Interpretable Biometrics: Should We Rethink How Presentation Attack Detection is Evaluated?

Ana F. Sequeira
INESC TEC
Porto, Portugal
ana.f.sequeira@inesctec.pt

João Ribeiro Pinto
INESC TEC & Universidade do Porto
Porto, Portugal
joao.t.pinto@inesctec.pt

Wilson Silva
INESC TEC & Universidade do Porto
Porto, Portugal
wilson.j.silva@inesctec.pt

Tiago Gonçalves
INESC TEC
Porto, Portugal
tiago.f.goncalves@inesctec.pt

Jaime S. Cardoso
INESC TEC & Universidade do Porto
Porto, Portugal
jaime.cardoso@inesctec.pt

***Abstract*—Presentation attack detection (PAD) methods are commonly evaluated using metrics based on the predicted labels. This is a limitation, especially for more elusive methods based on deep learning which can freely learn the most suitable features. Though often being more accurate, these models operate as complex black boxes which makes the inner processes that sustain their predictions still baffling. Interpretability tools are now being used to delve deeper into the operation of machine learning methods, especially artificial networks, to better understand how they reach their decisions. In this paper, we make a case for the integration of interpretability tools in the evaluation of PAD. A simple model for face PAD, based on convolutional neural networks, was implemented and evaluated using both traditional metrics (APCER, BPCER and EER) and interpretability tools (Grad-CAM), using data from the ROSE Youtu video collection. The results show that interpretability tools can capture more completely the intricate behavior of the implemented model, and enable the identification of certain properties that should be verified by a PAD method that is robust, coherent, meaningful, and can adequately generalize to unseen data and attacks. One can conclude that, with further efforts devoted towards higher objectivity in interpretability, this can be the key to obtain deeper and more thorough PAD performance evaluation setups.**

I. Introduction

Machine learning (ML) based systems are excelling in most of the artificial intelligence (AI) fields, outperforming other methods as well as humans. Undoubtedly, the success of AI systems is mainly due to: improvements in deep learning (DL) methodology, availability of large databases, and computational gains obtained with powerful GPU cards. [1], [2]

However, some challenges related to AI remain, especially the lack of transparency of the algorithms [3]–[5]. After the euphoria around artificial neural networks and their over-performing accuracy rates, the research community is starting to understand the importance of being made accountable for what are these outstanding models in fact learning and deciding upon. The fact is that most of the AI methods still operate as complex black boxes and the inner processes which sustain their predictions are still unattainable, making such methods untrustworthy.

One very illustrative situation was described by Lapuschkin *et al*. [6]. In this work, the authors observed that for the detection of the class "horse", by a deep neural network, the model assigned relevance to the bottom left corner of the images, where a careful inspection revealed the presence of a copyright tag. This is an example of how a ML model can make a correct decision based on parts of the image that are arbitrary for the task at hand and overly specific to the data seen during training. As a results, efforts have been devoted across several pattern recognition research topics to develop models that are more interpretable or to use interpretations to improve the design of the decision algorithms. [6], [7]

In particular, biometric recognition (BR) systems and anti-spoofing techniques may be positively impacted by the use of interpretability. Most BR systems can be spoofed by presenting fake or altered samples of the biometric trait at the sensor by an intruder that is trying to mischievously access the system. Presentation attack detection (PAD) methods (such as liveness detection techniques and tamper detection methods) are intended to detect spoofing attacks. When designing a PAD method, it can be very rewarding to know more about the rationale behind its predictions instead of relying on the output of a black box and risking putting high stakes decisions on the hands of models that behave like the aforementioned one. Many advantages will come from studying what a model learns and which information it uses to decide about a threat.

As in several other pattern recognition tasks, the use of DL-based PAD methods is increasingly common [8], [9]. This creates a pressing need for the use of interpretability since the involved processes become more elusive as the artificial neural networks become deeper. Moreover, traditional evaluation setups, which adequacy regarding the robustness to unseen attacks has been questioned [10]–[12], may not be able

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", and within the PhD grants "SFRH/BD/137720/2018" and "SFRH/BD/139468/2018".

to see in enough depth to capture the model's behavior.

