Machine Vision and Applications (2025) 36:44
https://doi.org/10.1007/s00138-025-01666-z

RESEARCH

Interpretability of fingerprint presentation attack detection systems: a look at the "representativeness" of samples against never-seen-before attacks

Simone Carta[1] · Roberto Casula[1] · Giulia Orrù[1] · Marco Micheletto[1] · Gian Luca Marcialis[1]

Received: 8 November 2024 / Revised: 13 January 2025 / Accepted: 17 January 2025

Abstract
Nowadays, fingerprint Presentation Attack Detection systems (PADs) are primarily based on deep learning architectures subjected to massive training. However, their performance decreases to never-seen-before attacks. With the goal of contributing to explaining this issue, we hypothesized that this limited ability to generalize is due to the lack of "representativeness" of the samples available for the PAD training. "Representativeness" is treated here from a geometrical perspective: the spread of samples into the feature space, especially near the decision boundaries. In particular, we explored the possibility of adopting three-dimensionality reduction methods to make the problem affordable through visual inspection. These methods enable visual inspection and interpretation by projecting data into two-dimensional spaces, facilitating the identification of weak areas in the decision regions estimated after the training phase. Our analysis delineates the benefits and drawbacks of each dimensionality reduction method and leads us to make substantial recommendations in the crucial phase of the training design.

Keywords Fingerprints · Generalization · Interpretability · Presentation attack · Visual inspection

1 Introduction

Because of their uniqueness and immutability, fingerprints are among the biometric traits most commonly used for personal authentication [1]. Although the maturity of Automated Fingerprint Identification Systems (AFISs), their exposure to presentation attacks is well-known [2]. Presentation Attacks (PAs) require the design of an additional detection module to prevent their occurrence.

The current Presentation Attack Detection systems (PADs) performance is challenged by never-seen-before Presentation Attack Instruments (PAIs). To ensure the robustness of PADs, the designer must anticipate and incorporate a variety of known PAI types during model training; this process is frequently hampered by the prohibitive costs associated with obtaining PAIs crafted with state-of-the-art techniques and materials [3]. Due to this limitation, PADs fail to generalize to PAIs made with new materials or techniques [4, 5].

The present work hypothesizes that this low ability to generalize is linked to the poor "representativeness" of the model's training data. We associated a geometrical meaning to the term "representativeness": the sample is representative if it falls into a region of the feature space where the training samples are "enough" dense, making the decision boundary stable even if other samples are added. Accordingly, detecting sampling "gaps" in the feature space regions would allow us to address the generalization problem.

✉ Roberto Casula
roberto.casula@unica.it

Simone Carta
simone.carta97@unica.it

Giulia Orrù
giulia.orru@unica.it

Marco Micheletto
marco.micheletto@unica.it

Gian Luca Marcialis
marcialis@unica.it

[1] DIEE, University of Cagliari, Piazza d'Armi I, 09123 Cagliari, Italy

Published online: 24 January 2025

Our hypothesis involves examining how samples are distributed in the feature space and their relation to the machine-learning decision boundary. This is impossible to achieve in the original feature space, usually very large, and only indirect measurements of clustering level or density estimation may help. However, they do not allow direct feedback on the sample distribution in the feature space and its relationship to the decision boundary [6].

In this regard, algorithms to obtain a significantly reduced feature space from the original have been proposed because of the human impossibility of visualizing such samples in a multidimensional feature space [7–10]. To our knowledge, these methods were adopted to observe how the classes were spread or grouped but never used actively in the design process.

In this paper, we showed that these methodologies can be helpful in the training phase and can strongly impact the accuracy of fingerprint PADs. The designer can directly identify underrepresented classes in the feature space and determine which samples are needed to improve the network's decision boundaries.

A model capable of providing such insights may be called "interpretable" [11]. This paper defines the term "model" by the couple of network architecture-training data. In fact, the architecture alone can be highly complex, embedding numerous parameters, making it unlikely to provide significant insights to the designer. In addition, the ablation technique is widely used to assess the importance of individual modules within the architecture. However, the training data used with a given architecture are equally important, as different training sets can produce varying responses, even for the same task. Similarly, when conducting ablation studies, it is essential to use the same training data for consistent evaluation. Therefore, understanding the relationship between the training data and the final model is just as critical as selecting the best architecture for the task.

The primary goal of this work is to make fingerprint PADs more robust to unseen attacks by identifying techniques that provide deeper insights into how the PAD system maps various attacks in the feature space, enabling a clear understanding of the logic behind the predictions. These techniques must be effective through visual inspection, simplifying the designer's job of providing the most appropriate training set for the given architecture. As depicted in the flowchart in Fig. 1, the work begins by identifying the challenge of generalization in fingerprint PAD systems due to insufficient training data. This is followed by a geometric analysis of sample distributions in the feature space, with the ultimate goal of enhancing PAD robustness against unseen attacks through better training set design.

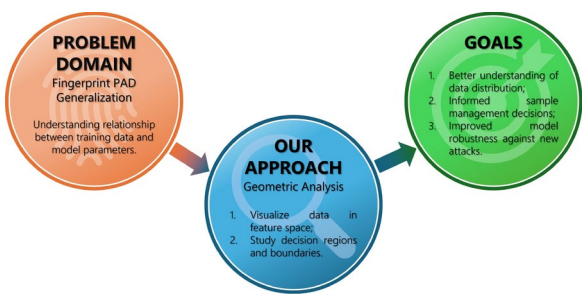


Fig. 1 Overview of the problem domain, proposed geometric analysis approach, and research goals

In particular, we detailed the advantages and limitations of each visualization technique by providing substantial recommendations for their application in advancing the knowledge and development of PAD models. Furthermore, we simulated conditions where training data may be limited or incomplete, providing insights into how much this may compromise the effectiveness of PADs.

