A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2019,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 1

Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations

Pedro M. Ferreira[1,2], Ana F. Sequeira [1], Diogo Pernes[1,3], Ana Rebelo[1,4], Jaime S. Cardoso[1,2]

Abstract: Despite the high performance of current presentation attack detection (PAD) methods, the robustness to unseen attacks is still an under addressed challenge. This work approaches the problem by enforcing the learning of the bona fide presentations while making the model less dependent on the presentation attack instrument species (PAIS). The proposed model comprises an *encoder*, mapping from input features to latent representations, and two classifiers operating on these underlying representations: (i) the *task-classifier*, for predicting the class labels (as bona fide or attack); and (ii) the *species-classifier*, for predicting the PAIS. In the learning stage, the *encoder* is trained to help the *task-classifier* while trying to fool the *species-classifier*. Plus, an additional training objective enforcing the similarity of the latent distributions of different species is added leading to a 'PAI-species'-independent model. The experimental results demonstrated that the proposed regularisation strategies equipped the neural network with increased PAD robustness. The adversarial model obtained better loss and accuracy as well as improved error rates in the detection of attack and bona fide presentations.

Keywords: Iris presentation attack detection, open-set, adversarial learning, transfer learning.

1 Introduction

Biometric recognition systems are considered reliable enough to be deployed in government and civilian applications. The shift from controlled samples acquisition to a more autonomous one increased the vulnerabilities of these systems. Unfortunately, presentation attack detection (PAD) measures had not grown robustly along with this quick evolution and several weak points can be exploited when performing unsupervised biometric identification as such in mobile biometrics, for example. Successful spoofing attempts have been made public in a matter of days, or even hours, after the release of high-tech devices equipped with biometric recognition. The iris recognition sensor of Samsung S8 was reportedly spoofed by German researchers by simply printing a photo of the authorised user and placing a contact lens in it [Ch17]. More recently, the quick hack of Samsung Galaxy S10 ultrasonic fingerprint sensor suggests no presentation attack detection measures of any kind. It is fair to conclude that industry does not share the same enthusiasm as academic community on anti-spoofing measures denoted by the good amount of research continuously produced [RB15, CB18, GFC19, Sc19].

The first two authors contributed equally to this work.

[1] INESC TEC, Porto, Portugal, pmmf@inesctec.pt
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
[3] Faculdade de Ciencias da Universidade do Porto, Porto, Portugal
[4] Universidade Portucalense, Porto, Portugal

Fortunately, exceptions are starting to show in commercial products, like the recent case of the Apple iPhone 'Face ID" case [5] or the FaceTec ZoOm technology [Fa19]. Undoubtedly this change is motivated and supported by initiatives that encourage the development and 'open testing' of spoofing coutermeasures such as 'The National Voluntary Laboratory Accreditation Program" (NVLAP) from NIST [6].

Nevertheless, research-wise there are still open problems to address. This work focus on the fact that most PAD techniques are based on falsely optimistic evaluation methodologies [Se16]: traditionally, the classification models are designed and then evaluated using datasets comprising bona fide presentations and a specific species of presentation attack instruments (PAI). The case when a PAI in the test set is significantly different from the ones used for training is overlooked. What if such sample has a higher probability to circumvent the system than the ones drawn from the original training dataset? To solve this research question it is necessary to develop robust methods to cope with sophisticated and unseen attacks as our eventual intruders become more capable and successfully develop new spoofing techniques.

The aforementioned problem has in fact been addressed before regarding iris, fingerprint and face (often targeted under the open-set or anomaly detection contexts). However, it still remains a challenging topic. Despite the importance of iris as a biometric trait for recognition purposes, in our view, the study of iris PAD generalization problem to unseen PAI species (PAIS) has not been yet fully studied in literature.

In this work, it is proposed an artificial neural network (ANN) along with an adversarial training (AT) objective which are specifically designed to improve the robustness and generalisation capacity of the PAD method to new presentation attacks. The idea behind this approach is to enforce the model to learn latent representations that preserve the liveness properties of the bona fide presentations while discarding the PAIS variability between different types of species. The use of invariant representations that capture the PAIS invariance and disregard the intra-PAIS variance will force the PAIS representations to be 'closer' and 'further away' from the bona fide representations. Thus, the influence of the 'PAI-species'-specific aspects that may hamper the PAD classification task will be minored and the model will be more robust to differentiate an attack presentations, even new unseen ones, from the bona fide presentations.