Traditional metrics quantify the performance solely relying on predicted labels, without looking into the information used to reach these predictions. This assessment is quite limited especially for DL-based approaches. While using handcrafted features, one has more control over the information used by the method to perform a decision. However, using DL approaches, the method is free to learn the most adequate features to discriminate classes using the training data, which can translate into the use of largely meaningless or overly specific features that would hardly generalize.

Considering the aforementioned limitations of the current way to assess performance in PAD, it is argued in this work that the evaluation frameworks need to be reformulated to become more thorough and meaningful. Interpretability, with its ability to give insights on the operation of complex models can be the key to achieve this goal.

As such, the aim of this work was the novel analysis of a face PAD method through the lenses of interpretability. The insights on PAD performance offered by both traditional metrics and the Grad-CAM interpretability tool were compared, to assess whether the latter would offer more meaningful information on the PAD model's performance.

The main contributions of this work are:

1) A pioneer study of interpretability on face PAD methods regarding a wide range of attacks;
2) The comparison between traditional performance metrics and interpretability tools in different evaluation frameworks (one-attack vs. unseen-attack);
3) An evaluation on robustness to unseen data by analyzing the explanations produced for the same samples when seen or not during training (data swap experiment);
4) Formulation of guidelines supporting the need for incorporating interpretability in a better and more meaningful PAD performance assessment.

The remainder of this paper is organized as follows: Section II an overview on interpretability; Section III presents the methodology; Section IV describes the experimental setup; Section V presents the results and their discussion and in Section VI some conclusions are drawn from this work.

II. An Overview on Interpretability

There is no standard definition regarding interpretability and explainability in AI. Some authors define these as follows: *Interpretability*: an interpretation is the mapping of an abstract concept (*e. g.*, a predicted class) into a domain that a human can grasp; *Explainability*: an explanation is the collection of features of the interpretable domain, that have contributed for the produced decision (*e. g.*, classification or regression) [13]. Other authors use the terms interchangeably, regarding both interpretability and explainability as a three-stage process in the development cycle of an ML model, with these stages being named as pre-, in-, and post-model [5].

The efforts of the research community posed on the field of explainable artificial intelligence (XAI) resulted in the development of both interpretable models (pre-, and in-model stage) and explanation methods (post-model stage) over the past few years [3], [5], [7], [14]–[16]. Nonetheless, most efforts have been made on these explanation methods, where the focus is more on understanding an unconstrained and previously built model than on creating intrinsically interpretable models. The XAI contributions span over the fundamental research in machine learning to applications in other fields such as Medicine [15], [17], [18] or Finance [16]. It is not a coincidence that the pioneer application fields are ones where it is of huge importance to foster awareness for the advantages and the necessity of transparent decision making.

Now that the dust raised by the euphoria around artificial neural networks and their over-performing accuracy is starting to settle down, the research community is alert to the reality of being made accountable for what these outstanding models are in fact learning and deciding upon. It is agreed that a lot can be learned by understanding the powerful, black-box-like, deep learning models which achieve remarkable accuracy but fail to provide any information about what exactly makes them reach their predictions. A growing number of works can be found in the literature devoted to interpret and explain the behavior of machine learning systems for diverse problems [19]–[22].

For biometrics, Zee *et al*. [23] combined face recognition and a face PAD method, and tried to use the interpretations to enhance the performance of the recognition method. Concerning PAD, Seibol *et al*. [24] focused on a specific face spoofing attack (morphing attack) and used the interpretation of the model decision to improve their training scheme and therefore the robustness of the PAD system to those attacks.

To the extent of the authors' knowledge, there is a void in the literature regarding the analysis of PAD techniques from the XAI perspective. It is a statement of the present work that the research in PAD methods may be remarkably impacted by the outcomes of following this path. The first effects in the PAD field will be obtained via the reinforcement of trust as an outcome of model validation as well as the improvement of PAD models' robustness through the detection of their hidden vulnerabilities.

III. Methodology

A. Presentation Attack Detection Network

A presentation attack detection method receives as input a biometric trait measurement and returns as output a prediction of the classification of that measurement as belonging to a live individual (referred to as a *bona fide presentation*) or as being a spoof attempt to intrude the system (in this case referred to as an *attack presentation*). In this work, an end-to-end convolutional neural network was trained and used to discriminate between *bona fide* and *attack* presentations.

