and R. Singh. Revisiting iris recognition with color cosmetic contact lenses. In *IEEE Int. Conf. on Biometrics (ICB)*, pages 1–7, Madrid, Spain, June 2013. IEEE.

[20] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore. Detecting medley of iris spoofing attacks using desist. In *IEEE Int. Conf. on Biometrics: Theory Applications and Systems*

(BTAS), pages 1–6, Niagara Falls, NY, USA, Sept 2016. IEEE.

[21] S. J. Lee, K. R. Park, Y. J. Lee, K. Bae, and J. H. Kim. Multifeature-based fake iris detection method. *Optical Engineering*, 46(12):1 – 10, 2007.

[22] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[23] D. Pal, R. Sony, and A. Ross. A parametric approach to adversarial augmentation for cross-domain iris presentation attack detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5719–5729, 2025.

[24] G. W. Quinn, J. Matey, E. Tabassi, and P. Grother. IREX V: Guidance for iris image collection. Technical Report NISTIR 8013, National Institute of Standards and Technology, 2014.

[25] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, pages 1–10, 2022.

[26] R. Sharma and A. Ross. D-NetPAD: An explainable and interpretable iris presentation attack detector. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020.

[27] J. E. Tapia, L. J. González-Soler, and C. Busch. Towards iris presentation attack detection with foundation models. *arXiv preprint arXiv:2501.06312*, 2025.

[28] P. Tinsley, S. Purnapatra, M. Mitcheff, A. Boyd, C. Crum, K. Bowyer, P. Flynn, S. Schuckers, A. Czajka, M. Fang, N. Damer, X. Liu, C. Wang, X. Sun, Z. Chang, X. Li, G. Zhao, J. Tapia, C. Busch, C. Aravena, and D. Schulz. Iris Liveness Detection Competition (LivDet-Iris) – The 2023 Edition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2023.

[29] P. Tinsley, S. Purnapatra, M. Mitcheff, A. Boyd, C. Crum, K. Bowyer, P. Flynn, S. Schuckers, A. Czajka, M. Fang, et al. Iris liveness detection competition (livdet-iris)–the 2023 edition. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[30] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Assessment of iris recognition reliability for eyes affected by ocular pathologies. In *IEEE Int. Conf. on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–6, 2015.

[31] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Post-mortem iris recognition with deep-learning-based image segmentation. *Image and Vision Computing*, 94:103866, 2020.

[32] Z. Wei, T. Tan, and Z. Sun. Synthesis of large realistic iris databases using patch-based sampling. In *Int. Conf. on Pattern Recognition (ICPR)*, pages 1–4, Tampa, FL, USA, Dec 2008. IEEE.

[33] D. Yambay, B. Becker, N. Kohli, D. Yadav, A. Czajka, K. W. Bowyer, S. Schuckers, R. Singh, M. Vatsa, A. Noore, D. Gragnaniello, C. Sansone, L. Verdoliva, L. He, Y. Ru, H. Li, N. Liu, Z. Sun, and T. Tan. LivDet Iris 2017 – iris liveness detection competition 2017. In *IEEE Int. Joint Conf. on Biometrics (IJCB)*, pages 1–6, Denver, CO, USA, 2017. IEEE.

[34] D. Yambay, B. Becker, N. Kohli, D. Yadav, A. Czajka, K. W. Bowyer, S. Schuckers, R. Singh, M. Vatsa, A. Noore, D. Gragnaniello, C. Sansone, L. Verdoliva, L. He, Y. Ru, H. Li, N. Liu, Z. Sun, and T. Tan. LivDet-Iris 2017 – Iris Liveness Detection Competition 2017. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 733–741, 2017.

[35] D. Yambay, J. S. Doyle, K. W. Bowyer, A. Czajka, and S. Schuckers. Livdet-iris 2013 - iris liveness detection competition 2013. In *IEEE Int. Joint Conf. on Biometrics (IJCB)*, pages 1–8, Clearwater, FL, USA, Sept 2014. IEEE.

[36] D. Yambay, B. Walczak, S. Schuckers, and A. Czajka. Livdet-iris 2015 - iris liveness detection competition 2015. In *IEEE Int. Conf. on Identity, Security and Behavior Analysis (ISBA)*, pages 1–6, New Delhi, India, Feb 2017.