The paper is organized as follows: Sect. 2 provides an overview of the current literature, listing the main approaches to fingerprint presentation attack detection and dimensionality reduction for data visualization. Section 3 describes the theoretical background of our proposed approach in detail. Section 4 describes our experiments, and discusses the achieved results. Conclusions are drawn in Sect. 5.

2 Related works

2.1 Generalization in fingerprint PAD

A considerable number of PAs have been successful over the last few decades, demonstrating the weakness of AFISs and the urgency of taking action to strengthen defenses for the protection of both corporate and private assets [2, 12].

The Deep Learning boom affected the development of PAD systems too, increasingly starting to replace algorithms based on hand-crafted feature extraction due to its remarkable effectiveness [13]. However, both these approaches are primarily hampered by a lack of generalization, particularly when detecting unknown attacks. Developing techniques for dealing with never-seen-before attacks remains a critical challenge [14, 15].

Over the past decade, various approaches have aimed to enhance performance against PAs created from unknown materials, often referred to as *cross-material* performance [4]. These include techniques like Weibull-calibrated SVMs, ensemble one-class SVMs, and the use of Generative Adversarial Networks (GANs) on training data consisting solely of genuine samples [3], showing improved performance across different fabrication materials. Notably, the choice of PAI materials used during training significantly impacts performance against unknown PAs. Some efforts have been made to analyze the characteristics of different

PAI materials and create a Universal Material Generator (UMG) that interpolates textural characteristics from known materials to synthesize images of unknown materials, thus improving generalization [10].

Synthetic data augmentation has gained prominence as it enhances model generalization against unseen data, addresses data scarcity, reduces biases in real-world datasets, and mitigates privacy concerns [16]. Recent research has been predominantly focused on these directions [17, 18]. To ensure a significant improvement in generalization performance, synthetic data should accurately fill the gaps in the training set to cover most of the PA feature space, in the spirit of Chugh et al. [5]. Fingerprint PAD can undoubtedly benefit from augmentation with synthetic images [10]; however, clear guidelines on the type (e.g., gelatine- or latex-like fingerprints) and quantity of synthetic data needed are lacking. The empirical rule is simply the more, the better.

According to this, we believe it is of the utmost importance to increase data representativeness using synthetic or real data and accurately identify undersampled areas of the feature space. Because of this, the generation of synthetic data, or acquisition of further real data that does not consider training data representativeness could be of limited help.

Some works include visualization of multi-dimensional data or feature vectors as points in 2D or 3D space, to simplify them into a more interpretable form by means of dimensionality reduction techniques, such as Principal Component Analysis (PCA) [19] and t-distributed Stochastic Neighbor Embedding (t-SNE) [20]. The main limitation of their current use is that, to the best of our knowledge, they do not appear to be used to extract and analyze meaningful map characteristics, such as the distribution density and the presence of non-sampled areas. We believe researchers and security professionals can gain easily interpretable insights into the training data representativeness by leveraging dimensionality reduction techniques.

2.2 Interpretability models

As machine learning (ML) models have become prevalent in various sensitive applications, the need for interpretability of models has become increasingly important [21]. According to Ref. [11], *interpretability* focuses on understanding the internal mechanisms of a model. In other words, an interpretable model has peculiar properties that allow the designer to adjust its parameters according to well-defined rules or guidelines.[1]

[1] Although this term is sometimes interchanged with *explainability* [22], we will always use *interpretability* according to what is clarified in [11]. There, *explainability* is referred to justify the predictions made by a model in a human-understandable manner.

Current interpretable algorithms [23] are primarily concerned with obtaining importance scores for each feature, visualizing semantic relationships between them [24], bringing a deeper understanding of model behavior. Typically, these methods are based on visualization techniques obtained with simple bar graphs representing feature importance scores, dimensionality reduction techniques, saliency maps [25] in vision and text models, and semantic maps. Of these, dimensionality reduction (DR) methods [8], such as PCA and t-SNE, are widely used to interpret PADs ability to generalize across domains [5]. These reduction methods are numerous and can be divided into linear and non-linear methods [26]. Linear methods [27] can capture critical data properties such as covariance patterns, dataset correlations, and input–output interactions due to simple geometric interpretations and effective computational capabilities. Non-linear methods [28] can preserve high-dimensional neighborhoods and relative distances during reduction. This is especially useful when dealing with the complexities of real-world data. However, this comes at the trade-off of decreased interpretability of the generated visualization.

Since the differences between these methods are many, and their behavior changes depending on the nature of the data analyzed, it is important to understand which of these is most suitable in relation to different interpretative purposes. For this reason, this work aims to explore the use of reduction techniques to evaluate the representativeness of the training set samples during the PAD design and to provide the designer with guidelines to best interpret the resulting visualization.

3 Data visualization for PAD design and interpretation

As mentioned in the previous sections, a fingerprint presentation attack detection model comprises the pair architecture-training set. In this paper, we focus on the insights that can be extracted from the observation of training data by crossing them with the architecture parameters set after the training phase.