The network implemented is an end-to-end convolutional neural network (CNN). Thus, it is able to freely learn the most appropriate features for the task at hand. Settings where the model is not constrained to certain features and has freedom to learn the most useful representations are the most interesting for interpretability studies to gain insight of the inner workings of a deep neural network. A relatively simple architecture was

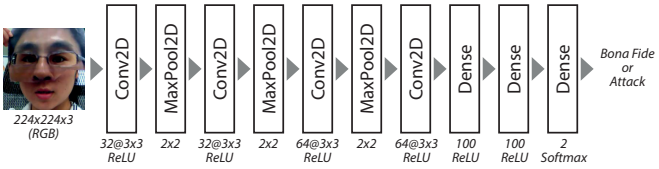

Fig. 1. Architecture of the implemented PAD model.

chosen (see Fig. 1), as the emphasis of this work is to study the interpretability of the face PAD model.

The input to the network is a 224×224 RGB image and the output is a two dimensional softmax layer providing two probability scores. The network is composed of four convolutional layers, with three max-pooling layers interposed between them, and three fully-connected layers. The four convolutional layers are composed of 32, 32, 64, and 64 filters, respectively, with size 3×3, unit stride, and padding. The max-pooling is performed in 2×2 regions with stride 2. The dense layers are composed of 100, 100, and 2 neurons, respectively. All convolutional and fully-connected layers are followed by rectified linear unit (ReLU) activations, except for the last dense layer, which is followed by softmax activation.

B. Interpretability Method

The Gradient-weighted Class Activation Mapping (Grad-CAM) [19] method is inspired by the Class Activation Mapping (CAM) [25], introduced for the identification of discriminative regions in CNNs without fully-connected layers and restricted to the architectures that perform global average pooling over convolutional maps immediately before prediction. Grad-CAM is a generalization of CAM in the sense that it was designed to be used with any type of CNN architecture. This algorithm consists of the combination of feature maps using the gradient signal. Therefore, this gradient information flows into the last convolutional layer of the CNN, thus, assigning different importance values to each neuron for a particular decision of interest. Also, this approach permits to generate explanations for any layer of the network. Moreover, it is possible to obtain explanations per class, allowing the analysis of the model predictions at a class-level.

IV. Experimental setup

A. Data and Pre-processing

The data used in this work was drawn from the ROSE-Youtu Face Liveness Detection Dataset [26] from the Rapid Object Search Lab. This dataset is composed by 3497 videos from twenty subjects. Each subject has several "genuine" videos, where they were directly recorded, and "attack" videos consisting of worn paper masks, paper photos, and replayed recordings (see Table I).

The samples from subjects $\{2, 3, 4, 5, 6\}$ were reserved for testing, while the data from remaining subjects (15) were used for training and validations. From each video, frames were extracted every 5 s and faces were detected using a Multi-Task Convolutional Neural Network (MTCNN) [27]. Frames were

TABLE I
Characteristics of the presentation attack instruments in the ROSE Youtu dataset (N.I. stands for "number of images", *i.e.*, frames extracted from the videos).

Attack	Type of presentation attack instruments	N.I.
-	Genuine (*bona fide*)	2794
#1	Still printed paper	1136
#2	Quivering printed paper	1188
#3	Video of a Lenovo LCD display	923
#4	Video of a Mac LCD display	1113
#5	Paper mask without cropping	1194
#6	Paper mask with two eyes and mouth cropped out	608
#7	Paper mask with the upper part cut in the middle	1162

cropped into square regions around the detected faces, resized to 224×224 and normalised to $[0, 1]$ intensity range. The prepared dataset included a total of 10 118 frames, of which 7396 composed the training set (1977 genuine and 5419 attack frames) and 2722 were use for testing (817 genuine and 1905 attack frames).

B. Implementation Details

The PAD model was implemented, trained, and evaluated using Keras with Tensorflow. The network was trained using the Adam optimizer, with initial learning rate $lr = 0.0001$, a maximum of 150 epochs, and batches with size 8. Early stopping was used, monitoring the validation loss, with patience of 20 epochs. For regularization, dropout (0.5) was used between each pair of consecutive dense layers. Horizontal flips, rotations with range of 20°, and width and height shifts with range 0.2 were used for data augmentation.