The proposed adversarial training objective combines representation learning and artificial neural networks and is specifically designed to address the generalisation capacity to an unseen attack problem. In an innovative approach, the proposed model jointly learns the representation and the classifier from the data, while explicitly imposing 'PAI-species'-invariance in the high-level representations for a robust presentation attack classification.

[5] www.biometricupdate.com/201812/android-devices-facial-recognition-fooled-by-3d-printed-head-but-not-face-id

[6] The NVLAP provides third-party accreditation to testing and calibration laboratories in response to legislative actions or requests from government agencies or private-sector organizations. NVLAP-accredited laboratories are assessed against the management and technical requirements from ISO/IEC 17025:2017

Concretely speaking, the proposed model consists of an *encoder*, mapping from input features to latent representations, and two classifiers operating on these underlying representations: (i) the *task-classifier*, for predicting the class labels (as bona-fide or attack), and (ii) the *species-classifier*, for predicting the type of species of the PAI. During the learning stage, the *encoder* is simultaneously trained to help the *task-classifier* as much as possible while trying to fool the *species-classifier*. To further discourage the latent representations of retaining any 'PAI-species'-specific traits, an additional training objective is introduced that enforces the latent distributions of different species to be as similar as possible. The result is a truly 'PAI-species'-independent model robust to detect unseen PAI species presented in the testing step.

The proposed adversarial training framework builds on those initially introduced by Ganin *et al* [GL15] and Ferreira *et al* [Fe19] in the context of domain adaptation; and of Feutry *et al* [Fe18] to learn learn anonymized representations. This work adds the following main contributions:

- The application of the adversarial training concept to the generalisation to unseen attacks problem in iris PAD;
- A training objective that is minimum if and only if the adversarial classifier - the *species-classifier*, produces a uniform distribution over the PAI species, meaning that our model is invariant to the PAI species seen in the training data.
- The introduction of an additional term to the adversarial training objective that further discourages the learned representations of retaining any species-specific information, by explicitly imposing similarity in their latent distributions.

The main definitions related to PAD concepts which will be used throughout this paper are the ones stated in the International Standard ISO/IEC 30107-3 Information Technology Biometric presentation attack detection Part 3: Testing and reporting [IS17].

This paper is organised as follows. This section summarises the proposed work and how it addresses the research question posed. Section 2 summarizes the related work. In section 3 the proposed methodology is detailed. Section 4 contains all the experimental setup including the results and discussion. Finally, the work is concluded in section 5.

2 Related work

Recent PAD methods in general, and iris-focused ones in particular, have demonstrated remarkable performances. However, a methodological limitation can be pointed as it is recurrently found that these results are obtained when training and test data comprise the same type of attacks - same PAIS. This problem has been addressed and proved that the performance rates of these PAD methods typically decrease significantly when the PAIS is new to the system [MS11, BD14, Se16]. This performance drop may be result of the large inter-'PAI-species' variability. A practical PAD system must operate in a 'PAI-species'-independent scenario, which means that the type of PAIS of the test set must not be seen during the training routine of the models. This problem is one of the crucial problems for

the development of real-world PAD systems and it has frequently been tackled in literature as an open-set or anomaly detection problem.

The pioneer work that raised the evaluation of PAD methods across different types and unseen PAIS appeared in the fingerprint domain with the work of Marasco and Sansone [MS11]. The works of Rattani & Ross [RSR15] and Sequeira & Cardoso [SC15], despite using different approaches, both relied on the idea of enforcing the knowledge of the bona fide presentations over the attacks to better deal with unseen PAIS. Bowyer and Doyle [BD14] studied the evaluation of a binary classification on contact lenses iris spoofing attacks. By using an unseen type on the test set the authors showed that using the same lens types in both the training and testing data can give a very misleading idea of the accuracy of the method. A step forward was made by combining methodologies designed for print and contact lenses attack [SMC14]. Eventually, the construction of a new database comprising several types of iris PAIS [RB15] allowed new evaluation scenarios. In [Se16] it is stated that whenever a new PAIS is presented in the test step, the performance of the classifier drops significantly and that an improvement can be obtained when a one-class classifier is trained only with bona fide presentations. One-class classification was also used for face in [AKC17]. With the rise of deep learning (DL) techniques, PAD methods have been proposed applying deep representations for iris, face and fingerprint [Me15, Pi18], following the same binary approach. Recent works investigate the robustness of DL fingerprint PAD methods to deal with unseen PAI species [To18].

Until recently, most of the proposed approaches, either assume overly optimistic assumptions about the attacker - binary classification approaches - or only use part of the data (and therefore, of the knowledge) available at training time to design the models - one-class approaches. Therefore, the goal of this work is to present an iris PAD method that uses the information of both bona fide and available attack presentations and is robust to unseen PAI species. This objective will be achieved by enforcing the learning of the task of distinguishing the bona fide from the attack presentations while at the same time ensuring the invariance between the different type of the PAI species.

