Pre-print of article that will appear in T-IFS, vol. 14, no. 6, pp. 1419-1431, June 2019.

The published article is available on https://doi.org/10.1109/TIFS.2018.2878542

# Ensemble of Multi-View Learning Classifiers for Cross-Domain Iris Presentation Attack Detection

Andrey Kuehlkamp, *Student Member, IEEE,* Allan Pinto, *Student Member, IEEE,*
Anderson Rocha, *Senior Member, IEEE*, Kevin W. Bowyer, *Fellow, IEEE,* Adam Czajka, *Senior Member, IEEE*

***Abstract*—The adoption of large-scale iris recognition systems around the world has brought to light the importance of detecting presentation attack images (textured contact lenses and printouts). This work presents a new approach in iris Presentation Attack Detection (PAD), by exploring combinations of Convolutional Neural Networks (CNNs) and transformed input spaces through binarized statistical image features (BSIF). Our method combines lightweight CNNs to classify multiple BSIF views of the input image. Following explorations on complementary input spaces leading to more discriminative features to detect presentation attacks, we also propose an algorithm to select the best (and most discriminative) predictors for the task at hand. An ensemble of predictors makes use of their expected individual performances to aggregate their results into a final prediction. Results show that this technique improves on the current state of the art in iris PAD, outperforming the winner of LivDet-Iris 2017 competition both for intra- and cross-dataset scenarios, and illustrating the very difficult nature of the cross-dataset scenario.**

## I. Introduction

HOW can we distinguish between two individuals without reasonable doubt? This question has motivated biometric researchers for centuries — from the pioneer works of Bertillon and Galton to recent advances in biometric-enabled mobile payments and wearable devices. Numerous biometric characteristics emerged, and sometimes faded away, as the field progressed, but one biometric trait that has surely withstood the test of time is iris recognition. The iris pattern is unique and "determined epigenetically by random events in the morphogenesis of this tissue" [1], and thus offers high discrimination power, making it useful in distinguishing even identical twins [2].

The recognition power and matching speed of iris recognition has propelled it into use in large-scale applications, *e.g.*, Unique ID in India [3], [4], and the NEXUS system operated jointly by the Canada Border Services Agency and the U.S. Customs and Border Protection to speed up the identification of pre-screened travelers [5]. As iris recognition becomes more pervasive, the number and variety of attempted attacks naturally intensify, and the problem of *presentation attack detection* (PAD) becomes an essential research topic.

According to the standardized vocabulary in ISO/IEC 30107-1, a *presentation attack* is a "presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system" [6]. An impostor may seek to obtain unauthorized access to an authentication system by *impersonating* a legitimate user, or may want to intentionally *conceal* his or her identity to avoid recognition. Presentation attacks can be realized in various ways, in particular by presenting artifacts, such as paper printouts or textured contact lenses, non-conformant use of a biometric sensor, or even presenting cadaver eyes to the sensor. In this paper, we focus on attacks related to presentation of artificial objects.

The Iris Liveness Detection Competition (LivDet-Iris, www.livdet.org) was begun in 2013. The third and the most recent edition took place in 2017 [7]. This competition is focused on properly measuring how well current technology withstands presentation attacks employing artifacts. The sequence of

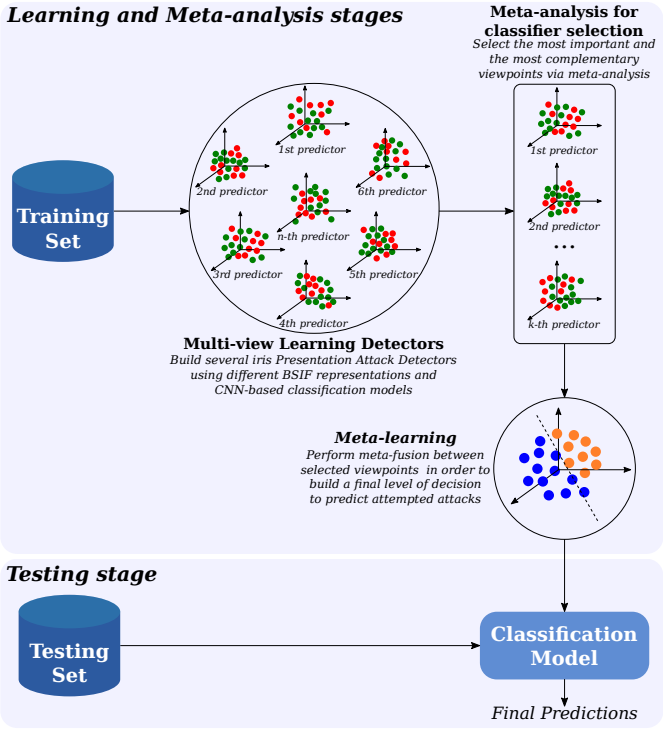



Fig. 1: Overview of the proposed method. The first step is generating multiple views from dataset by training lightweight CNNs fed with different BSIF representations, referred to as multi-view-CNN predictors. Next, we select the most promising predictors according to their relevance and complementarity, and combine them via meta-fusion approach.

A. Kuehlkamp, K. Bowyer and A. Czajka are with the University of Notre Dame, USA. E-mail: {akuehlka,kwb,aczajka}@nd.edu

A. Pinto and A. Rocha are with the Institute of Computing, University of Campinas (Unicamp), Av. Albert Einstein, 1251, Campinas, SP, Brazil, 13083-852. E-mail: {allan.pinto,anderson.rocha}@ic.unicamp.br.

Manuscript received ...; revised ...

competitions provides an important insight into the pace of evolution of iris PAD methods. The results reported in the 2017 competition show that iris PAD algorithms are still far from achieving acceptable detection rates. Moreover, the LivDet 2017 results suggest that challenging evaluation protocols, such as cross-dataset and cross-sensor setups — hereinafter referred to as cross-domain — can be considered the major limitation of current PAD algorithms and a current open research problem.

In the face of the evident need for better iris PAD, this paper introduces a new technique based on exploiting multiple transformations of the input data so as to enhance complementary patterns, leading to a more discriminant manifold separating genuine authentication attempts from attacks. The multiple transformations are obtained through the combination of hand-crafted and data-driven approaches. While Binarized Statistical Image Features (BSIF) [8] is a popular, effective and partially hand-crafted texture descriptor, Convolutional Neural Networks (CNN) complement the repertoire with powerful description methods capable of learning even the subtlest discrimination clue present in the available training data through a series of non-linear transformations on the input data.

Although BSIF- and CNN-based methods have been used before in iris PAD, to our knowledge there are no published papers dealing with the challenges of cross-domain deployments, or exploiting complementary properties of various feature extractors and classification strategies when defining a discriminative manifold for the problem under cross-domain constraints. In this vein, we introduce a way of combining both methodologies, hand-crafted and data-driven, in order to offer an iris PAD algorithm that better generalizes to unknown attack types. In addition, we also present a novel meta-analysis algorithm for selecting and aggregating the most prominent data views (*i.e.*, transformations), based on two well-known techniques for feature selection: the random forest importance feature weighting and the inter-rater agreement measures, so as to provide the most accurate detection method with the least possible computational impact. Fig 1 gives an overview of the proposed method.

In summary, the main contributions and novelties of this work are:

- A new approach that leverages multiple pre-trained BSIF filters to effectively train lightweight CNNs;
- A new fusion algorithm that selects and combines multiple classifiers, considering their importance and complementarity;
- A thorough cross-domain evaluation of the problem on datasets currently used to document the state of the art in the field;
- A new PAD method that outperforms the winner of the most recent LivDet-Iris competition, the authoritative international challenge on the subject.

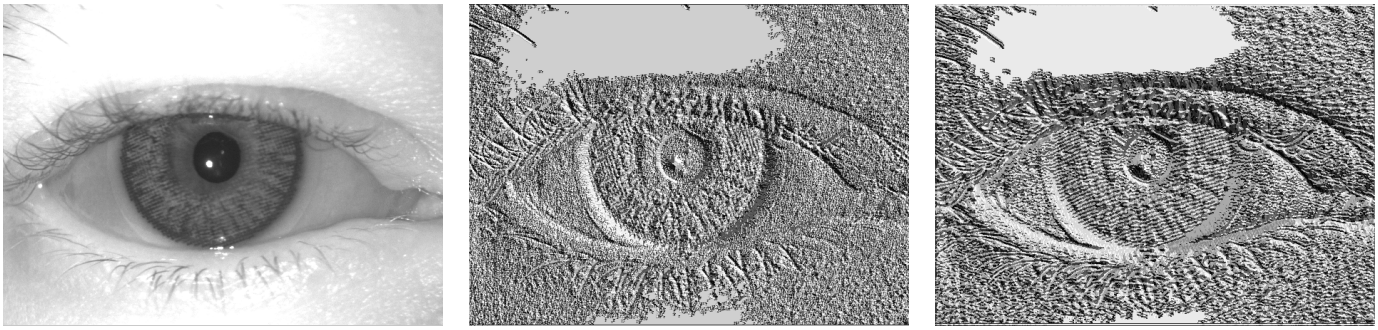
To encourage reproducibility, the source code of our implementation will be publicly available on GitHub. The datasets are already available through the LivDet competition. In the remainder of this paper, we briefly survey important iris PAD methods in prior art (Sec. II) and introduce the proposed methodology (Sec. III). Then we present experiments and validation (Sec. IV) and, finally, draw conclusions and present possible future work (Sec. V).

## II. Related Work

The first ideas for countermeasures against iris presentation attacks were proposed some 18 years ago by Daugman [9] and became a basis of many current effective PAD methods. Early Daugman's concepts include finding anomalies in Fourier spectrum to detect printed irises, either on a paper or on a contact lens, detection of specular "Purkinje" reflections from both the cornea and the lens, or investigating pupil size variations, either spontaneous ("hippus") or stimulated by visible light.