As it is well known, the model parameters broadly define the estimation of *a posteriori* probability classes for the training data spread over the feature space. The analysis of this relationship may be carried out by a statistical approach or by a geometric approach. In this paper, we focus on a geometric approach where the data are visualized in their feature space locations, and we study their relationship with the decision regions and boundaries. In fingerprint PAD, such decision regions are linked with the *bona fide* and the PA classes [29]. From the geometrical viewpoint, the decision regions depend on the distributions of samples

and their possible clusters, or groups, at varying degrees of mutual overlapping.

We aim to provide guidelines to understand and improve PAD systems using feature space visualization techniques and correctly interpret distances between samples and clusters. Given the potentially infinite feature space of fingerprints (including unknown PAI materials and techniques), each training dataset represents a finite and limited sample. Moreover, being PAD an arms race problem, it is expected that new PAIs will be designed and used over time, and related samples may concur with the feature space population. Assuming that fingerprint images, characterized by user information, specific acquisition conditions, materials, or manufacturing techniques, can be represented as points in a generic feature space, we can suppose that fingerprints sharing common characteristics are grouped in clusters. Visualizing this information is essential for a PAD designer. In fact, viewing the reciprocal positions of the two clusters makes it possible to evaluate whether they are too overlapped or too distant. However, information and noise are strongly mixed in a high-dimensional feature space. Reducing this complex space to a more manageable representation is necessary. Feature reduction techniques help achieve this goal by:

1. Enabling the preservation of significant characteristics: the transformation should retain critical information from the high-dimensional space essential for distinguishing between different types of fingerprints.
2. Ensuring visual separability of clusters: in the reduced two-dimensional space, clusters corresponding to different types of fingerprints should be visually distinct.In this passage from R^n to R^2, designers can empirically observe the effect of training with respect to the geometric distribution of the samples in the feature space.

Techniques well-suited for this purpose include Principal Component Analysis (PCA) [19], t-Distributed Stochastic Neighbor Embedding (t-SNE) [20], Uniform Manifold Approximation and Projection (UMAP) [30] and Isomap Embedding (IsoMAP) [31]. We employed them as reference methods in our study.

PCA is the common approach to dimensionality reduction, employed for both visualization and data de-noising purposes. Being a linear method, it is expected to highlight global structures in data. t-SNE was created to stagger distances in the neighborhood structure. While some implementations, such as the Barnes-Hut approximation, offer efficiency gains, t-SNE has several disadvantages, including high computational costs and stochastic nature. UMAP and IsoMAP, like t-SNE, allows multidimensional data visualization via non-linear dimensionality reduction. Through uniform distributions of data on the Riemannian manifold, locally constant Riemannian metric, and local connectivity of the manifold, UMAP allows the visualization of a weighted graph where the edge weights represent the probability that two points are connected. It aims to maintain the global structure and continuity of clusters but prioritizes local distances over long-range ones. IsoMAP, on the other hand, is based on the standard linear Multidimensional Scaling (MDS) method [32]. In essence, MDS serves as a tool for assessing similarities or dissimilarities within data, aiming to locate low-dimensional points such that their pairwise distances in the reduced space closely resemble the distances in the original high-dimensional feature space. Instead of using Euclidean distances, which may not adequately represent the underlying low-dimensional manifold structure, as PCA and MDS do, it uses geodesic manifold distances to attempt to capture the intrinsic geometry of the data. We expect t-SNE and UMAP, broadly popular non-linear data visualization approaches, to recreate local cluster shape and composition better. However, mutual distances between clusters and their relative sizes could lead to erroneous observations [33].

As a result, it is essential to carefully consider which technique provides the most reliable insights for accurately visualizing and interpreting the geometric distribution of the samples. The more accurate the visualization, the better designers can understand the data and make informed decisions grounded in empirical analysis, such as adding or removing samples [6]. By observing how these interventions affect the training process, designers can refine the model and ultimately improve its robustness against never-seen-before attacks. The following sections will demonstrate how these techniques are applied in our experiments to achieve these goals.

4 Experimental observation

4.1 Protocol

To effectively map high-dimensional feature vectors to a two-dimensional space with minimal information loss and assess a model's ability to generalize, it is essential to understand the differences between linear methods, such as PCA, and non-linear methods, such as t-SNE and UMAP. Acknowledging the strengths and weaknesses of each approach is crucial for selecting the appropriate technique for a given dataset. To this aim, we applied the PCA, UMAP, and t-SNE methods on the feature vector obtained on four different descriptors:

- a custom and simple *CNN* (white-box), consisting of two convolutional layers, each followed by max pooling and dropout for regularization, and employing global average pooling before passing through two fully connected layers.
- *JLWLivDetL* [34], proposed by Hangzhou Jinglianwen Technology Co., Ltd., relies on a combination of hand-crafted and Deep Learning-based features (black-box). The base network, Slim-ResCNN [35], is specifically adapted and designed to be lightweight and suitable for real-time systems. It incorporates several enhanced residual blocks, with a dropout layer added to each kernel pair to mitigate overfitting.
- *Megvii*,[2] submitted to the LivDet 2021 competition by Megvii Technology Co., Ltd., adopts a feature extraction approach purely based on Deep Learning (black-box). The backbone of this detector is the assembly of two different networks, ResNetxt and MobileNetV3 pretrained on ImageNet [36]. The goal of this algorithm is to fit unknown data by decreasing the generalization error, thus simultaneously lowering the variance and squared bias. However, as these two elements decrease, the complexity of the model increases. The authors employ regularization techniques such as data augmentation, model ensemble, and score averaging to manage this trade-off.
- *PADUnk* [37], proposed by the University of Applied Sciences of Darmstadt, relies solely on hand-crafted features (black-box). The encoding is performed using the Fisher Vector (FV) technique, which combines local and global information from various local feature descriptors to enhance the generalization ability of the PAD system. These descriptors capture different fingerprint characteristics, including gradients, intensities, and textures. The extracted features are encoded using a parametric generative model, such as a Gaussian Mixture Model (GMM) or a Bernoulli Mixture Model (BMM), to produce an FV representation. This representation describes how the feature distribution of a fingerprint deviates from the previously learned distribution. The final decision regarding the fingerprint's liveness is then made using a linear Support Vector Machine (SVM) classifier.The reduced vectors are then displayed as data points on a 2D feature space. The arrangement of the points in the plane and the quality of the reproduction of the decision function are evaluated in order to select the most suitable representation to interpret the distribution of the training set.