3 Proposed Methodology

The proposed methodology combines an artificial neural network (ANN) with an adversarial training (AT) scheme. Specifically, the network is a Multilayer Perceptron that uses as input features extracted with a state of the art method [Se16], detailed in section 4.

The ultimate goal of our model is to learn latent 'PAI-species'-invariant representations, that preserve relevant information about the liveness properties and discard the 'PAI-species'-specific aspects - which may hamper the PAD classification task. To accomplish this purpose, we introduce a deep neural network along with an adversarial training scheme that is able to learn feature representations that combine both liveness discriminativeness and 'PAI-species'-invariance.

General description

Let $\mathbb{X} = \{\boldsymbol{X}_i, y_i, s_i\}_{i=1}^{N}$ denote a labeled dataset of N samples, where $\boldsymbol{X}_i$ represents the i-th feature vector, and y_i and s_i denote the corresponding class label and the PAI species,

respectively. $\mathbb{X}$ comprises elements from two classes: *bona fide* or *attack*. Let $\mathbb{X}^{bf}$ and $\mathbb{X}^{a}$ be these partitions and N^{bf} and N^{a} their cardinality, respectively.

With the aim at learning 'PAI-species'-invariant representations, the architecture of the proposed model is composed, as illustrated in Figure 1, by three main sub-networks or blocks, i.e. an *encoder*, a *task-classifier* and a *'PAI-species'-classifier*:

- an *encoder* network, which aims at learning an encoding function $h(\mathbf{X};\theta_h)$, parameterized by θ_h, that maps an input feature vector $\mathbf{X}$ to a latent representation $\boldsymbol{h}$;
- a *task-classifier* network to learn a task-specific function $f(\boldsymbol{h};\theta_f)$, parameterized by θ_f, that maps from $\boldsymbol{h}$ to the predicted probabilities $p(\mathrm{y}|\boldsymbol{h};\theta_f)$ of the two classes;
- a *species-classifier* network to learn a 'PAI-species'-specific function $g(\boldsymbol{h};\theta_g)$, parameterized by θ_g, that maps the same hidden representation $\boldsymbol{h}$ to the predicted probabilities $p(\mathrm{s}|\boldsymbol{h};\theta_g)$ of each PAI species.

During the learning stage, the parameters of both classifiers are optimized to minimize their specific errors on the training set. In addition, the parameters of the *encoder* network are optimized in order to minimize the loss of the *task-classifier* network while forcing the *species-classifier* to be a random guessing predictor. In the course of this AT procedure, the learned latent representations $\boldsymbol{h}$ are encouraged to be 'PAI-species'-invariant and highly discriminative for the PAD classification. To further discourage the latent representations of retaining any 'PAI-species'-specific traits, an additional training objective is introduced that enforces the latent distributions of different species to be as similar as possible. The result is a truly 'PAI-species'-independent model robust to new test PAI species.

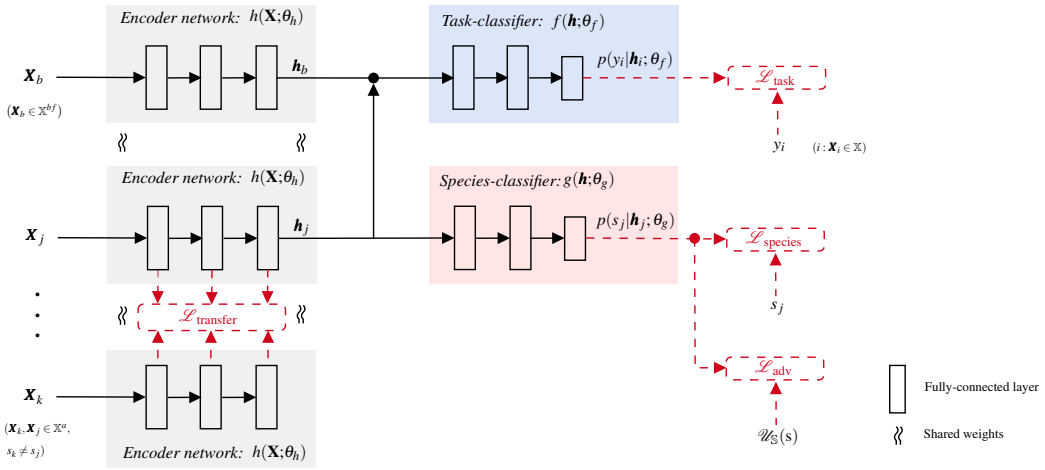


Fig. 1: The architecture of the proposed species-invariant neural network.