In general, there are two goals in presentation attacks, impersonation or identity concealment. The first successful demonstration of impersonation with the use of a commercial sensor was shown by Thalheim *et al.* [10]. They used iris images printed on a paper, with a hole cut where the pupil was printed, to make a successful impersonation attack on a commercial iris recognition system with authentic eyes previously enrolled. The first use of one's eye to evade recognition observed in an operational environment was recorded at the border crossing point employing iris recognition in the United Arab Emirates [11]. The attackers administered eye drops to make the pupil excessively dilated. This made the iris texture deformed too severely to be compensated by feature extraction and matching algorithms, and hence generating false non-matches. Since these early demonstrations of vulnerabilities, other presentation attack instruments have been studied, including use of textured contact lenses that partially occlude the actual iris texture [12], presentation of iris images displayed on a screen [13], or use of prosthetic eyes [14]. After recent studies by Trokielewicz *et al.* [15] presenting that post-mortem iris recognition is possible for a couple of weeks after death, cadaver eyes can also be considered as a potential presentation attack instrument.

There is a rich literature on PAD methods presenting various levels of sophistication, and putting different requirements on the sensors' configuration and signals necessary to detect presentation attacks. In the largest group of PAD methods, a single near-infrared iris image, compliant to ISO/IEC 19794-6, is used in both identity verification and presentation attack detection. The *hand-crafted* approaches use various image descriptors to calculate image features, which are used to distinguish between authentic irises and artifacts typically through the use of Support Vector Machine classifiers. Popular techniques used in calculation of PAD-related iris image features are Binarized Statistical Image Features (BSIF) [16], Local Binary Patterns (LBP) [12], Binary Gabor Patterns (BGP) [17], Local Contrast-Phase Descriptor (LCPD) [18], Local Phase Quantization (LPQ) [19], Scale Invariant Descriptor (SID) [20], Scale Invariant Feature Transform (SIFT) and DAISY [21], Locally Uniform Comparison Image Descriptor (LUCID) and CENsus TRansform hISTogram (CENTRIST) [22], Weber Local Descriptor (WLD) [18], Wavelet Packet Transform (WPT) [23] or image quality descriptors proposed by Galbally *et al.* [24]. Instead of "hand-crafting" effective



(a) Raw iris image (b) $B_{3\times3\times5}$ (c) $B_{7\times7\times12}$

Fig. 2: Examples of different BSIF representations $B_{n\times n\times l}$ (2b,2c) of the same image (2a). Note how the structure of the contact lens is highlighted through the different representations.

feature extractors, one may also benefit from recently popular *data-driven* approaches that learn directly from the data how to process and classify iris images to solve the PAD task [21], [25]–[29].

The above PAD methods rely upon an iris image that is typically used later for biometric recognition, and hence they can be implemented in the existing iris recognition sensors. However, if some hardware adaptations are possible, and some more complicated static features of the eye can be measured, one may consider multi-spectral analysis [23], [30]–[32] or estimation of three-dimensional iris features [33]–[36] as potential PAD techniques. Making the PAD more complex, one may consider measuring spontaneous dynamic features of the eye, such as micro-movements of an eyeball, either using Eulerian video magnification [37] or by using an eye-tracking device [38]. If there is a possibility to additionally stimulate the eye with varying visible light, and measure its reaction, the use of pupil dynamic models may help to easily detect static or oddly-behaving artifacts [32], [39]. A recent survey by Czajka and Bowyer provides a comprehensive assessment of the state of the art in iris PAD [40].

In [41], a PAD approach based on deep features extracted from different iris regions and classified by feature and score-level fusion of SVM is described. The authors report very low error rates on a subset of the LivDet-Iris 2017 datasets. Kontschieder *et al.* [42] published a work that seeks to combine deep convolutional networks for feature extraction, with the classification power of Random Forests. They describe an alternative approach to train Random Forest classifiers, which consists of a stochastic version of decision trees that are trainable through backpropagation. Random Forests trained with this method can either be standalone classifiers or act as alternative classifiers on top of a CNN. The authors claim to have outperformed the state of the art in image classification when integrating it to a GoogLeNet network.

Evaluation of PAD reliability significantly differs from statistical evaluation of biometric recognition. ISO/IEC JTC1 subcommittee 37 issued both the PAD-related vocabulary [6], and recommendations on how to evaluate and report the PAD-related performance [43]. An important effort related to iris PAD evaluation is the LivDet-Iris competition series (http://livdet.org/), which has had editions in 2013 [44], 2015 [45] and 2017 [7]. The 2017 edition of LivDet-Iris is the most recent, global, independent evaluation of PAD algorithms for detection of iris printouts and textured contact lenses. This paper follows exactly the LivDet-Iris 2017 evaluation protocol, and the results presented herein are directly comparable with the LivDet-Iris 2017 winning solution.

## III. Proposed Method

The task of the PAD algorithm is to capture differences between an actual live iris and either non-iris object, non-conformant use of an actual iris, or a cadaver eye. In the simplest scenario, when only static features of a single iris image are used, the PAD method recognizes anomalies in the presented iris texture. However, given different fabrication processes and the richness of details in an iris, it is likely that any single texture descriptor cannot capture all necessary leads hinting at a possible attack. Therefore fusions of different texture descriptors (*e.g.*, LBP, BSIF, Gabor filters) and classifiers (*e.g.*, Support Vector Machines, Neural Networks, etc.) have also been explored in prior literature.

Although texture analysis has been a staple in iris research over the years, development of image processing methods that capture such intricate textures has been a challenge. Therefore, more recently some researchers have brought to bear data-driven techniques, especially deep CNNs, to learn directly from training data the iris features useful in PAD. Normally, the input of such networks are the raw pixels themselves. Even though such data-driven methods have led to good results, these models do not work well in the so-called cross-domain setup, *i.e.*, when different training/testing conditions are part of the problem. As such, the conditions during training are not enough to allow robust generalization during testing.

This is where our work herein comes into play. In this section, we present a way of accelerating and empowering CNNs to capture texture patterns in such a way that differences among genuine and attack samples can be more easily spotted. Conscious of the representational power of CNNs to properly learn discriminative features, but at the same time of their innate need of large amounts of training data, we facilitate the process of learning by feeding the network with transformed input that highlights texture features important to make a distinction between a live iris and a presentation attack iris. We

achieve this by transforming the input image into Binarized Statistical Image Features (BSIFs) [46]. Second, it is likely that looking at the iris texture patterns from different vantage points, including different scales, might allow us extracting more features capturing differences between authentic iris patterns and artifacts.

Thus, we consider using multiple BSIF filter sets, characterized by a scale $l$ and a depth $n$ (number of filter kernels in a single set) to create BSIF representations $B_{n\times n\times l}$ for each $(i, j)$ pixel of the original image $I$:

$$B_{n\times n\times l}(i,j) = \sum_{k=0}^{n-1} b_k(i,j)2^k$$

where

$$b_k(i,j) = \begin{cases} 1 & \text{if } s_k(i,j) \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

$$s_k(i,j) = \sum_{u=0}^{l-1}\sum_{v=0}^{l-1} w_{k;n,l}(u,v)I(i+u, j+v),$$

and $w_{k;n,l}$ is the $k$-th filter kernel in the set of filters derived for a given $n$ and $l$. Using original sets of BSIF filters, as proposed by Kannala and Rahtu [46], allows calculating 60 different BSIF representations of a single image. Each of these 60 image transformations is then used by one CNN to learn higher-level features for discriminating authentic and presentation attack irises. Finally, with different learned features at hand, a natural question is how to effectively select the best ones for PAD detection while eliminating the non-representative ones. For this we exploit two different fusion schemes: one based on random-forest feature weighting and meta-learning strategies, and the second relying upon inter-rater agreement measures.

BSIF kernels included into each of 60 sets were calculated from natural images in a way that maximizes statistical independence of filter responses, and the Independent Component Analysis was used for this purpose [46]. Assuming that patterns observed in artifacts are statistically independent of textures observed in authentic irises, such decomposition of images, and then building BSIF representations of image based on binarized filter responses, may facilitate calculating PAD-related features from BSIF representations instead of raw images.

While it is theoretically possible to train a CNN to obtain the same filters as BSIF in its first convolutional layer, the cost function and the training process would be significantly different from our goals and would likely require more training data than one normally has for solving a PAD problem. To reduce the amount of training data, we thus transform the images to a new space represented through BSIF operations and this new input serves as an additional transformation layer for the CNNs. As we go toward the output layers in the CNN, such features are further specialized to higher-level representations. The idea is that feeding the network with a transformed input would allow it to more quickly achieve such deeper representations without extensive training. However, some useful features may still not be learned from this transformed space. Therefore, drawing on good results from previous studies in the PAD literature [26], [47], in addition to learning features from these transformed spaces, we also consider learning features directly from raw image inputs. At the end, with fusion strategies, we can show how such different treatments are complementary and how some of them do not contribute to separate authentic samples from artifacts. The key aspect of our methodology is that different filters learned from transformed spaces capture richer details than just a single representation and translate to better results in the difficult cross-dataset setup, the one in which training and testing conditions are different.

Finally, we end up with 61 CNN-based predictors: 60 fed by BSIF representations and 1 fed by the raw image. We refer to such complementary CNNs learned with different input representations as multi-view-CNN predictors. Finding the optimal decomposition of the numerous possible configurations of the BSIF filter sets is not straightforward. That is exactly why the combination of CNNs is powerful.