C. Interpretable Artificial Intelligence Framework

All the experiments were performed using the *Keras Visualization Toolkit* [28] for Python, which is a library that permits to visualize and debug a trained Keras model. Currently, it supports the visualization of class activation maps, saliency maps, and activation maximization. This framework is important for the Grad-CAM method, which uses class activation maps. In these maps, each pixel is assigned a probability value that will correspond to a specific color, within the range $\{\text{blue}, \text{yellow}\}$, where blue corresponds to the less activated/important pixel and yellow to the most activated/important pixel. Explanations presented on the results section always correspond to the predicted label.

D. Performance Metrics

The metrics used for the evaluation of the face PAD models are, as defined in [29], the *Bona fide Presentation Classification Error Rate (BPCER)*: the proportion of bona fide presentations erroneously classified as attacks; and the *Attack Presentation Classification Error Rate" (APCER)*: the proportion of attack presentations erroneously classified as bona fide. The *Equal Error Rate (EER)* is given by the error at the operation point where the APCER and PBCER take the same value.

TABLE II
PAD PERFORMANCE OF THE MODELS FOR ONE-ATTACK AND UNSEEN-ATTACK EVALUATION FRAMEWORKS. (EER, APCER, AND BPCER IN %)

Attack	One-Attack			Unseen-Attack		
	EER	*APCER*	*BPCER*	*EER*	*APCER*	*BPCER*
#1	7.29	12.15	3.06	5.90	6.94	4.90
#2	3.62	6.67	1.35	5.55	**3.00**	10.65
#3	2.79	8.37	0.12	10.38	26.29	4.28
#4	12.66	30.38	1.84	25.34	45.73	**3.92**
#5	1.61	**1.61**	1.59	**4.84**	3.55	7.10
#6	4.46	5.10	1.10	10.19	12.74	7.71
#7	**0.73**	5.23	**0.00**	15.49	34.31	7.71

V. RESULTS AND DISCUSSION

The results presented in this section were obtained in different experiments. Three different evaluation frameworks were used: *Mix-Attack*, *One-Attack* and *Unseen-Attack*:

a) Mix-Attack: The model is trained and tested with *bona fide* samples and all the varieties of attacks available.

b) One-Attack: The model is trained and tested with *bona fide* samples and only one type of attack (at a time for each type of attack). Therefore, the only type of attack shown to the network during the test phase was already seen in the training step.

c) Unseen-Attack: The model is trained with all but one type of attack and tested with this remaining attack, besides the *bona fide* samples in train and test steps. Therefore, during the test phase, the network is evaluated only with one type of attack that was not present in the training step - referred to as the unseen attack. Whereas in the training phase, all the other types of attacks were available.

Additionally, another experiment was done within the One-Attack evaluation. The *Data Swap experiment* comprised two stages: first, the samples of one random subject were present in the training set and the evaluation was run; and secondly, the split of data was maintained except for that subject which was put on the test set and new evaluations were run.

A. Performance of the face PAD method

Even though the focus of the present work is not on the performance of the face PAD model but on the interpretation of its decisions, it is important to state that the proposed method performs at the state-of-the-art methods' level. One should not aim to interpret a model that lacks PAD abilities by design, as this will impair and bias the drawn conclusions, and the authors had the aim to avoid this.

Table II presents the performance results for One-Attack and Unseen-Attack frameworks. The Mix-Attack evaluation framework provided values of EER, $APCER$ and $BPCER$ equal to 4.90%, 3.41% and 7.40%, respectively. The results generally worsened from the One-Attack to the Unseen-Attack scenario as pointed out before in the literature [11]. This can be seen as a result of the large variability between different types of attacks. Thus, a practical PAD system must be developed in a scenario where the types of attacks of the test set are not seen during the training routine of the models [12].