Adversarial training

By definition, the 'PAI-species'-invariant representations discard all 'PAI-species'-specific information and, as such, no function exists that maps such representations into the correct species. This naturally leads to an adversarial problem, in which: (i) a *species-classifier*

network $g(\cdot;\theta_g)$ receives latent representations $\boldsymbol{h} = h(\mathbf{X};\theta_h)$ from an *encoder* network $h(\cdot;\theta_h)$ and tries to predict the species s corresponding to feature vector $\mathbf{X}$ and (ii) the *encoder* network tries to fool the *species-classifier* network while still providing good representations for the *task-classifier* network $f(\cdot;\theta_f)$, which in turn receives the same representations $\boldsymbol{h}$ and aims to predict the task label y of $\mathbf{X}$. Therefore, the *species-classifier* network shall be trained to minimize the negative log-likelihood of correct task predictions:

$$\min_{\theta_g} \mathscr{L}_{\text{species}}(\theta_h,\theta_g) = \min_{\theta_g} \left\{ -\frac{1}{N^a}\sum_{i=1}^{N^a} \log p(s_i|h(\boldsymbol{X}_i;\theta_h);\theta_g) \right\}, \forall i : \boldsymbol{X}_i \in \mathbb{X}^a \tag{1}$$

So, in the perspective of the *encoder*, the predictions of the *task-classifier* should be as accurate as possible and the predictions of the *species-classifier* should be kept close to uniform, meaning that this latter classifier is not capable of doing better than random guessing the species type. Formally, this may be translated into the following constrained objective:

$$\min_{\theta_h,\theta_f} \mathscr{L}_{\text{task}}(\theta_h,\theta_f) = \min_{\theta_h,\theta_f} \left\{ -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|h(\boldsymbol{X}_i;\theta_h);\theta_f) \right\}, \tag{2}$$

$$\text{subject to } \frac{1}{N^a}\sum_{i=1}^{N^a} D_{\text{KL}}(\mathscr{U}_{\mathbb{S}}(\text{s})||p(\text{s}|h(\boldsymbol{X}_i;\theta_h);\theta_g) \leq \varepsilon, \forall i : \boldsymbol{X}_i \in \mathbb{X}^a \tag{3}$$

where D_{KL} is the Kullback-Leibler (KL) divergence and $\mathscr{U}_{\mathbb{S}}(\text{s})$ denotes the discrete uniform distribution on the random variable s, defined over the set of PAI species in the training set ($\mathbb{S}$). Here, $\varepsilon \geq 0$ determines how far from uniform the *species-classifier* predictions are allowed to be (as measured by the KL divergence). The choice of the uniform distribution implies the underlying assumption that the training set is balanced relatively to the number of examples per species (which should be true for most practical datasets).

The constraint inequality (3) may be rewritten as the equation (4)

$$\mathscr{L}_{\text{adv}}(\theta_h,\theta_g) = -\frac{1}{N^a|\mathbb{S}|}\sum_{i=1}^{N^a}\sum_{\text{s}\in\mathbb{S}} \log p(\text{s}|h(\boldsymbol{X}_i;\theta_h);\theta_g) \leq \varepsilon + \log|\mathbb{S}|, \tag{4}$$

Then, the constrained optimization problem may be equivalently formulated as in (5).

$$\min_{\theta_h,\theta_f} \mathscr{L}(\theta_h,\theta_f,\theta_g) = \min_{\theta_h,\theta_f} \left\{ \mathscr{L}_{\text{task}}(\theta_h,\theta_f) + \lambda \mathscr{L}_{\text{adv}}(\theta_h,\theta_g) \right\}, \tag{5}$$

where $\lambda \geq 0$ depends on the value ε and $\mathscr{L}_{\text{adv}}$ plays the role of an adversarial loss with respect to the species classification loss $\mathscr{L}_{\text{species}}$.

'Species'-transfer training objective

In addition to the adversarial training, a species-transfer training objective is added to further encourage the latent representations $\boldsymbol{h}$ to be species-invariant. Thus, an additional term is introduced in objective (5), the so-called species-transfer loss $\mathscr{L}_{\text{transfer}}$. The core

idea of $\mathscr{L}_{\text{transfer}}$ is to enforce the latent distributions of different species to be as similar as possible. In practise, this is achieved by minimizing the difference between the hidden representations of different species, at each layer of the *encoder* network.