Since one of the CNN-based predictor operates on a raw image, it is interesting to compare its first-layer filters with those used in BSIF transformation. Visual comparison of these filters reveals some similarities: at a small scale ($3 \times 3$), most of them resemble edge, corner or dot detectors, similarly to what we would expect in the V1/V2 regions of the brain specialized in roughly describing visual inputs through edges and corners. However, BSIF $3 \times 3$ filters and CNN $3 \times 3$ filters are different. This means that the CNN operating on raw images developed its own way to preprocess the images in the first layer when compared to BSIF-based transformation, which may suggest that this predictor is complementary to the remaining 60 predictors, and it is worth using in the fusion.

Fig. 1 illustrates the main steps in our approach, which are explained in details in the following sections.

### A. Feature Extraction

We compute BSIF representations by exploiting filter width size ranges from $3 \times 3$ to $17 \times 17$, and depth (*i.e.*, number of filters in a set) ranges from 5 to 12. Fig. 2 shows some examples of BSIF views of the same image presenting an iris with a textured contact lens. Note how different BSIF representations highlight the pattern of a contact-lens.

We use OSIRIS [48] (version 4.1) [1] to find the center of the iris in the original (not BSIF-transformed) image. Next, we crop a $260 \times 260$ region around the iris in both the BSIF-transformed images and the original image. These cropped samples are input to the CNNs in the next step. This $260 \times 260$ size is based on the size of iris images in commercial sensors ($640 \times 480$) and the ISO/IEC 19794-6 recommendation for a minimum 120 pixels across the iris diameter.

To substantiate the descriptive advantage offered by BSIF filtering over the raw image input, we performed a baseline classification using a linear SVM to compare classification accuracies. While the best accuracy obtained with raw images

[1]Source code is available in http://svnext.it-sudparis.eu/svnview2-eph/ref_syst/Iris_Osiris_v4.1/

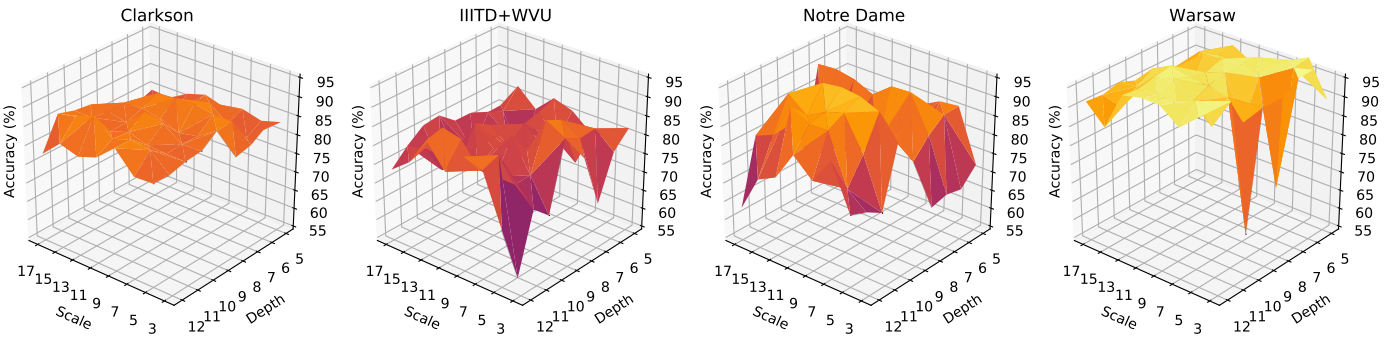



Fig. 3: Performance comparison of the proposed CNN operating on BSIF inputs of varying scale and depth. Accuracies are estimated over the *test unknown* partition of each dataset.

was below 80%, some of the BSIF filters achieved performance superior to 95%. Besides, the mean accuracy of BSIF images was higher, confirming the potential benefit of some of the filters.

With the intention of better understanding the results of BSIF images as input features, we performed an ablation analysis on the effect of scale and depth of the BSIF filters on the final classification. A CNN classifier (described in Section III-B) was trained on different input images, generated by BSIF filtering, first varying filter sizes, and then their depths. Results of this analysis are shown in Figure 3. It is possible to observe that some particular scales or depths have some (positive or negative) impact on the classification accuracy, but it is not possible to infer a generalizable trend.

In the same vein, we also tried to vary the size of the filter in the first convolutional layer of the CNN, while training it to perform classification on original images, and compared these results with BSIF (Fig. 4). In this case, the BSIF classification accuracy was consistently higher, presenting also a smaller variability in the results than the Raw input image. These results suggest a better capability of BSIF filters to capture discriminating features of attack images. Along with the BSIF scale/depth analysis, results indicate there is a potential gain when using some BSIF filters, and this encouraged us to develop our fusion strategy explained ahead.

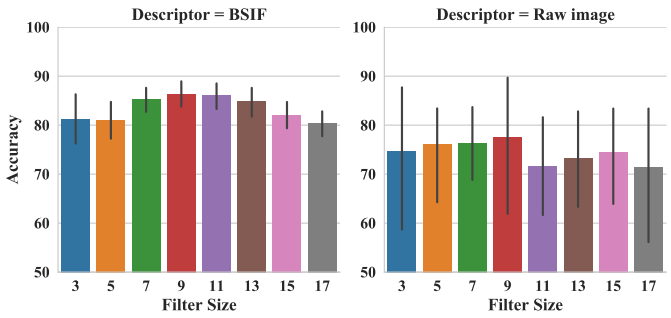



Fig. 4: Performance comparison between BSIF and raw image as the input to the CNN. Accuracies are estimated over the *test unknown* partition of each dataset. The length of the whiskers equals to one standard deviation of the obtained results.

### B. Classification using lightweight CNNs

As already mentioned, we train a CNN model for each of 60 BSIF transformations of the original image and a CNN for the original image, for a total of 61 CNNs. As we will show in our experiments, the ensemble of the most complementary CNN models allows us to exploit different aspects of the iris image texture, which is important to achieve high detection rates, especially in challenging scenarios, such as cross-sensor and cross-dataset evaluation protocols.

Given that we do not have large training datasets, and also to avoid possible overfitting to specific datasets through the use of millions of parameters, we opt to use a not-too-deep network architecture. Experiments with deeper configurations did not translate into improved results in our case (Fig. 5). Inspired by Menotti *et al.* [26], our CNN architecture comprises two convolutional layers, and two fully-connected layers. Batch normalization [49] is applied after each convolutional layer, to optimize the training procedure. The input of the network is an image of configurable size. The first convolutional layer consists of 16 filters of size $3 \times 3$ and Rectified Linear Unit (*ReLU*) activation. Next, MaxPooling on a $9 \times 9$ pixel window and stride 2 is applied, just before the first batch normalization. The second convolutional layer comprises 32 filters of size $3 \times 3$ and *ReLU* activation. MaxPooling on a $9 \times 9$ area with stride 8 and batch normalization follow this layer. The output of convolutional layers is fed into a fully-connected layer composed of 1,024 *ReLU* activated neurons. Finally, the output layer is formed by two units, corresponding to two classes we want to recognize (authentic iris vs. presentation attack iris) following the SoftMax transformation.

### C. Ensemble Fusion of Multiple Views

In this section, we present the proposed approaches for fusing the result of the ensemble of CNNs.

*1) Random Forest Fusion:* Random Forests is a well-known ensemble method that has been used for classifier fusion [50] as well as for variable importance ranking [51], [52]. It is composed of multiple decision trees built on different random subsets of the training data. Some of these decision trees tend to grow deep, sometimes leading to overfitting. This effect can be reduced by balancing the output of the decision trees and through tree pruning. By using a random forest classifier on

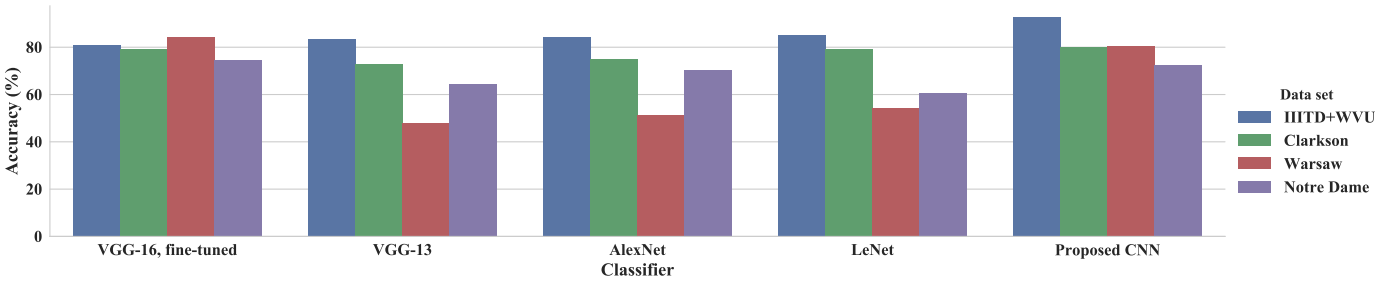



Fig. 5: Performance of CNN architectures on the *test unknown* partition, fine-tuned or trained from scratch in the *train* partition of each dataset. Our proposed architecture has fewer layers, takes less time to train, and results in similar or superior accuracy to other known CNN architectures.

top of the results of the different multi-view-CNN predictors, we can consolidate their outputs into a single prediction measure. As byproduct of this fusion, the random forest is also able to provide a ranking of predictor's importance, which can help us to select the most relevant predictors for our task.

Random Forests can estimate the test error (generalization error) of the ensemble, without needing to keep a separate test data partition. This is called out-of-bag (OOB) error estimation: the prediction error is calculated on all the samples that are left out of the bootstrap for each of the decision trees. During the process of training a Random Forest, after each tree is constructed, OOB error rate (of all variables) is compared to the classification error of the permutation of out-of-bag examples for each of the variables. As a result, we get an estimate of how much the misclassification error increases if each of the variables is disturbed. This measure is called variable importance.