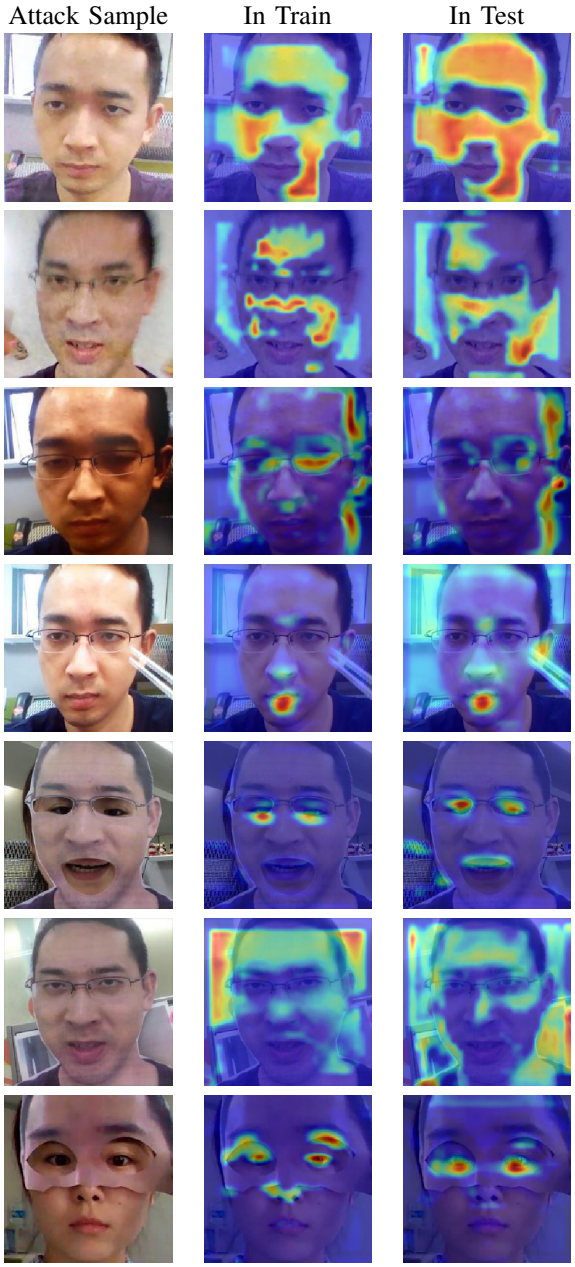

Fig. 2. Grad-CAM Explanations for correctly classified attack samples when a subject is in the train set (2^{nd} column) or in the test set (3^{rd} column). The rows correspond to One-attack #1 to #7.

One example of the inability to generalize is Attack #7 (upper face mask with eyes cropped out), which the model finds very challenging to detect when has been trained with the other types of attacks ($APCER$ goes from 5.27% to 34.31%). Some exceptions stand out, such as the case of Attack #1 (full face printed photo) in which $APCER$ drops from 12.15% to 6.94%. In this case, the model probably learns most of the needed features from the remaining attacks and can, thus, take advantage of the greater availability of data and achieve greater performance in the Unseen-Attack settings.

Fig. 3. Grad-CAM Explanations for correctly and incorrectly classified attack samples (TP/FN) in the Swap experiment for One-Attack #1. The identity was correctly classified when present in the train (2nd column) and misclassified when presented in testing step (3rd column).

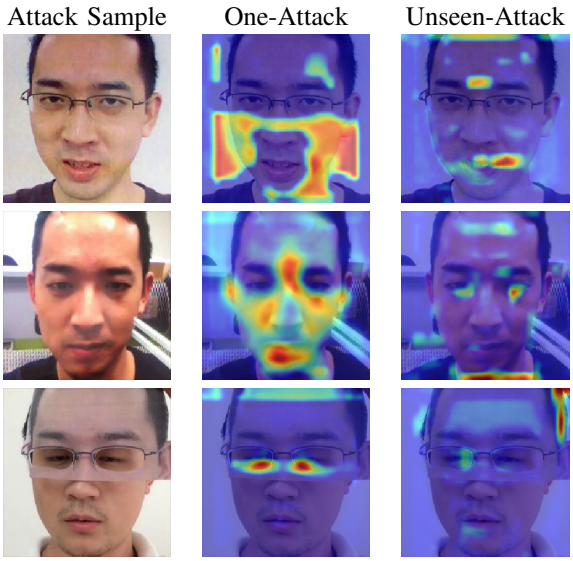

Fig. 4. Explanations for correctly classified attack samples (TP) in the One-Attack (2nd column) or Unseen-Attack (3rd column) frameworks. Each row corresponds to one specific type of attack, top to bottom: #1, #4, and #7.