[2] https://www.youtube.com/watch?v=WrlR1XFdyXU&t=43 s.

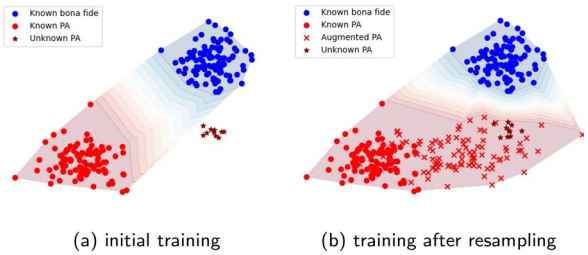


Fig. 2 Toy diagram that exemplifies how feature reduction techniques can be exploited to obtain the interpretation of a PAD model. From the initial training, the designer realizes that the training set has a subsampling area (**a**); after adding new samples in that area the model is able to correctly classify some samples (stars) that were initially classified incorrectly (**b**)

Table 1 Characteristics of scanners used in LivDet 2019 and LivDet 2021 acquisitions

Scanner	Resolution (dpi)	Image size	Format	Type
Green Bit DactyScan84C	500	500 × 500	BMP	Optical
Dermalog LF10	500	500 × 500	PNG	Optical

To better highlight the predictions of the PADs associated with the projected feature vectors, we used a different color in the graphs for each label (BF and different PAI materials) and we drew the classification regions with a diverging colormap ranging from blue to red (high score, i.e., $\geq 50\%$ in blue) or PA (low score, i.e., $< 50\%$ in red) in Euclidean space. As is evident from the example played in Fig. 2, the designer can evaluate the degree of separation or overlap of the two clusters BF and PA from the analysis of the training set; moreover, using a validation set, if the position of new samples falls in an undersampled area, he/she can add samples to the training set to cover those areas. This analysis can then guide the sampling phase of the training set or an augmentation phase.

We adopted the LivDet 2019 [38] and 2021 datasets [39] for experiments. We have also included the results on LivDet 2023 [40] in the supplementary materials. Such datasets represent a noteworthy and reliable benchmark for testing the performance of PAD algorithms, both because of the modern acquisition hardware technologies and because of the presence of PAIs created with state-of-the-art techniques, such as the semi-consensual ScreenSpoof fabrication technique [41], which has been dangerously capable of causing notable performance drops in the vast majority of competitors. In Table 1 we report the characteristics of the acquisition sensors, i.e., Green Bit DactyScan84C and Dermalog LF10, of the datasets used in this analysis.

Table 2 reports the training set composition, comprising 2750 fingerprint images. Green Bit and Dermalog test sets, namely, the set of samples to be mapped in the feature

Table 2 LivDet 2021 and 2019 training set composition

Dataset	Bona fide	PAI materials				
		Latex (glue)	RProFast (silicone)	Woodglue (glue)	Ecoflex(silicone)	Gelatine (hybrid)
LivDet 2021 Green Bit	2050	1250	750	–	–	–
LivDet 2021 Dermalog	2050	1250	750	–	–	–
LivDet 2019 GreenBit	1000	250	–	250	250	250

PAs are fabricated using the consensual method only

Table 3 LivDet 2021 test set composition

Dataset	Bona fide	PAI materials				
		Body Double (silicone)	Mix1 (hybrid)	Elmers Glue (glue)	GLS20 (silicone)	RFast30 (silicone)
Green Bit	2050	820/820	820/820	820/820	–	–
Dermalog	2050	–	–	–	1230/1230	1230/1230

The number of PAs created using the consensual (CC) and ScreenSpoof (SS) methods is reported for each PAI material

Table 4 Feature vector size of PAD algorithms and liveness accuracy evaluated against the LivDet 2021 test set

PAD	Feat. vec. size	GreenBit					Dermalog				
		BPCER [%]	CC		SS		BPCER [%]	CC		SS	
			APCER [%]	Acc.[%]	APCER [%]	Acc. [%]		APCER [%]	Acc. [%]	APCER [%]	Acc. [%]
JLWLivDetL	1280	2.59	8.21	94.35	54.11	69.31	0.68	2.8	98.16	95.12	45.41
Megvii	64	0.29	6.3	96.43	13.94	92.26	0.83	0.77	99.2	29.07	83.77
PADUnk	65,536	1.46	37.2	79.05	18.42	89.29	2.68	4.8	96.16	24.72	85.3
Simple-CNN	64	10.24	21.83	83.44	45.12	70.73	13.17	3.94	91.86	49.96	66.76

space, consist of 6970 fingerprint images each, acquired from 41 different subjects. PA samples in the test set are fabricated using PAI materials unseen during training for generalization ability evaluation. Table 3 summarizes the test set composition.