*2) Voting Fusion:* Another strategy we exploit for decision-making is fusion through voting. We considered *majority* and *weighted voting*. In its simplest form, majority voting [50], [53] considers all 61 CNNs as inputs to decide whether or not a given sample should be considered as an authentic or an attack. The Condorcet Jury Theorem states that if all classifiers produce independent predictions, and if each has a probability of correct prediction that is greater than 0.5, the addition of more voters will increase the probability that the consensus will make the correct decision [50]. Veering away from simple majority voting, in some cases it is desirable to give greater weights to classifiers that are more likely to yield the right decision. With this in mind, we also use a strategy called *Best-Worst Weighted Vote* (*BWWV*) [54]. The best and the worst ranking predictors are identified and receive maximum and minimum weight (1 and 0, respectively). All remaining predictor weights are linearly cast between these extremes.

We consider two ways to attribute weights to CNN predictors: classification accuracy and rank importance. For the former, predictors are sorted by decreasing accuracy for the weight assignment. For the latter, predictors are sorted by their rank importance calculated through the Random Forest classifier. We refer to these techniques as `BWWVA` and `BWWVI`, respectively.

*3) Classifier selection and Meta-fusion:* One would naturally expect that some predictors provide complementary

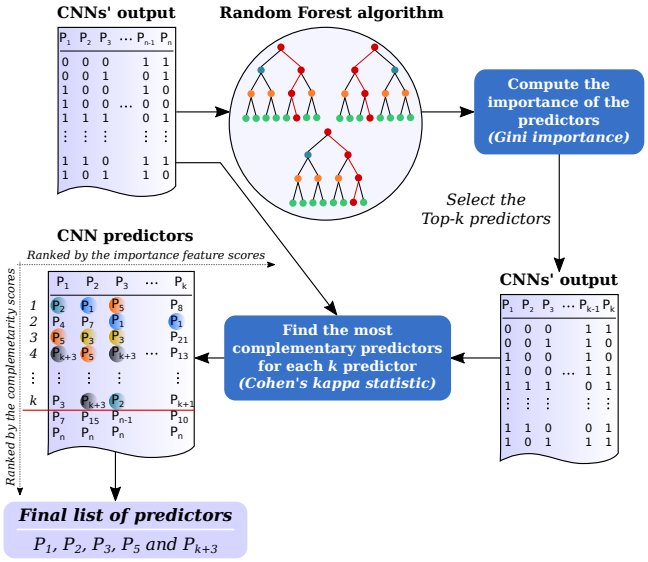



Fig. 6: Proposed algorithm for selecting predictors: after generating CNN outputs, a Random Forest is trained on the binary results of all predictors (feature vector $v \in R^{61}$ dimensions), to rank predictors by importance. The most complementary to each of the $k$ most important predictors are calculated through Cohen's kappa statistic. Finally, a predictor list is generated by selecting those which appear in two or more columns in the calculated matrix.

views on the problem while other predictors are highly correlated. In this section, we present our proposed algorithm for selecting the most relevant subset of predictors to use as the final ensemble for the PAD. We take advantage of Gini importance [55], measured with tree-based models, and inter-rater agreement, measured using Cohen's kappa statistic [56].

Gini importance should give us the most promising predictors in terms of classification accuracy, while the inter-rater agreement measure should give us the most complementary predictors. Fig. 6 illustrates the proposed approach.

Our algorithm performs a score analysis of all multi-view-CNN predictors' outputs using the Random Forest algorithm to find the $k$ most important predictors. This step focuses on which models are important in the classification task in terms of their Gini importance, also known as mean decrease in

impurity (MDI) [55]. In the context of our meta-analysis, each node of a tree represents a single feature, which is the output of a given predictor.

Gini importance measures how much the predictors decrease the weighted impurity in a tree. More precisely, the RF algorithm is a collection of decision trees aiming at splitting the data into two branches so that similar patterns end up in the same branch. The RF algorithm computes, during the training process, how much each predictor decreases the weighted impurity in a tree, and this metric is used to measure the quality of such splits. Our proposed algorithm takes advantage of that by ranking the multi-view-CNN predictors according to this measure and selecting the top-$k$ predictors, which will be our initial pool of good candidates for the fusion step.

After finding the $k$ most important models, we compute, through the Cohen's kappa statistic, which gives an agreement estimation between two predictors, their $l$ most complementary predictors, ending up with a $k \times l$ matrix $\mathbf{C}$ of predictors. We compute a final list of predictors by selecting those that appear in two or more columns of $\mathbf{C}$. The final decision-making is accomplished through SVM meta-fusion, which takes $k$ selected predictor outcomes as the input.

## IV. Experimental Results

In this section, we describe the datasets, validation protocols and experimental results.

### A. Datasets

Our work was performed on datasets made available in the Iris Liveness Detection Competition 2017 [7]. There is one set from each of four universities involved — Clarkson University, Warsaw University of Technology, IIITD-WVU and University of Notre Dame.

The Clarkson dataset [7], [44], [45] contains images of live irises, textured contact lenses, and iris printouts. The Warsaw dataset [7] comprises images of live irises and iris printouts. The Notre Dame dataset [57] contains images of live irises without textured contact lenses and of irises wearing textured contact lenses. Finally, the IIITD-WVU dataset [58]–[61] contains images of live irises, textured contact lenses, iris printouts, and printouts of textured contact lenses. Table I summarizes the composition of the datasets. For some of our experiments, we also formed a merged dataset by combining the four sets, which we will refer to as the "Combined" dataset. The original cross-validation partitioning was kept the same in the "Combined" dataset.

Each dataset is composed of a *train* partition, made available to the participants to facilitate training their algorithms, and a *test* partition, not distributed before the competition was ended, and used by the organizers to evaluate the submissions. LivDet-Iris 2017 co-organizers marked their test samples to form two groups of images. In the first group, referred to as *test known*, both live images and images of artifacts had the same "known" properties as train samples. The images belonging to a second group, *test unknown*, have different, or "unknown", properties than pictures included in the train subsets. Competition organizers applied different strategies when producing the *test unknown* samples. Clarkson University included visible-light image printouts and new patterns of textured contact lenses, Warsaw University of Technology used different equipment to prepare and photograph iris printouts. The images of patterned contact lenses offered by University of Notre Dame are of different brands than those in the train set. Finally, the whole test partition of the IIITD-WVU benchmark is considered as *test unknown*, since it was collected by a different institution (WVU) than the train set (IIITD), by a different sensor, and included outdoor acquisitions. Figure 7 shows sample images of each dataset.

To keep our evaluation protocol compliant with LivDet-Iris 2017, training of our methods uses solely pre-defined training partitions of the datasets, while the final performance is estimated both on the *test known* and *test unknown* partitions.

### B. Experimental Protocol

Our experiments followed the LivDet-Iris 2017 competition protocol [47], using the same datasets and train/test partitions as described in Section IV-A. A sub-partition consisting of a randomly selected 20% of the images from each train set was created to serve as a validation set during training. The system is trained to perform binary classification: authentic iris images should be labeled as *live*, while attack images (textured contact lenses, printouts of live images or printouts of textured contact lenses) should be labeled as *attack*.

We measure the performance of classifiers using four metrics:

- *Accuracy*, which is the ratio between the number of correctly classified images and the total number of images classified,
- *Bona-Fide Presentation Classification Error Rate* (BPCER), which is the proportion of *live* images that were incorrectly classified as *attacks*,
- *Attack Presentation Classification Error Rate* (APCER), which is the proportion of *attack* images incorrectly classified as *live* samples, and
- *Half Total Error Rate* (HTER), which corresponds to the average of BPCER and APCER.

BPCER and APCER error rates were defined by ISO/IEC 30107-3 [43] and adopted in LivDet-Iris 2017 for evaluation of submissions. The *accuracy* and HTER are used in the training stage for ranking the obtained solutions.

The CNN classifiers output a liveness score in the range of $\langle 0, 1 \rangle$ and the decision threshold was 0.5, as defined in the LivDet-Iris 2017 protocol. The source-code to reproduce all of our experiments is available to researchers [2].

### C. Evaluation of BSIF Representations and CNN-based Detectors

The first part of our work was to train and evaluate 61 lightweight CNNs, which act as primary predictors in our system. As it could be anticipated, individual CNN predictors result in a very good accuracy on the *train* and *test known* partitions, but they do not generalize well to *test unknown*

[2] Available on GitHub: https://github.com/akuehlka/mvlc_ipad.