B. Swap Experiment

Fig. 2 depicts the Grad-CAM explanations for the data Swap experiment. No relevant differences are observed in the explanations, for correct decisions, in one step or the other. This illustrates a desirable behavior of a PAD method that can be verified through interpretability: a robust model should avoid overfitting, using similar information to reach its decisions, regardless of whether or not the image is seen during training.

Despite the robustness shown in the previous examples, these models produced discrepancies and classification mismatches between the two stages of the experiment. In Fig. 3 is shown a sample that was correctly classified when present in the training phase and misclassified when present in the testing phase. However, in a more thorough investigation, it should be noted that, in the case of the FN, the classifier is producing an explanation with significant similarities to the ones produced for a TN example. Specifically in the first row, second column of Fig. 5 is shown the *bona fide* example for the same case, One-Attack#1. It can be noted that in both situations the model uses similar information to reach its decision for *bona fide* label (parts of the top head contour, region around the mouth) thus verifying the aforementioned desirable property.

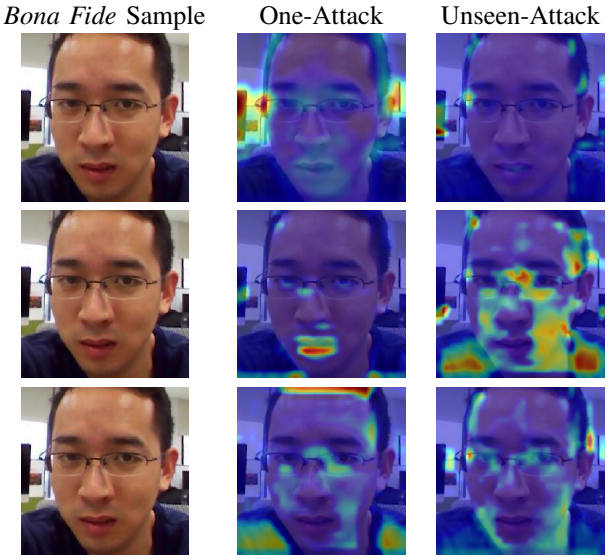

Fig. 5. Explanations for correctly classified *bona fide* samples (TN) for One-Attack (2nd column) and Unseen-Attack (3rd column). Each row corresponds to one specific type of attack, top to bottom: #1, 3 and 5.

C. One-Attack vs. Unseen-Attack

In Fig. 4 and Fig. 5 are depicted examples of explanations obtained for *attack* samples through the One-Attack and the Unseen-Attack evaluation frameworks for *attack* and *bona fide* presentations, respectively. It can be observed that comparing the results of the One-Attack *vs.* Unseen-Attack frameworks, the method does not show coherence on the information that is relevant to making the decision. Despite the fact that the model is producing correct decisions, this observation raises a serious red flag about its reliability.

Arguably, a robust PAD model is expected to present coherence between One-Attack and Unseen-Attack explanations. This coherence is a sign that the model is learning discriminative features of the *bona fide* samples and also of the model's capability to generalize to new attacks. Attackers are constantly devising new and improved ways to spoof biometric systems, and a suitable PAD model should be as much as possible able to generalize from the known attacks. At the same time, the knowledge about the *bona fide* samples should be determined by their "liveness" characteristics and not impacted by changes in the attacks that are displayed in the training. On the other hand, the results in Fig. 5 suggest that, when trained under the Unseen-Attack settings, the model does not show a predictable behaviour when learning about the *bona fide* samples when the changes in the training are only about the types of attack samples. The high variability in the types of attacks in the training in the Unseen-Attack is apparently making the model prioritizing the learning of features for the detection of attacks clearly jeopardizing the learning of liveness features.

VI. Conclusions

In this work, interpretability tools were explored to obtain insights on the inner workings of a deep neural network trained for presentation attack detection (PAD) in face biometrics. The present study supports the statement that interpretability does offer more insight into the intricacies of a PAD deep learning model, thus advisable to be used for its deeper and more thorough performance evaluation.

In the studied models, the traditional metrics did not denote overfitting but through interpretability it was possible to grasp the limitations of the models to satisfyingly generalize from the training data. These observations highlight the necessity for performing model validation using interpretability tools. From this study, the acquired knowledge should then be used to interpret the decision making process and therefore anticipate vulnerabilities and ultimately to adapt the model to overcome the anticipated vulnerabilities.