To measure the species distribution difference at the m-th layer, $m = 1, ..., M$, we compute a distance $\mathscr{D}^{(m)}$ between the hidden representations $h^{(m)}(\cdot;\theta_h)$ of two species s and t at the output of that layer, such that:

$$\mathscr{D}^{(m)}(s,t;\theta_h) = \left|\left| \frac{1}{N_s} \sum_{i:\, s_i=s} h^{(m)}(\boldsymbol{X}_i;\theta_h) - \frac{1}{N_t} \sum_{j:\, s_j=t} h^{(m)}(\boldsymbol{X}_j;\theta_h) \right|\right|_2^2, \tag{6}$$

where $||\cdot||_2$ is the ℓ^2-norm, and N_s and N_t denote the number of training examples of species s and t, respectively.

The overall species-transfer loss $\mathscr{L}_{\text{transfer}}$ is then a weighted sum of the losses computed at each layer of the *encoder* network, such that:

$$\mathscr{L}_{\text{transfer}}(\theta_h) = \sum_{m=1}^{M} \beta^{(m)}\, \mathscr{L}^{(m)}_{\text{transfer}}(\theta_h) = \sum_{m=1}^{M} \beta^{(m)} \sum_{s\in\mathbb{S}} \sum_{\substack{t\in\mathbb{S},\\ t\neq s}} \mathscr{D}^{(m)}(s,t;\theta_h), \tag{7}$$

where $\beta^{(m)} \geq 0$ is a hyperparameter that controls the relative importance of the loss obtained at the m-th layer and the species-transfer loss at the m-th layer is the sum of the pairwise distances between all species.

By combining (5) and (7), the *encoder* and *task-classifier* networks are trained to minimize the following loss function:

$$\min_{\theta_h,\theta_f} \mathscr{L}(\theta_h,\theta_f,\theta_g) = \min_{\theta_h,\theta_f} \left\{ \mathscr{L}_{\text{task}}(\theta_h,\theta_f) + \lambda \mathscr{L}_{\text{adv}}(\theta_h,\theta_g) + \gamma \mathscr{L}_{\text{transfer}}(\theta_h) \right\}, \tag{8}$$

where $\gamma \geq 0$ is the weight that controls the relative importance of the species-transfer term.

Summing up, the adversarial training is done by alternatively training both the *encoder* and the *task-classifier* to minimize (8) or training the *species-classifier* to minimize (2).

4 Experimental setup

PAD Performance Evaluation Metrics:

Bona-fide Presentation Classification Error Rate (BPCER) and *Attack Presentation Classification Error Rate" (APCER)* as defined in the ISO/IEC 30107-3 [IS17]. The *Average Classification Error Rate (ACER)*, given by their mean, though deprecated, is used to allow comparison with the literature.

Dataset and Evaluation protocol:

Visible Spectrum Iris Artefact (VSIA) Database [RB15] which comprises five different presentations combining print and electronic screen attacks: (i) Print Attack (PA); (ii) iPad

Tab. 1: Hyperparameters sets.

Hyperparameters	Acronym	Set
Leaning rate	-	$\{1e^{-04}, 1e^{-03}\}$
ℓ^2-norm coefficient	-	$\{1e^{-05}, 1e^{-04}\}$
encoder dense layers	L_e	$\{3,4\}$
$\mathscr{L}_{\text{adv}}$ weight	λ	10 values n $\in \{n : n = log_{10}C \wedge n \in [1e^{-03}, 1]\}$
$\mathscr{L}_{\text{transfer}}$ weight	γ	10 values n $\in \{n : n = log_{10}C \wedge n \in [1e^{-03}, 1]\}$

Electronic Screen Attack (ESA); (iii) Samsung Galaxy Tab ESA; (iv) combined PA & ESA using iPad; and (v) combined PA & ESA using Samsung Pad. The methods are evaluated by leaving out one PAI species for testing. The training set is therefore divided in one specie for validation and the remaining used for training. Also the same set of samples are used for testing across the different experiments to allow precise comparison of the results.

Feature Extraction and Classifiers:

Weighted Local Binary Patterns **(wLBP)** [ZST10] method combines LBP with a Scale Invariant Feature Transform (SIFT). The wLBP was chosen to allow comparison with the literature, the work [Se16] used several handcrafted feature extraction methods combined with a Support Vector Machine (SVM) to study the generalisation capability in face of an unseen PAI specie (wLBP provided the best results). In the proposed work, the SVM classifier was replaced by an artificial neural network - a Multilayer Perceptron (MLP). It is worth mentioning that, for a fair comparison, the MLP in the baseline method has the same architecture as the *task-classifier* module of the proposed model.