TABLE I: Composition of the Datasets

| Dataset | Train | | | Test known | | | Test unknown | | |
|---|---|---|---|---|---|---|---|---|---|
| | Live | Contacts | Printouts | Live | Contacts | Printouts | Live | Contacts | Printouts |
| Clarkson | 2,469 | 1,122 | 1,346 | 1,485 | 765 | 908 | 638 | 494 | 144 |
| IIITD-WVU | 2,250 | 1,000 | 3,000 | – | – | – | 702 | 701 | 2,806 |
| Notre Dame | 600 | 600 | – | 900 | 900 | – | 900 | 900 | – |
| Warsaw | 1,844 | – | 2,669 | 974 | – | 2,016 | 2,350 | – | 2,160 |
| Combined | 7,163 | 2,722 | 7,015 | 3,359 | 1,665 | 2,924 | 4,590 | 2,095 | 5,110 |

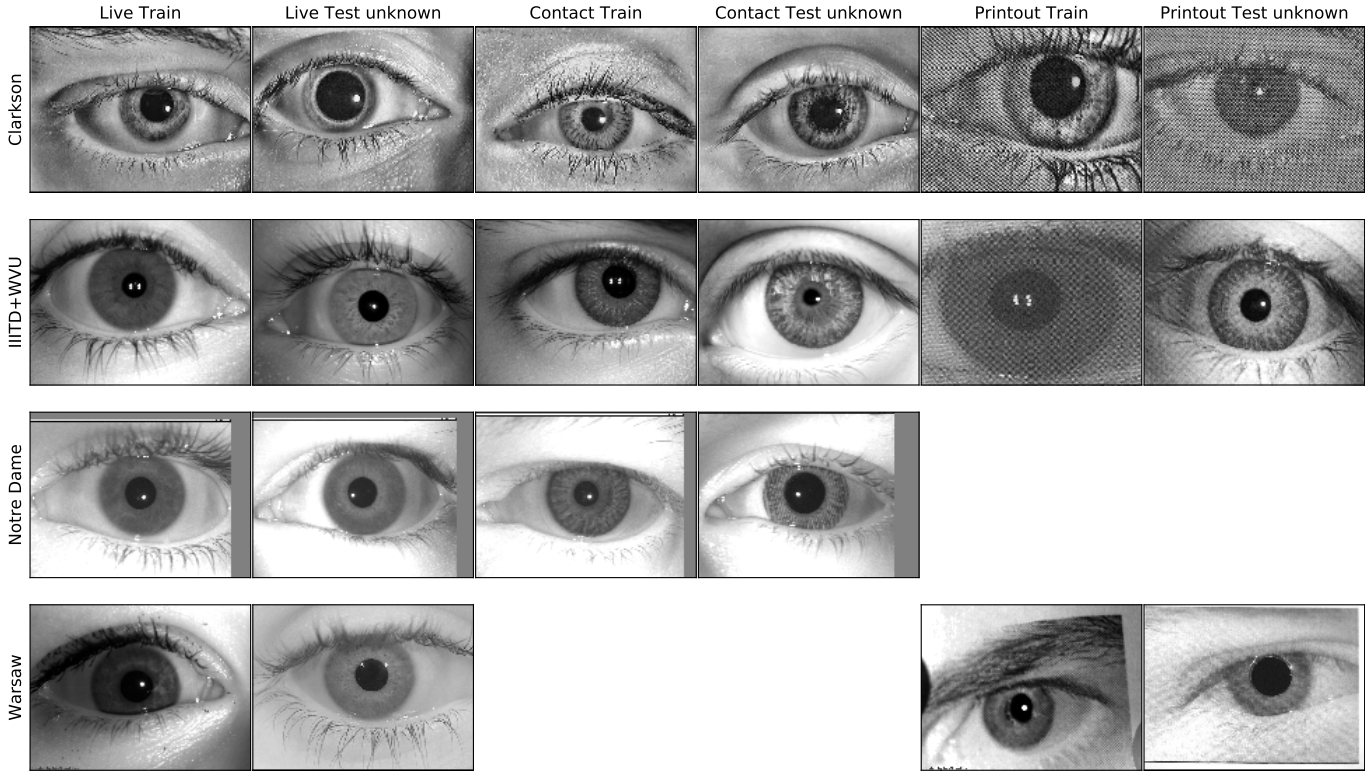



Fig. 7: Image samples from all datasets, from the *train* and *unknown* partitions. The difference between train and unknown images, especially in the case of attacks, illustrates situations where classifiers commonly fail.

data. This can be seen in Fig. 8, which shows the distribution of accuracies obtained by all 61 CNN-based predictors.

Since the CNN is different for different feature descriptors, one can assume the variation in performance is caused by the ability of such descriptors to capture particular textures of the images. This may suggest that some filters are particularly good at classifying a certain type of attack (a certain textured lens brand or a type of a printout), but they are inadequate for others. The proposed method identifies good predictors and aggregates them in order to create a more robust cross-dataset PAD system.

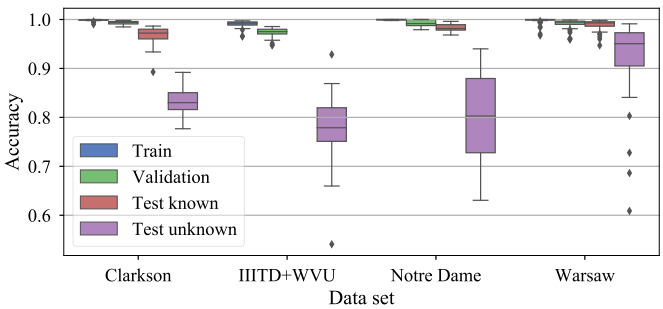



Fig. 8: Distribution of classification accuracies for individual CNN predictors, on each dataset and partition.

Table II shows the top three CNN classifiers on each data set. The size of the BSIF filters that performed best in each data set may offer some insight about the types of features that are being used for classification. While CNN predictors achieved very high accuracy on Warsaw dataset using BSIF filters of size ranging from 5 to 7, the best performing filters on Notre Dame dataset are of size ranging from 3 to 5, while for Clarkson data the best filters are larger, and their size ranges from 7 to 15. The range of BSIF filter sizes that performed best in IIITD+WVU goes from 5 through 17, which is consistent with the wide diversity of images found in this dataset.

It is important to observe a basic difference in the datasets: while Warsaw attack images are printouts, Notre Dame's

attack images are textured contact lenses. On the other hand, Clarkson and IIITD+WVU attack subsets are composed of a mix of textured contacts and printouts. Another fact that should be noted is that these data sets do not have similar proportions regarding the number of samples for each class in each partition. As an example, while there are only 144 printouts ($\sim 11\%$) in Clarkson unknown, IIITD+WVU has 2,806 ($\sim 66\%$) printouts in the same partition. These differences in data set composition may help to explain how different BSIF filter sizes achieve different ranges of performance for each dataset.

### D. Fusion Evaluation

Before deciding that a more elaborate method for fusion would be required, we experimented with basic methods. Table III summarizes results of four basic fusion methods: Random Forest (RF), Majority Voting (MV), Best-to-worst Weighted Voting by Accuracy (BWWVA), and by Importance (BWWVI). While there is no clear trend towards a specific fusion method, all of them seem to perform best in specific datasets and partitions. With regard to the data sets, the same trend seen in the CNN predictors happens here: Warsaw had the highest accuracy, followed by IIITD+WVU, Clarkson, and Notre Dame. However, in some cases simple fusion led to an obvious gain in accuracy that was not always the case: some fusion methods were not able to obtain a better result than the best individual classifier, typically in the unknown test partitions.

RF fusion was the most successful method, outperforming the others in three situations. With the exception of BWWVA, which tied with RF in the unknown partition of Warsaw, all other fusion methods were best in a single dataset/partition.

These somehow contradictory results motivated us to further investigate this aspect. Fig. 9 shows the relationship between the number of predictors and the output accuracy, for each weighted voting fusion technique and for each data set. For each fusion method and dataset, after the optimal number of predictors is reached, the addition of extra predictors does not improve the results, and in some cases it may even be detrimental. In some cases, the optimal accuracy is reached using a single predictor, but other cases may require up to 57 predictors.

TABLE II: Classification accuracy (%) of the top 3 predictors for each dataset on the *Test unknown* partition.

| Dataset | BSIF Filter | Accuracy | HTER |
|---|---|---|---|
| Clarkson | 15x15x11 | 87.08% | 12.91% |
| | 07x07x10 | 82.69% | 17.29% |
| | 13x13x12 | 80.74% | 19.25% |
| IIITD+WVU | 05x05x07 | 81.87% | 47.45% |
| | 09x09x11 | 78.69% | 27.03% |
| | 17x17x07 | 77.62% | 19.41% |
| Notre Dame | 05x05x11 | 71.78% | 28.22% |
| | 05x05x10 | 69.11% | 30.89% |
| | 03x03x06 | 67.83% | 32.17% |
| Warsaw | 05x05x11 | 98.00% | 2.08% |
| | 07x07x12 | 93.84% | 6.24% |
| | 05x05x08 | 91.95% | 8.26% |

A method for predictor selection is necessary so that we can obtain the best accuracy from the ensemble. However, this process is not trivial: the analysis of the predictor-accuracy relation in other data partitions (train and validation) reveals significantly different trends. Consequently, it is not always possible to determine the optimal number of predictors to be used in the unknown partition by simply using information from other partitions.

Accuracy is the obvious metric of choice for selecting the best base classifiers, but it does not tell us about the complementarity relations between different classifiers. The *Gini* importance measure calculated by training a Random Forest can suggest the best classifiers, based on how much each predictor helps to reduce the impurity in a decision tree. Therefore, it can also be used as a criterion for predictor selection. However, neither of these estimations helps us to determine an optimal number of predictors. Furthermore, our results show there is no distinct advantage in using either accuracy or importance. These results motivated us to search for a method for selection of predictors that is based on their already known properties, but also based on how different base classifiers integrate with each other.

### E. Cross-Domain Evaluation

With the exception of Clarkson, all LivDet-Iris 2017 datasets were designed to allow cross-sensor evaluation. Images in their unknown partitions were captured with different sensors and environments. Additionally, we conducted cross-dataset experiments to verify the possibility of transfer learning across these datasets. However, direct cross-dataset evaluation using simple fusion did not result in good accuracy. In fact, training CNN predictors in one data set and testing them in another resulted in accuracy no better than random prediction.

One of the premises of our approach is the use of multiple views of the data, in order to capture texture subtleties. We can confirm this if we consider the ranges of BSIF filter sizes that had better accuracy in the different data sets. This suggests there is enough difference in texture scale from one data set to another to cause the individual CNN predictors not to be able to recognize them. The variation of these data sets in nature (texture contacts, printouts, or a mix of both), composition (regarding sizes of partitions and classes), acquisition devices and environments can explain the limited ability for transfer learning here.