Table 4 reports further details of the PADs analyzed, including the size of the generated feature vectors and the presentation attack detection accuracy for each test set. Performance is evaluated using standard measures such as APCER (rate of misclassified presentation attacks) and BPCER (rate of misclassified *bona fide* fingerprints) [42], along with a more generic accuracy metric that indicates the percentage of samples correctly classified by the PAD, by inverting the weighted average of APCER and BPCER.

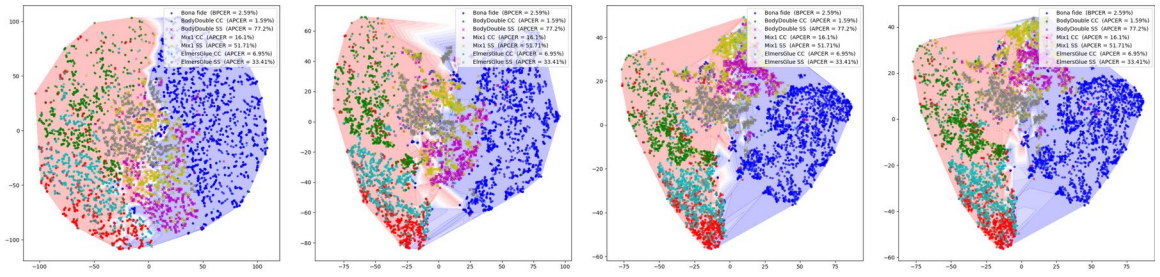
4.2 Interpretation of the feature mapping

In this section, we conduct three analyses using the proposed 2D representation. First, we examine the parameterization of non-linear methods in order to ensure that parameters are tailored to the specific characteristics of the data and objectives of the analysis (Sect. 4.2.1). Next, we examine different DR methods applied to feature vectors from three PAD models to determine the most effective approach for interpretability (Sect. 4.2.2). Finally, we investigate the impact of feature space subsampling during training (Sect. 4.2.3), driven by the hypothesis that insufficient sample representativeness, particularly near decision boundaries, weakens PADs' generalization to unseen attacks. To simulate downsampling, we re-trained the models with the same random seed to ensure consistent initialization, progressively excluding portions of outlier samples to observe the effect on model robustness.

4.2.1 Parameter analysis

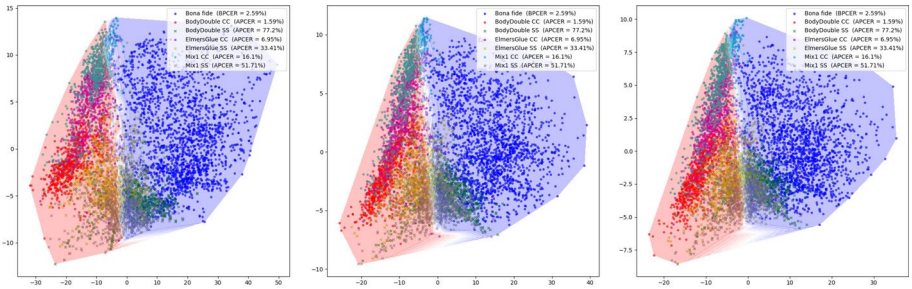
The t-SNE, UMAP and IsoMAP algorithms are highly sensitive to parameter selection as non-linear dimensionality reduction methods. To better preserve global data structure, we initialized t-SNE with PCA and UMAP with the Laplacian Eigenmaps (LE) technique [43]. IsoMAP does not rely on an explicit initialization but constructs a lower-dimensional embedding by leveraging geodesic distances in the neighborhood graph and applying multidimensional scaling on the resulting distance matrix. In addition to initialization, various parameters may still be tuned.

For t-SNE, the most critical parameter is perplexity (denoted as P in this study), which approximates a target number of neighbors for each point and balances the preservation of local versus global data structures. According to the original authors [20], P values between 5 and 50 generally yield stable and reliable visualizations, as the algorithm effectively preserves important relationships in the data across this range. As a rule of thumb, they also suggest that larger datasets may require higher values, although P should



(a) $P = 5$ (b) $P = 30$ (c) $P = 70$ (d) $P = 200$

Fig. 3 Visualization of feature space separation using t-SNE algorithm at varying levels of *perplexity* (*P*): **a** 5, **b** 30, **c** 70, and **d** 200



(a) $N = 5$ (b) $N = 15$ (c) $N = 50$

Fig. 4 Visualization of feature space separation using IsoMAP algorithm at varying levels of the *n_neighbors* (*N*) parameter: **a** 5, **b** 15, **c** 50

not exceed the total number of samples. The default value proposed by the scikit-learn implementation is $P = 30$.