Implementation details:

All deep models were implemented in PyTorch and trained with the Adam optimization algorithm with batch size equal to 64. For reproducibility purposes, the source code as well as the weights of the trained models and the features used will be made publicly available online[7]. The hyperparameters common to all the implemented models (i.e., learning rate and ℓ^2 regularization weight) as well as some hyperparameters specific to the proposed model (i.e., λ and γ) were optimized by means of a grid search approach and cross-validation on the training set (see Table 1 for more details). The species-transfer penalty $\mathscr{L}_{\text{transfer}}$ is applied to the last two layers of the *encoder* network with a relative weight of 1. Regarding the architecture of the proposed model, the *encoder* simply consists of a sequence of L_e fully-connected layers with 128 neurons, followed by a Rectified Linear Unit (ReLU) activation function. As depicted in Table 1, L_e was also optimized by means of a grid search approach and cross-validation on the training set. Both classifiers, i.e. the *task-classifier* and the *species-classifier*, follow the same network topology. In particular, it comprises a total of 3 hidden layers with 256 neurons, also with a ReLU, along with a softmax output layer. The number of nodes of the output layer of the *species-classifier* is defined accordingly to the number of species in the training set.

Results and discussion:

[7] github.com/pmmf/DeepAdvNN-IrisPAD.

The experimental results were obtained for: i) baseline method ($wLBP+MLP$); and ii) proposed method ($wLBP+MLP_{reg}$) - that adds the adversarial training and transfer learning objective to the MLP.

In Table 2 the Baseline ($wLBP+MLP$) and the Proposed ($wLBP+MLP_{reg}$) methods are compared with the state-of-the-art results [Se16] obtained with a similar evaluation (training a SVM leaving-one-out PAIS for testing). Despite the differences between the works, this comparison is, to the best of our knowledge, the possible one to make of the proposed method with the literature as no other works in iris PAD perform a similar evaluation.

Tab. 2: Results of one state-of-art and the evaluated approaches. (ACER in % was used for comparison with state of the art method.)

Method	**ACER (%)**					
	Attack1	***Attack2***	***Attack3***	***Attack4***	***Attack5***	***Average***
$wLBP+SVM$ [Se16]	21.15	9.61	**1.92**	**4.32**	2.88	7.98
Baseline $wLBP+MLP$	22.00	**7.00**	5.50	10.00	4.50	9.80
Proposed $wLBP+MLP_{reg}$	**18.00**	**7.00**	2.00	5.50	**2.50**	**7.00**

Comparing the ACER for each attack and their Average values in Table 2, it can be claimed that uniquely the replacement of the *SVM* for a *MLP* does not result in an improvement. This can be explained by the fact that the dataset has a very limited size and therefore the *MLP* method tends to overfit due to the lack of training samples. It was not for no reason that *SVMs* ruled for a long time in the pattern recognition domain. However, the regularization methods added to the *MLP* lead to an improvement of the average error providing the best value of 7.00%.

The PAD error rates of the baseline and the proposed methods are compared in Table 3. It is clear that the adversarial learning added to the *MLP* lead to a significant improvement in the PAD robustness of the method as it decreased the APCER and BPCER in most cases. These results clearly enforce the idea that the application of deep learning techniques with additional strategies will provide breakthroughs in this challenge.

Tab. 3: PAD error rates of the baseline and the proposed approaches. (APCER and BPCER are in %.)

Method	**PAD metrics (%)**											
	Attack1		***Attack2***		***Attack3***		***Attack4***		***Attack5***		***Average***	
	APCER	BPCER	APCER	BPCER	APCER	BPCER	APCER	BPCER	APCER	BPCER	APCER	BPCER
Baseline $wLBP+MLP$	39.00	5.00	9.00	**5.00**	**0.00**	11.00	12.00	8.00	3.00	6.00	12.60	7.00
Proposed $wLBP+MLP_{reg}$	**33.00**	**3.00**	**6.00**	8.00	**0.00**	**4.00**	**6.00**	**5.00**	**0.00**	**5.00**	**9.00**	**5.00**

Table 4 compares the loss ($\mathscr{L}_{\text{task}}$) and accuracy of the baseline ($wLBP+MLP$) and the proposed ($wLBP+MLP_{reg}$) methods. Again, there is a clear gain when the regularization methods are applied as the later provides the best values in every case.