Despite the fact that direct cross-dataset evaluation was not successful, training CNN predictors in a combined dataset produced much better results. Table IV shows a comparison of all predictors and their fusion in the combined dataset. Both individual CNNs and simple fusion methods achieved typically over $95\%$ classification accuracy in the known test partition. In the unknown partition, while most CNN predictors achieved HTER of $12\%$ or lower, all simple fusion methods obtained HTER below $8\%$, with classification accuracies higher than $91\%$ even on the unknown set.

Fig. 10 shows the distribution of HTER for fusion methods, in a comparison between training on each individual dataset, and training on the combined dataset. Performing the training on the combined dataset was particularly favorable in

TABLE III: Results for four simple fusion strategies on the *test known* (K) and *test unknown* (U) partitions; best HTERs in bold.

| | | RF | | | | MV | | | | BWWVA | | | | BWWVI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | APCER | BPCER | HTER | Accuracy | APCER | BPCER | HTER | Accuracy | APCER | BPCER | HTER | Accuracy | APCER | BPCER | HTER |
| Clarkson | K | 97.74 | 4.45 | 0.74 | 2.59 | 99.48 | 0.00 | 0.87 | **0.44** | 99.44 | 0.00 | 0.94 | 0.47 | 99.40 | 0.09 | 0.94 | 0.52 |
| | U | 78.70 | 41.94 | 0.63 | 21.28 | 85.59 | 28.17 | 0.63 | 14.40 | 86.37 | 26.60 | 0.63 | **13.61** | 85.75 | 27.70 | 0.78 | 14.24 |
| IIITD+WVU | K | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | U | 83.34 | 15.94 | 20.23 | **18.08** | 83.82 | 12.83 | 32.91 | 22.87 | 84.25 | 12.29 | 33.05 | 22.67 | 83.06 | 9.21 | 55.56 | 32.38 |
| Notre Dame | K | 99.44 | 0.44 | 0.67 | **0.56** | 99.33 | 0.22 | 1.11 | 0.67 | 99.39 | 0.33 | 0.89 | 0.61 | 99.27 | 0.44 | 1.00 | 0.72 |
| | U | 85.11 | 29.33 | 0.44 | **14.89** | 82.39 | 34.22 | 1.00 | 17.61 | 81.61 | 35.78 | 1.00 | 18.39 | 82.28 | 34.44 | 1.00 | 17.72 |
| Warsaw | K | 99.80 | 0.05 | 0.51 | 0.28 | 99.87 | 0.04 | 0.31 | 0.18 | 99.87 | 0.05 | 0.31 | 0.18 | 99.90 | 0.00 | 0.30 | **0.15** |
| | U | 99.51 | 0.32 | 0.64 | **0.48** | 99.49 | 0.37 | 0.64 | 0.50 | 99.51 | 0.37 | 0.59 | **0.48** | 99.46 | 0.23 | 0.81 | 0.52 |

TABLE IV: HTER (%) on the Combined dataset.

| | No Fusion (avg.) | RF | MV | BWWVI | BWWVA |
|---|---|---|---|---|---|
| K | 2.57 | 0.74 | 0.65 | 0.59 | 0.63 |
| U | 12.47 | 7.02 | 6.89 | 7.12 | 6.88 |

IIITD+VWU dataset, in which HTER was reduced from more than 25% to nearly 10%.

### *F. Evaluation of the Proposed Meta-Analysis: Selection and Meta-Fusion of Classifiers*

In this section, we evaluate the proposed method designed to automatically select the most relevant multi-view-CNN predictors, in terms of their importance and complementarity, and to fuse their individual results via meta-fusion approach, which was performed using the Support Vector Machine.

First, we used our proposed algorithm described in Section III-C3 to select the most relevant predictors from a pool of 61 predictors. Next, we performed a meta-fusion using the SVM with a radial basis function kernel. Both the parameters of the SVM and the parameter $k$ (see Sec. III-C3) were found through grid search in the aggregated training sets of all datasets considered in this work. The *known* and *unknown* test sets were used only to report the final performance results.

We performed a fine grid-search of parameter $k$ on the neighborhood of (5, 20), and the best HTER was achieved with $k = 16$. According to one-sample Wilcoxon signed-rank test, the observed differences in the HTER values obtained for each $k$ are statistically significant at the significance level $\alpha = 0.05$ ($p\text{-}value = 0.0004$). Henceforth, we use $k = 16$ to report performance results of our proposed method.

Furthermore, Table V shows the effectiveness of our meta-fusion approach compared to the best multi-view-CNN predictor and the majority vote fusion technique. In addition to the results for *test known* and *test unknown* partitions, Table V also presents *overall test*, which corresponds to the accuracy obtained on both test partitions combined. These results are in agreement with those reported in the literature [50], which states that both accuracy and complementarity are the foundation for effective fusion.

### *G. Comparison with the State of the Art*

In this section, we compare our results with the LivDet-Iris 2017 best performing method. As we can see, our method outperforms the competition winner. Table VI summarizes the results of the competition in contrast to our best solution developed solely on the LivDet-Iris 2017 train partitions.

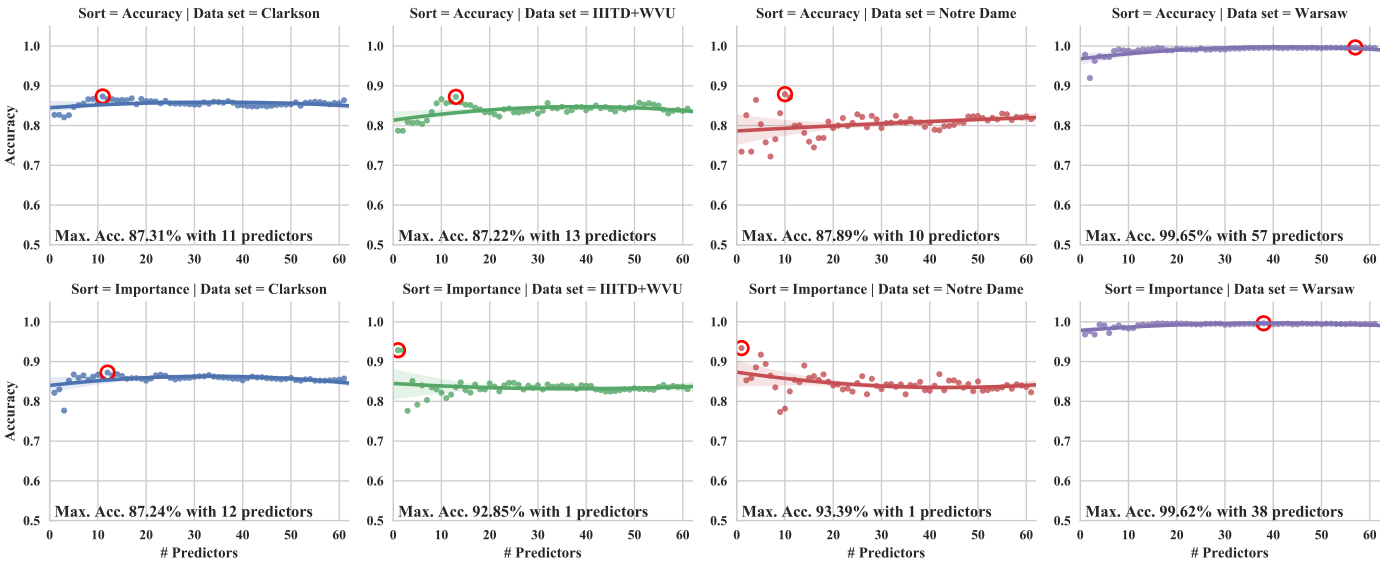



Fig. 9: Relationship between classification accuracy and number of predictors used in the fusion. The top row shows results for *BWWVA*, and the bottom row for *BWWVI*.

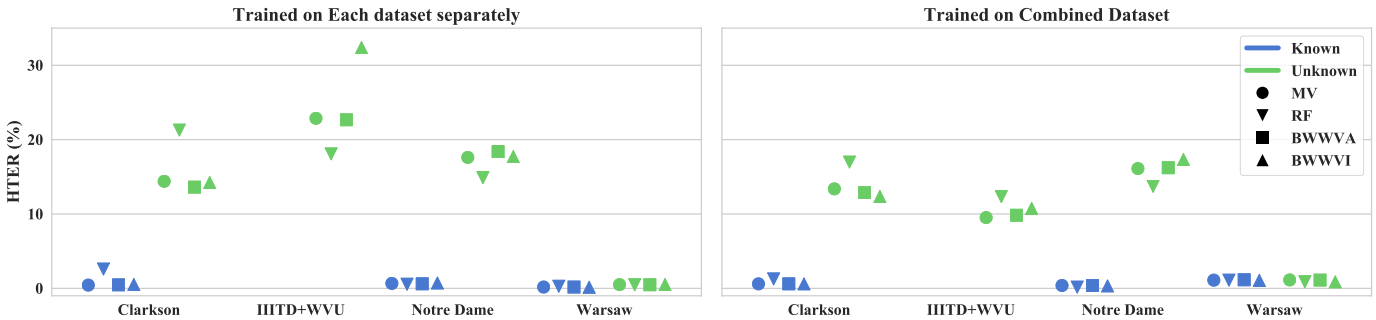



Fig. 10: HTER distribution for fusion methods. On the left, training and evaluation is performed on each individual dataset. On the right, training is performed on the combined dataset, and evaluation is performed on each individual dataset.