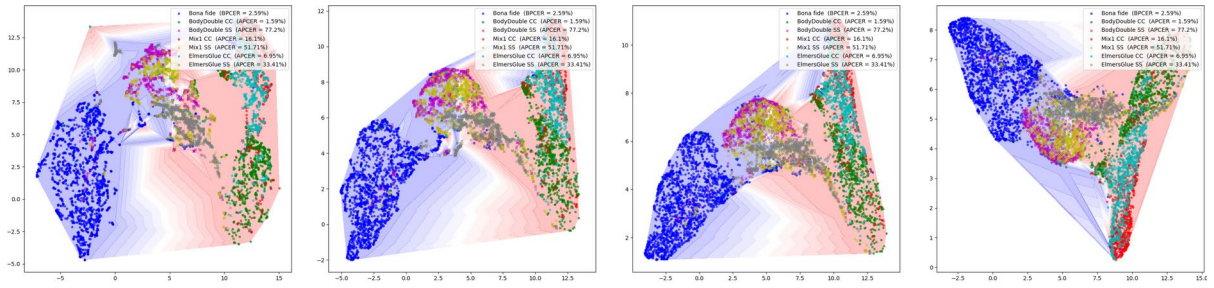
However, other researchers [9] have observed that, for larger datasets, higher values of $P = 30$ are often necessary, as they induce long-range attractive forces during t-SNE optimization, which reduce fine details but help retain the visualization of larger structures. Consequently, they recommend $P = n/100$, where n is the number of samples in the dataset. Following these studies, we applied four P values ([5, 30, 70, 200]) to reduce the feature vectors of the LivDet2021 test set across the four PAD models analyzed in this work and evaluate how this parameter influences the feature space visualization. In our case, applying the rule $P = n/100$ we obtained the value 70.

Due to space constraints, we present only the results for the JLWLivDetL model, as all four PAD models yielded similar outcomes across feature vector sizes. Results in Fig. 3 show that as P increases, points progressively concentrate into well-defined clusters, occupying smaller regions in the feature space. However, higher values of P also increase computational time.

At very low values, such as $P = 5$, points tend to be evenly spaced, preventing the formation of compact clusters. In contrast, no significant differences were observed between $P = 30$ and $P = 70$ to justify the additional computational cost associated with the higher value. Based on these observations, we adopted $P = 30$ for the remainder of the analysis, as it offers a balance between identifying subsampling areas and maintaining low computational costs. For larger datasets, however, we recommend using a higher P value (Fig. 4).

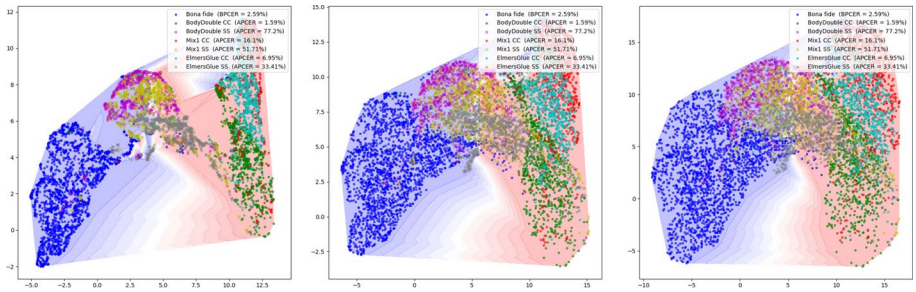
With UMAP and IsoMAP, the parameter *n_neighbors* (which we denote as N in this study) can be seen as analogous to t-SNE's *perplexity*, as it similarly balances the preservation of local and global data structures. The UMAP results (Fig. 5) obtained by varying N across values [5, 15, 50, 200] indicate that well-defined clusters are visible even at relatively low values. However, minimal differences in cluster visualization are observed between $N = 15$ and $N = 50$, except for an increase in computation time as N grows. Some visualizations created with different N values may appear rotated or flipped; this is due to the stochastic nature of UMAP, which does not alter the relative distances between points. Based on these observations, we adopted $N = 15$ for the remainder of the UMAP analysis. However, for IsoMAP, no difference is appreciated when varying N (Fig. 4); for this reason, the value that leads to the lowest computational complexity was selected, i.e. the default value $N = 5$.

Another UMAP parameter, D (corresponding to *min_dist*), also affects feature display by controlling the minimum distance between points. As D increases, clusters become less dense, and points overlap less. Rather than causing clusters to expand spatially, increasing D widens the gaps between distinct groups. This parameter has no notable impact on computational time, and its variation (Fig. 6) allows different visual aspects to be highlighted: lower D values emphasize cluster compactness and inter-cluster distance, while



(a) $N = 5$ (b) $N = 15$ (c) $N = 50$ (d) $N = 200$

Fig. 5 Visualization of feature space separation using UMAP algorithm at varying levels of the *n_neighbors* (N) parameter: **a** 5, **b** 15, **c** 50 and **d** 200



(a) $D = 0.10$ (b) $D = 0.50$ (c) $D = 0.99$

Fig. 6 Visualization of feature space separation using UMAP algorithm at varying levels of the *min_distance* (D) parameter: **a** 0.10, **b** 0.50, **c** 0.99

higher values produce more sparse maps. For the rest of the analysis, we use $D = 0.10$.

4.2.2 Interpretability by visualization

This section examines how the DR techniques under investigation (PCA, t-SNE, UMAP and IsoMAP) can support interpretability in the PAD design phase. During this phase, the designer has access to both training and validation data. Analyzing the feature space of the training set helps assess the model's ability to separate the two classes, BF and PA, and identify any sparse, potentially under-sampled areas within the feature space. On the other hand, examining the feature space representation of the validation set provides insights for interpreting PAD outputs, as the positioning of samples and their proximity to others can help explain the model's behavior.