Additionally, we were able to identify some "desirable properties", assessable using interpretability tools, that should be verified by PAD methods: (1) explanations for the same sample should be similar whether or not it is seen during training (data swap); (2) explanations for the same sample should be similar whether or not the model is trained to detect that specific attack (One-Attack *vs.* Unseen-Attack); (3) explanations should be similar for different samples with the same predicted label (intraclass coherence); (4) explanations should be meaningful (a human would likely use them to provide the same decision). These are properties that should be verified by a PAD method that is robust, coherent, meaningful, and can adequately generalize to unseen data and attacks.

Interpretability is a flourishing research topic, still ridden with subjectivity and uncertainty. Thus, further efforts should be devoted to reducing subjectivity in the evaluation of explanations, through objective metrics and procedures [30]. Then, interpretability will be ready for its application for PAD, which should be further explored. At last, the standards for evaluation of PAD methods should be redefined through new objective metrics using explanations obtained with interpretability tools.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. of the IEEE CVPR*, 2014, pp. 1725–1732.

[3] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[4] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainabilty of artificial intelligence in medicine," *Wiley Interdisc. Reviews: Data Mining and Knowl. Disc.*, p. e1312, 2019.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[6] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proc. of the IEEE CVPR*, 2016, pp. 2912–2920.

[7] W. Samek and K.-R. Müller, *Towards Explainable Artificial Intelligence*. Springer International Publishing, 2019, pp. 5–22.

[8] D. Perez-Cabo, D. Jimenez-Cabello, A. Costa-Pazo, and R. J. Lopez-Sastre, "Deep anomaly detection for generalized face anti-spoofing," in *IEEE CVPR Workshops*, June 2019.

[9] S. Bhattacharjee, A. Mohammadi, A. Anjos, and S. Marcel, *Recent Advances in Face Presentation Attack Detection*. Cham: Springer International Publishing, 2019, pp. 207–228.

[10] A. Rattani, W. Scheirer, and A. Ross, "Open set fingerprint spoof detection across novel fabrication materials," *IEEE TIFS*, vol. 10, no. 11, pp. 2447–2460, Nov 2015.

[11] A. F. Sequeira, S. Thavalengal, J. Ferryman, P. Corcoran, and J. S. Cardoso, "A realistic evaluation of iris presentation attack detection," in *39th TSP*, June 2016, pp. 660–664.

[12] P. M. Ferreira, A. F. Sequeira, D. Pernes, A. Rebelo, and J. S. Cardoso, "Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations," in *2019 Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2019, pp. 1–7.

[13] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1 – 15, 2018.

[14] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," *arXiv preprint arXiv:1711.11279*, 2017.

[15] W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso, "Towards complementary explanations using deep neural networks," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, 2018, pp. 133–140.

[16] W. Silva, K. Fernandes, and J. S. Cardoso, "How to produce complementary explanations using an ensemble model," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[17] M. Graziani, V. Andrearczyk, and H. Müller, "Regression concept vectors for bidirectional explanations in histopathology," in *Underst. and Interp. ML in Medical Img.e Comp. Appl.* Springer, 2018, pp. 124–132.

[18] J. Zhuang, J. Cai, R. Wang, J. Zhang, and W. Zheng, "CARE: Class attention to regions of lesion for classification on imbalanced data," in *Int. Conf. on Medical Imaging with Deep Learning*, 2019, pp. 588–597.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[20] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, "What is relevant in a text document?: An interpretable machine learning approach," *PloS one*, vol. 12, no. 8, 2017.

[21] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," *arXiv preprint arXiv:1706.07206*, 2017.

[22] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *preprint arXiv:1704.07911*, 2017.

[23] T. Zee, G. Gali, and I. Nwogu, "Enhancing human face recognition with an interpretable neural network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[24] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Accurate and robust neural networks for security related applications exampled by face morphing attacks," *arXiv preprint arXiv:1806.04265*, 2018.

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[26] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[28] R. Kotikalapudi and contributors, "keras-vis," https://github.com/raghakot/keras-vis, 2017.

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

[30] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.