Even in the cases where our methods were not able to outperform one of the specific error rates, we were able to significantly reduce the combined error rate. This happens, for instance, in the overall performance on IIITD+WVU dataset: the best HTER among LivDet-Iris 2017 competitors is 16.7%. Although our methods did not outperform their BPCER of 3.99%, we managed to lower HTER to 7.9% on that specific dataset. Similar situations occurred on all datasets, showing a consistent improvement of our results over the state of the art.

TABLE VI: Comparison of HTER (%) with LivDet-Iris 2017, on the combined *test known* and *test unknown* partitions.

| Dataset | LivDet-Iris 2017 Winner | Proposed Meta-Fusion approach | Error Reduction (%) |
|---|---|---|---|
| Clarkson | 9.59 | 9.45 | 1.46 |
| IIITD+WVU | 16.70 | 14.92 | 10.66 |
| Notre Dame | 4.03 | 3.28 | 18.61 |
| Warsaw | 5.81 | 0.68 | 88.30 |
| Average | 9.03 | 7.08 | 21.59 |

Meta-Fusion results were further improved when applied to the combined dataset. Table VII presents these results in comparison with LivDet-Iris 2017, and also with our previous Meta-Fusion results. The overall HTER was reduced from 7% to 4% when Meta-Fusion was applied to the combined dataset. At this point, it is necessary to be clear about the LivDet-Iris 2017 comparison: the known and unknown HTER numbers presented in Table VII are estimated. Since BPCER is not available for all *known* and *unknown* test partitions in LivDet-Iris 2017, we assumed 0 (perfect score) in our estimation. Even with this optimistic assumption about the competition results, our method achieved an error reduction of more than 50% with regard to the former.

Our current implementation takes, on average, 0.028s to perform classification on a single image. Timing was performed on a 4-core Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz with 128GB of RAM, equipped with a GeForce GTX 1080 Ti GPU. This time includes the CNN classification of the 61 data views and the meta-fusion of their results. Since the LivDet-Iris protocol does not establish parameters for speed efficiency analysis, the main directive for our implementation herein was classification accuracy. Therefore our method's efficiency can certainly be optimized, if needed.

## V. Conclusions

In this paper, we addressed the iris Presentation Attack Detection problem, as defined by the LivDet-Iris 2017 Competition, using an approach based on an ensemble of multi-view learning detectors. Our method has advanced the state of the art in iris PAD and offered insight on the potential use of different BSIF filters to deal with different textures in the same domain.

TABLE V: Performance results (%) of the proposed method, majority vote, and the best individual CNN model for the *test known* (K), *test unknown* (U), and *overall test* (O) partitions from datasets used in this work. The classifiers were trained on the train partition of each dataset. † IIITD-WVU dataset contains only the *unknown test* partition.

| Dataset | Testing set | Meta-Fusion via SVM | | | BWWVA | | | RF | | | Best Individual CNN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | APCER | BPCER | HTER | APCER | BPCER | HTER | APCER | BPCER | HTER | APCER | BPCER | HTER |
| Notre Dame | K | 0.00 | 2.44 | 1.22 | 0.33 | 0.89 | 0.61 | 0.44 | 0.67 | 0.56 | 0.67 | 2.44 | 1.56 |
| | U | 9.22 | 1.44 | 5.33 | 35.78 | 1.00 | 18.39 | 29.33 | 0.44 | 14.89 | 32.44 | 2.44 | 17.44 |
| | O | 4.61 | 1.94 | 3.28 | 18.06 | 0.94 | 9.50 | 14.89 | 0.56 | 7.72 | 16.56 | 2.44 | 9.50 |
| Warsaw | K | 0.05 | 0.82 | 0.44 | 0.05 | 0.31 | 0.18 | 0.05 | 0.51 | 0.28 | 0.30 | 0.41 | 0.35 |
| | U | 0.09 | 1.49 | 0.79 | 0.37 | 0.60 | 0.48 | 0.32 | 0.64 | 0.48 | 20.97 | 0.68 | 10.83 |
| | O | 0.07 | 1.29 | 0.68 | 0.21 | 0.45 | 0.33 | 0.19 | 0.58 | 0.38 | 10.63 | 0.55 | 5.59 |
| Clarkson | K | 4.55 | 0.20 | 2.37 | 0.00 | 0.94 | 0.47 | 3.48 | 0.47 | 1.98 | 1.26 | 1.75 | 1.50 |
| | U | 41.54 | 0.31 | 20.92 | 26.30 | 0.63 | 13.47 | 39.12 | 0.63 | 19.88 | 32.71 | 1.88 | 17.29 |
| | O | 18.66 | 0.24 | 9.45 | 13.30 | 0.78 | 7.04 | 21.30 | 0.55 | 10.93 | 16.98 | 1.82 | 9.40 |
| IIITD+WVU † | U | 12.32 | 17.52 | 14.92 | 12.29 | 33.05 | 22.67 | 5.56 | 21.23 | 13.39 | 21.81 | 72.22 | 47.02 |
| | O | | | | | | | | | | | | |
| Average | O | **8.92** | **5.25** | **7.08** | 10.96 | 8.80 | 9.88 | 10.48 | 5.73 | 8.11 | 16.50 | 19.26 | 17.88 |

TABLE VII: Results in terms of HTER (%) for our Meta-Fusion approach trained on the Combined dataset. Error reduction (ER) values present the error decrease achieved by our approach with regard to LivDet-Iris 2017 winner.

| Dataset | LivDet-Iris 2017 Winner HTER | Proposed Meta-Fusion approach (%) | | | |
|---|---|---|---|---|---|
| | | Trained on each dataset | | Trained on Combined dataset | |
| | | HTER | ER | HTER | ER |
| K | 0.74[a] | 1.34 | -81.08 | 0.74 | 0.00 |
| U | 13.23[a] | 10.49 | 20.71 | 8.39 | 36.58 |
| O | 9.03 | 7.08 | 21.59 | 4.44 | 50.83 |

[a]Estimated

We presented a new approach combining two techniques for iris PAD: CNNs and Ensemble Learning. Extensive experimentation was conducted using the most challenging datasets publicly available. The experiments included cross-sensor and cross-dataset evaluations. Results show a varying ability for different BSIF+CNN representations to capture different aspects of the input images.

Simple fusion experiments show that although helpful, such techniques are not yet capable to provide optimal classification accuracy. In fact, we demonstrate how the continuous addition of classifiers to the fusion does not necessarily improve the classification performance. In that context, we presented a new method for selection of classifiers, based on the meta-analysis of their Gini importance and inter-classifier complementarity.

Our Meta-Fusion method was able to consistently outperform the LivDet-Iris 2017 competition winner, with an overall Reduction Error Rate of more than $21\%$. Specifically, the HTER in the Warsaw dataset was only $0.68\%$, corresponding to a reduction in error of more than 88% in relation to the top result reported in LivDet-Iris 2017. Although not as extreme as in Warsaw case, classification accuracy was also improved for other datasets, with reduction in error ranging from 1 to $19\%$. Experiments with the combined dataset showed an additional improvement of $37\%$ on the overall HTER.

As a suggestion for future work, an immediate alternative application for our method is face PAD. A significant portion of face recognition attacks are based on printouts, as we believe meta-fusion of multi-view-CNN predictors could be easily applied to it with good potential for success.

## VI. Acknowledgement

This research was partially supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) through grants BEX 12976/13-0 and DeepEyes as well as by the São Paulo Research Foundation (FAPESP) through the DéjàVu research project (Grant #2017/12646-3).

## References

[1] J. Daugman and C. Downing, "Epigenetic randomness, complexity and singularity of human iris patterns," *Proceedings of the Royal Society B: Biological Sciences*, vol. 268, no. 1477, pp. 1737–1740, 2001.

[2] K. W. Bowyer and P. J. Flynn, "Biometric identification of identical twins: A survey," in *IEEE Intl. Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sept 2016, pp. 1–8.

[3] Unique Identification Authority of India, Government of India, "Unique Identification Authority of India Website," https://uidai.gov.in/, 2018, [Online; accessed 02/12/2018].

[4] J. Daugman, "600 million citizens of India are now enrolled with biometric ID," in *SPIE Newsroom*, 2014, pp. 1–4.

[5] Canada Border Services Agency and U.S. Customs and Border Protection, "NEXUS," https://www.cbsa-asfc.gc.ca/prog/nexus/menu-eng.html, accessed January 19, 2018.

[6] ISO/IEC 30107-1:2016, "Information technology – Biometric presentation attack detection – Part 1: Framework."

[7] D. A. 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 2017 - Iris Liveness Detection Competition 2017," *Biometrics: Theory Applications and Systems (BTAS)*, pp. 0–5, 2017.

[8] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Intl. Conference on Pattern Recognition (ICPR)*, Nov 2012, pp. 1363–1366.

[9] J. Daugman, "Wavelet demodulation codes, statistical independence, and pattern recognition," in *Institute of Mathematics and its Applications, (IMA-IP)*, 2000, pp. 244–260.

[10] L. Thalheim, J. Krissler, and P.-M. Ziegler, "Biometric Access Protection Devices and their Programs Put to the Test, Available online in c't Magazine, No. 11/2002, p. 114," on-line, 2002.

[11] A. N. Al-Raisi and A. M. Al-Khouri, "Iris recognition and the challenge of homeland and border control security in uae," *Telematics and Informatics*, vol. 25, no. 2, pp. 117 – 132, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0736585306000360

[12] J. S. Doyle, P. J. Flynn, and K. W. Bowyer, "Automated classification of contact lens type in iris images," in *IEEE Int. Conference on Biometrics (ICB)*, June 2013, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICB.2013.6612954

[13] X. He, Y. Lu, and P. Shi, "A new fake iris detection method," in *IEEE Int. Conference on Biometrics (ICB)*, M. Tistarelli and M. S. Nixon, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1132–1139. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01793-3_114

[14] J. Zuo, N. A. Schmid, and X. Chen, "On generation and analysis of synthetic iris images," *IEEE Trans. Inf. Forens. Security*, vol. 2, no. 1, pp. 77–90, March 2007.