Figures 7, 8, 9, 10, 11, 12, 13 and 14 depict the training and validation set representations of the models described in Sect. 4.1. In the validation set representations (bottom row of each figure), APCER and BPCER values per material are also included to indicate model performance. Each model was trained on the LivDet2021 training set and validated on the LivDet2021 test set (both for the GreenBit and Dermalog sensors). From these reduced representations, a PAD designer can qualitatively evaluate the degree of separation of the clusters in the design phase. Observing the training set's feature space, it becomes clear that models with well-separated clusters tend to perform better on validation data than those with overlapping clusters. This is evident, for instance, when comparing the Simple-CNN model (Figs. 7, 8) and the JLW model (Figs. 9, 10). The Simple-CNN model displays more overlap between bona fide and attack samples in the training set representations, indicating weaker class separation. As a result, it shows higher error rates on the validation set, particularly an increased BPCER, as indicated in Table 4. In contrast, the JLW model achieves distinct clusters in the training set, which generally translates to lower validation error rates and improved robustness, especially against consensual attacks.

We can also conclude from the Megvii results (Figs. 11, 12 {a,d}) that such separation is not always evident when using PCA representations. When reducing spaces of very high dimensionality with PCA, the two dimensions of the reduced space are often insufficient to capture all the relevant information. In contrast, non-linear representations such as t-SNE and UMAP do not show this limitation (Figs. 11, 12 b,c,f,g); they spread data points across a significantly larger surface, enhancing the definition of each cluster and emphasizing the mutual distance between them. This also facilitates the analysis of the distance between different PAI materials in both the training set and the validation set (see, for example, Figs. 11f and 13f, where each material occupies a distinct area of the space). IsoMAP appears to fall somewhere between PCA and the more flexible nonlinear methods like t-SNE and UMAP. It generally performs better than PCA in separating clusters in the high-dimensional space because it preserves the global geometric structure

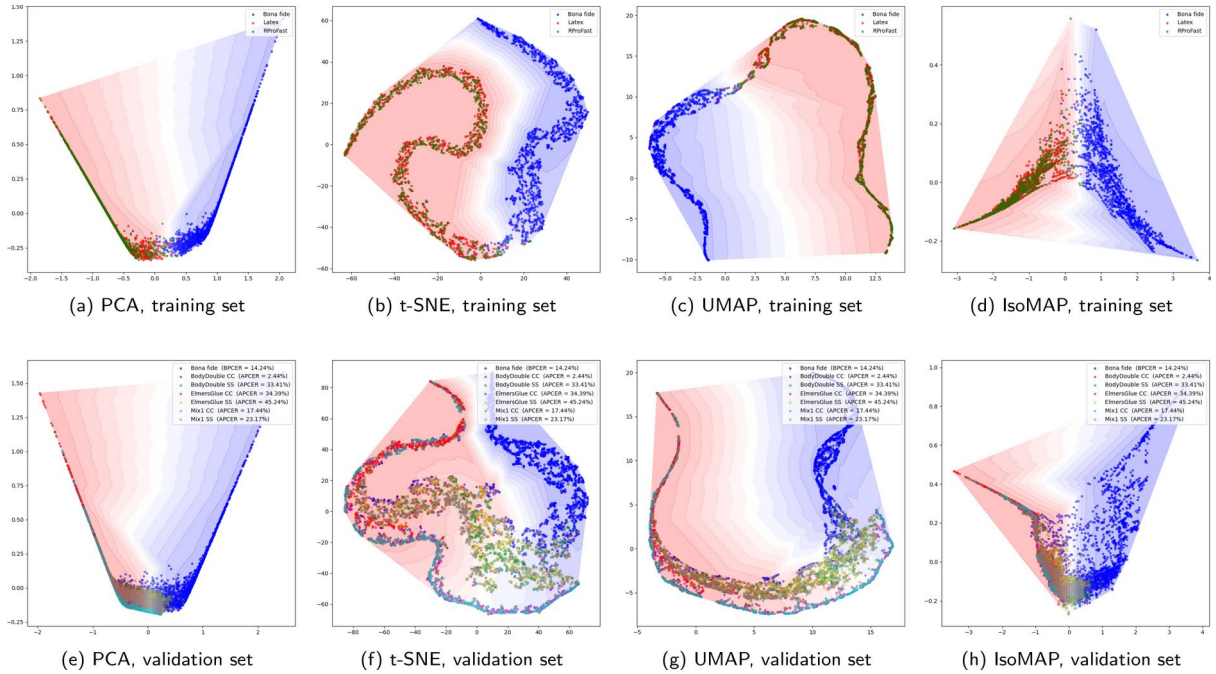


Fig. 7 Training set and validation set visualization for the Simple-CNN PAD (Green Bit scanner)