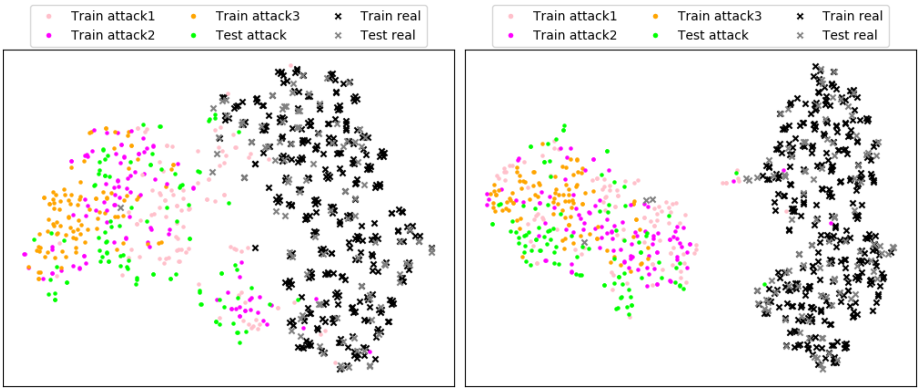
Tab. 4: Loss ($\mathscr{L}_{\text{task}}$) and Accuracy of the baseline and the proposed methods.

Method	**Model performance metrics (%)**											
	Attack1		***Attack2***		***Attack3***		***Attack4***		***Attack5***		***Average***	
	Loss	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss	Acc
Baseline $wLBP+MLP$	0.50	78.00	**0.27**	**93.00**	0.10	94.50	0.25	90.00	0.12	95.50	0.25	90.20
Proposed $wLBP+MLP_{reg}$	**0.42**	**82.00**	**0.27**	**93.00**	**0.07**	**98.00**	**0.18**	**94.50**	**0.10**	**97.50**	**0.21**	**93.00**

'PAI-species'-invariant latent space visualization:

To further demonstrate the effectiveness of the proposed model, the t-distributed stochastic neighbor embedding (t-SNE) [vdMH08] was used to perform a visual inspection of the latent representations. Figure 2 is presented to further demonstrate the effectiveness of the proposed model in promoting species-invariant latent representation spaces.

The t-SNE plots clearly demonstrate how the proposed model is more capable of imposing species-independence in the latent representations. In the latent representations space derived from the proposed model, representations of the same PAD class - bona fide or attack - are close to each other and well mixed, while it keeps latent representations of different classes far apart. In particular, samples belonging to the attack class are still close disregarding the fact that they may be originated from different types of PAI-species. Very relevant is the fact that the unseen species present in the test are better mixed with the other species (from the training) in the proposed model. By analyzing these plots, it is also possible to observe that the latent representations of the different species (belonging to the attack class) tend to be closer to each other in the latent space for the proposed model. In addition, there is some overlapping between clusters of different classes in the baseline model, whereas the proposed model achieved by far a better species-invariance and class separability.



(a) Baseline 1 (b) Proposed model

Fig. 2: Two-dimensional projection of the latent representation space with t-distributed stochastic neighbor embedding (t-SNE) (colored • denote different PAI species; × are bona fide presentations).

5 Conclusions

This work proposed a method to improve the robustness and generalisation capacity of an iris PAD method to new attacks. The goal of the proposed model is to learn latent representations invariant to the PAI species that preserve relevant information about the PAD properties while discarding the 'PAI-species'-specific aspects that may hamper the PAD classification task. The proposed regularisation strategies made the PAD method 'PAI-species'-independent and robust to new test PAIS. The experiments were based in comparing a baseline MLP and a MLP trained with adversarial strategies using as input highly discriminative features (wLBP) extracted from the images. When comparing the baseline

MLP to a SVM classifier the results are quite similar or even worse. This can be explained simply by the fact that the dataset has a very limited size and the *MLP* method will overfit. However, applying the regularisation strategies significantly improved the PAD robustness of the method. The obtained results clearly enforce the idea that the application of deep learning techniques with additional strategies will provide breakthroughs in this challenge.

Acknowledgements

This work was financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-030707".

References

[AKC17] Arashloo, S. R.; Kittler, J.; Christmas, W.: An Anomaly Detection Approach to Face Spoofing Detection: A New Formulation and Evaluation Protocol. IEEE Access, 5:13868–13882, 2017.

[BD14] Bowyer, K.W.; Doyle, J.S.: Cosmetic Contact Lenses and Iris Recognition Spoofing. Computer, 47(5):96–98, May 2014.

[CB18] Czajka, Adam; Bowyer, Kevin W.: PAD for Iris Recognition: An Assessment of the State-of-the-Art. ACM Comput. Surv., 51(4):86:1–86:35, July 2018.