[15] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Human iris recognition in post-mortem subjects: Study and database," in *IEEE Int. Conference on Biometrics: Theory Applications and Systems (BTAS)*, Sept 2016, pp. 1–6.

[16] J. Komulainen, A. Hadid, and M. Pietikinen, "Generalized textured contact lens detection by extracting bsif description from cartesian iris images," in *IEEE Int. Joint Conference on Biometrics (IJCB)*, Sept 2014, pp. 1–7.

[17] Lovish, A. Nigam, B. Kumar, and P. Gupta, "Robust contact lens detection using local phase quantization and binary gabor pattern," in *Int. Conference on Computer Analysis of Images and Patterns (CAIP)*, G. Azzopardi and N. Petkov, Eds. Springer International Publishing, 2015, pp. 702–714.

[18] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva, "An investigation of local descriptors for biometric spoofing detection," *IEEE Trans. Inf. Forens. Security*, vol. 10, no. 4, pp. 849–863, Apr. 2015.

[19] A. F. Sequeira, S. Thavalengal, J. Ferryman, P. Corcoran, and J. S. Cardoso, "A realistic evaluation of iris presentation attack detection," in *Int. Conference on Telecommunications and Signal Processing (TSP)*, June 2016, pp. 660–664.

[20] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva, "Contact lens detection and classification in iris images through scale invariant descriptor," in *Int. Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, Nov 2014, pp. 560–565.

[21] F. Pala and B. Bhanu, "Iris liveness detection by relative distance comparisons," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[22] Z. Akhtar, C. Micheloni, C. Piciarelli, and G. L. Foresti, "Mobio_livdet: Mobile biometric liveness detection," in *IEEE Int. Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2014, pp. 187–192.

[23] R. Chen, X. Lin, and T. Ding, "Liveness detection for iris recognition using multispectral images," *Pattern Recognition Letters*, vol. 33, no. 12, pp. 1513 – 1519, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865512001262

[24] J. Galbally, M. Savvides, S. Venugopalan, and A. A. Ross, *Iris Image Reconstruction from Binary Templates*. London: Springer

London, 2016, pp. 469–496. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-6784-6_20

[25] P. Silva, E. Luz, R. Baeta, H. Pedrini, A. X. Falcão, and D. Menotti, "An approach to iris contact lens detection based on deep image representations," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, August 2015, pp. 157–164.

[26] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, April 2015.

[27] D. Gragnaniello, C. Sansone, G. Poggi, and L. Verdoliva, "Biometric spoofing detection by a domain-aware convolutional neural network," in *Int. Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, Nov 2016, pp. 193–198.

[28] L. He, H. Li, F. Liu, N. Liu, Z. Sun, and Z. He, "Multi-patch convolution neural network for iris liveness detection," in *IEEE Intl. Conference on Biometrics Theory, Applications and Systems (BTAS)*, September 2016, pp. 1–7.

[29] R. Raghavendra, K. B. Raja, and C. Busch, "Contlensnet: Robust iris contact lens detection using deep convolutional neural networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp. 1160–1167.

[30] S. J. Lee, K. R. Park, and J. Kim, "Robust fake iris detection based on variation of the reflectance ratio between the iris and the sclera," in *Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, September 2006, pp. 1–6.

[31] J. H. Park and M. G. Kang, "Multispectral iris authentication system against counterfeit attack using gradient-based image fusion," *Optical Engineering*, vol. 46, no. 11, pp. 117 003–117 003–14, 2007. [Online]. Available: http://dx.doi.org/10.1117/1.2802367

[32] S. Thavalengal, T. Nedelcu, P. Bigioi, and P. Corcoran, "Iris liveness detection for next generation smartphones," *IEEE Trans. Consumer Electronics*, vol. 62, no. 2, pp. 95–102, May 2016.

[33] A. Pacut and A. Czajka, "Aliveness detection for iris biometrics," in *IEEE Int. Carnahan Conference on Security Technology (ICCST)*, October 2006, pp. 122–129.

[34] E. C. Lee and K. R. Park, "Fake iris detection based on 3D structure of iris pattern," *Intl. Journal of Imaging Systems and Technology*, vol. 20, no. 2, pp. 162–166, 2010. [Online]. Available: http://dx.doi.org/10.1002/ima.20227

[35] J. Connell, N. Ratha, J. Gentile, and R. Bolle, "Fake iris detection using structured light," in *IEEE Int. Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8692–8696.

[36] K. Hughes and K. W. Bowyer, "Detection of contact-lens-based iris biometric spoofs using stereo imaging," in *Hawaii Intl. Conference on System Sciences*, January 2013, pp. 1763–1772.

[37] K. Raja, R. Raghavendra, and C. Busch, "Video presentation attack detection in visible spectrum iris recognition using magnified phase information," *IEEE Trans. Inf. Forens. Security*, vol. 10, no. 10, pp. 2048–2056, October 2015.

[38] I. Rigas and O. V. Komogortsev, "Eye movement-driven defense against iris print-attacks," *Pattern Recognition Letters*, vol. 68, Part 2, pp. 316 – 326, 2015, special Issue on Soft Biometrics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865515001737

[39] A. Czajka, "Pupil dynamics for iris liveness detection," *IEEE Trans. Inf. Forens. Security*, vol. 10, no. 4, pp. 726–735, April 2015.

[40] A. Czajka and K. W. Bowyer, "Presentation attack detection for iris recognition: An assessment of the state-of-the-art," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 86:1–86:35, Jul. 2018. [Online]. Available: http://doi.acm.org/10.1145/3232849

[41] D. Nguyen, T. Pham, Y. Lee, and K. Park, "Deep learning-based enhanced presentation attack detection for iris recognition by combining features from local and global regions based on nir camera sensor," *Sensors*, vol. 18, no. 8, p. 2601, Aug 2018. [Online]. Available: http://dx.doi.org/10.3390/s18082601

[42] P. Kontschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo, "Deep neural decision forests," in *IEEE Int. Conference on Computer Vision (ICCV)*, 2015, pp. 1467–1475.

[43] ISO/IEC FDIS 30107-3:2017, "Information technology – Biometric presentation attack detection – Part 3: Testing and reporting."

[44] D. Yambay, J. S. Doyle, K. W. Bowyer, A. Czajka, and S. Schuckers, "Livdet-iris 2013 - iris liveness detection competition 2013," *IEEE Int. Joint Conference on Biometrics (IJCB)*, pp. 1–8, 2014.

[45] D. Yambay, B. Walczak, S. Schuckers, and A. Czajka, "Livdet-iris 2015 - iris liveness detection competition 2015," *IEEE Intl. Conference on Identity, Security and Behavior Analysis (ISBA)*, pp. 1–6, 2017.

[46] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Pattern Recognition (ICPR), 2012 21st Intl. Conference on*. IEEE, 2012, pp. 1363–1366.

[47] "LivDet-Iris 2017 – Iris Liveness Detection Competition," Website, 2017, last access: 02/21/2018. [Online]. Available: http://iris2017.livdet.org/

[48] N. Othman, B. Dorizzi, and S. Garcia-Salicetti, "OSIRIS: An open source iris recognition software," *Pattern Recognition Letters*, vol. 82, no. P2, pp. 124–131, Oct. 2016. [Online]. Available: https://doi.org/10.1016/j.patrec.2015.09.002

[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Intl. Conference on Machine Learning*, 2015, pp. 448–456.

[50] L. I. Kuncheva, "Combining label outputs," in *Combining Pattern Classifiers*. John Wiley & Sons, Inc., 2014, pp. 111–142. [Online]. Available: http://dx.doi.org/10.1002/9781118914564.ch4

[51] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[52] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[53] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[54] F. Moreno-Seco, J. M. Iñesta, P. J. P. de León, and L. Micó, "Comparison of classifier fusion methods for classification in pattern recognition tasks," in *Structural, Syntactic, and Statistical Pattern Recognition*, D.-Y. Yeung, J. T. Kwok, A. Fred, F. Roli, and D. de Ridder, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 705–713.

[55] L. Breiman, *Classification and regression trees*, ser. Wadsworth statistics/probability series. Wadsworth International Group, 1984.

[56] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, no. 3, pp. 613–619, 1973.

[57] J. S. Doyle and K. W. Bowyer, "Robust detection of textured contact lenses in iris recognition using BSIF," *IEEE Access*, vol. 3, pp. 1672–1683, 2015.

[58] N. Kohli, D. Yadav, M. Vatsa, and R. Singh, "Revisiting iris recognition with color cosmetic contact lenses," *Intl. Conference on Biometrics (ICB)*, pp. 1–7, 2013.

[59] P. Gupta, S. Behera, M. Vatsa, and R. Singh, "On iris spoofing using print attack," *Pattern Recognition (ICPR), 2014 22nd Intl. Conference on*, pp. 1681–1686, 2014.

[60] D. Yadav, N. Kohli, J. S. Doyle, R. Singh, M. Vatsa, and K. W. Bowyer, "Unraveling the effect of textured contact lenses on iris recognition," *IEEE Trans. Inf. Forens. Security*, vol. 9, pp. 851–862, 2014.

[61] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore, "Detecting medley of iris spoofing attacks using desist," *IEEE Int. Conference on Biometrics: Theory Applications and Systems (BTAS)*, pp. 1–6, 2016.