[Ch17] Chaos Computer Clubs breaks iris recognition system of the Samsung Galaxy S8. `www.ccc.de/en/updates/2017/iriden`, accessed on 29th of May of 2017.

[Fa19] Facetec liveness detection technology is iBeta/NIST Certified Anti-Spoofing Level 12. `www.zoomlogin.com/`, accessed on 10th of June of 2019.

[Fe18] Feutry, Clément; Piantanida, Pablo; Bengio, Yoshua; Duhamel, Pierre: Learning anonymized representations with adversarial neural networks. arXiv:1802.09386, 2018.

[Fe19] Ferreira, Pedro M.; Pernes, Diogo; Rebelo, Ana; Cardoso, Jaime S.: Learning signer invariant representations with adversarial training. In: 12th International Conference on Machine Vision (ICMV 2019). 2019.

[GFC19] Galbally, Javier; Fierrez, Julian; Cappelli, Raffaele: An Introduction to Fingerprint Presentation Attack Detection. In (Marcel, Sébastien; Nixon, Mark S.; Fierrez, Julian; Evans, Nicholas, eds): Handbook of Biometric Anti-Spoofing: Presentation Attack Detection. Springer International Publishing, Cham, pp. 3–31, 2019.

[GL15] Ganin, Yaroslav; Lempitsky, Victor: Unsupervised Domain Adaptation by Backpropagation. In: Proc. 32nd Int. Conf. ML. volume 37, Lille, France, pp. 1180–1189, 2015.

[IS17] ISO/IEC JTC1 SC37: Information Technology - Biometrics - Presentation attack detection Part 3: Testing and Reporting. ISO Int. Organization for Standardization, 2017.

[Me15] Menotti, D.; Chiachia, G.; Pinto, A.; Robson Schwartz, W.; Pedrini, H.; Xavier Falcao, A.; Rocha, A.: DeepRep.Iris,Face,and Fingerp.Spoof.Det. TIFS, 10(4):864–879, 2015.

[MS11] Marasco, Emanuela; Sansone, Carlo: On the Robustness of Fingerprint Liveness Detect. Alg. against New Materials used for Spoofing. In: BIOSIGNALS. pp. 553–558, 2011.

[Pi18] Pinto, Allan; Pedrini, Helio; Krumdick, Michael; Becker, Benedict; Czajka, Adam; Bowyer, Kevin W; Rocha, Anderson: Counteracting presentation attacks in face, fingerprint, and iris recognition. Deep Learning in Biometrics, 245, 2018.

[RB15] Raghavendra, R.; Busch, C.: Robust Scheme for Iris Pres. Attack Det. Using Multiscale Binarized Statistical Image Features. IEEE TIFS, 10(4):703–715, 2015.

[RSR15] Rattani, A.; Scheirer, W.J.; Ross, A.: Open Set Fingerprint Spoof Detection Across Novel Fabrication Materials. IEEE TIFS, 10(11):2447–2460, Nov 2015.

[SC15] Sequeira, Ana F.; Cardoso, Jaime S.: Fingerprint liveness detection in the presence of capable intruders. Sensors, 15:14615–14638, 2015.

[Sc19] Scherhag, U.; Rathgeb, C.; Merkle, J.; Breithaupt, R.; Busch, C.: Face Recognition Systems Under Morphing Attacks: A Survey. IEEE Access, 7:23012–23026, 2019.

[Se16] Sequeira, A. F.; Thavalengal, S.; Ferryman, J.; Corcoran, P.; Cardoso, J. S.: A realistic evaluation of iris presentation attack detection. In: 39th TSP. pp. 660–664, June 2016.

[SMC14] Sequeira, Ana F.; Murari, Juliano; Cardoso, Jaime S.: Iris Liveness Detection Methods in Mobile Applications. In: Proc. Int. Con. on CV Theory and Applic. pp. 22 – 33, 2014.

[To18] Tolosana, Ruben; Gomez-Barrero, Marta; Kolberg, Jascha; Morales, Aythami; Busch, Christoph; Ortega-Garcia, Javier: Towards fingerprint presentation attack detection based on convolutional neural networks and short wave infrared imaging. In: 2018 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–5, 2018.

[vdMH08] van der Maaten, Laurens; Hinton, Geoffrey: Visualizing Data using t-SNE. Journal of Machine Learning Research, 9:2579–2605, 2008.

[ZST10] Zhang, Hui; Sun, Zhenan; Tan, Tieniu: Contact lens detection based on weighted LBP. In: 20th ICPR. pp. 4279–4282, 23 - 26 August 2010.