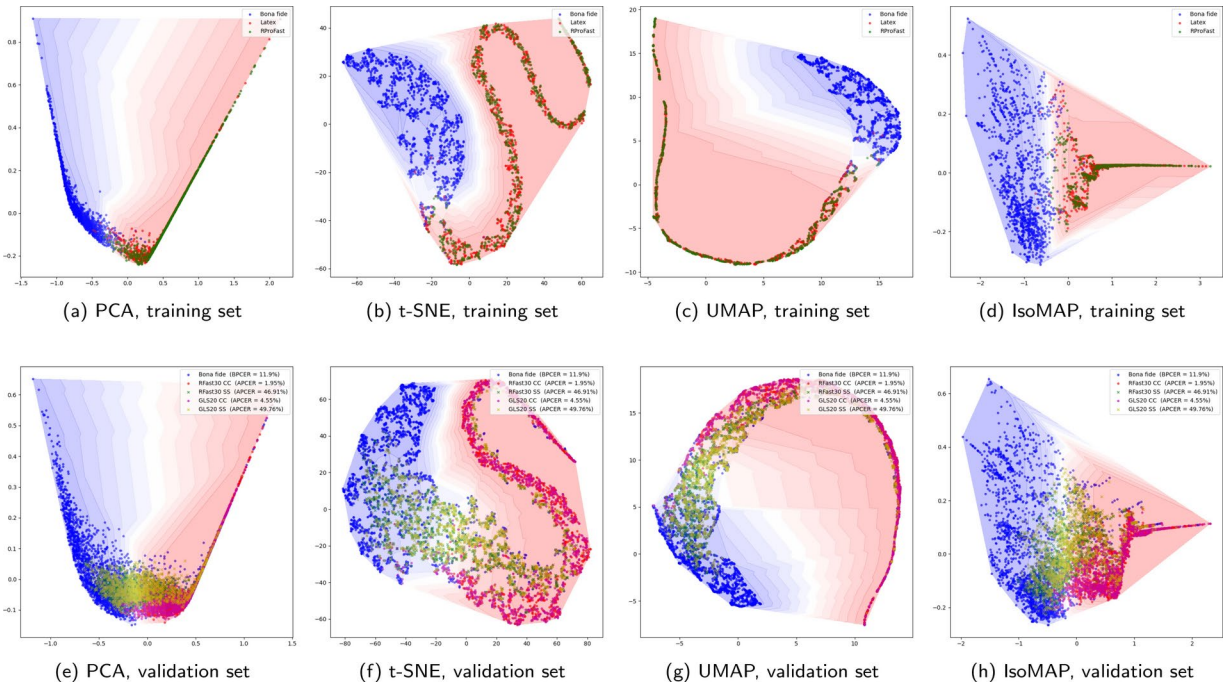


Fig. 8 Training set and validation set visualization for the Simple-CNN PAD (Dermalog scanner)

through geodesic distances. However, the separation is not as pronounced as with t-SNE or UMAP, particularly for overlapping or complex clusters. Unlike t-SNE and UMAP, IsoMAP may struggle with fine-grained local relationships within clusters due to its reliance on MDS for embedding, which emphasizes global structure over local density.

Examining the different PADs, the validation feature spaces also reveal that never-seen-before samples, such as those produced via the ScreenSpoof acquisition approach, tend to occupy areas that correspond to empty gaps close to the decision boundary. In particular, when these subsampled areas are into the BF cluster (such as for the Simple-CNN GreenBit case, Fig. 7 and the JLW Dermalog, Fig. 10), if the

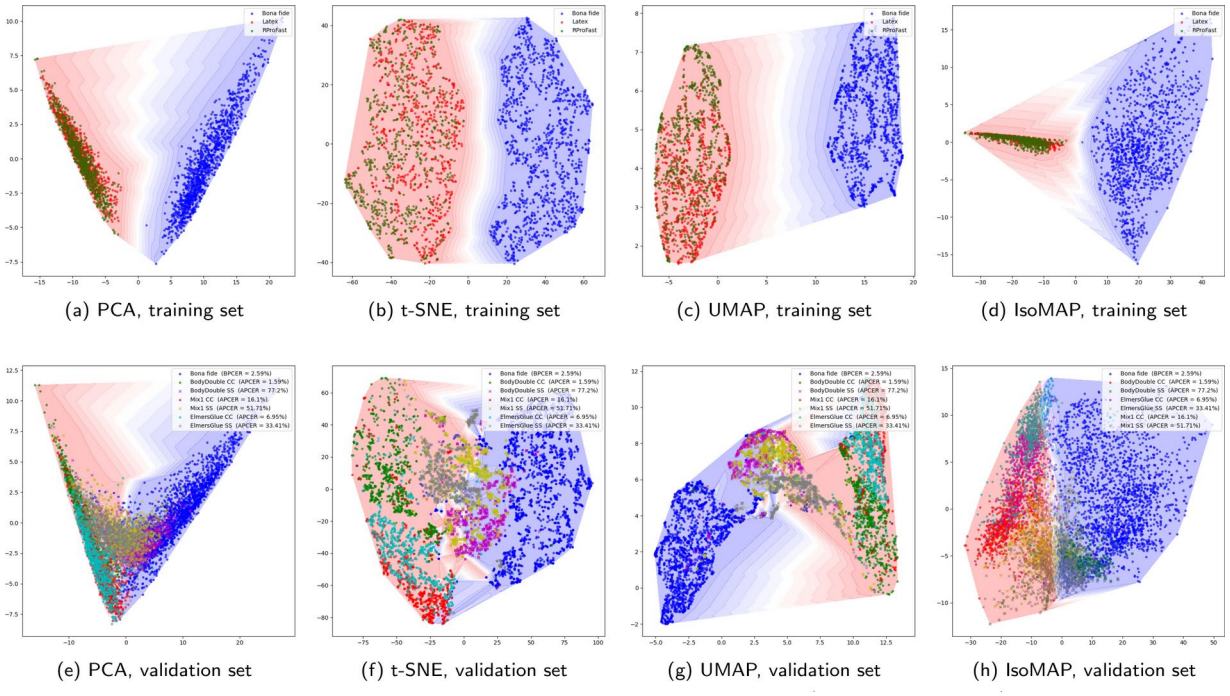


Fig. 9 Training set and validation set visualization for the JLWLivDetL PAD (Green Bit scanner)

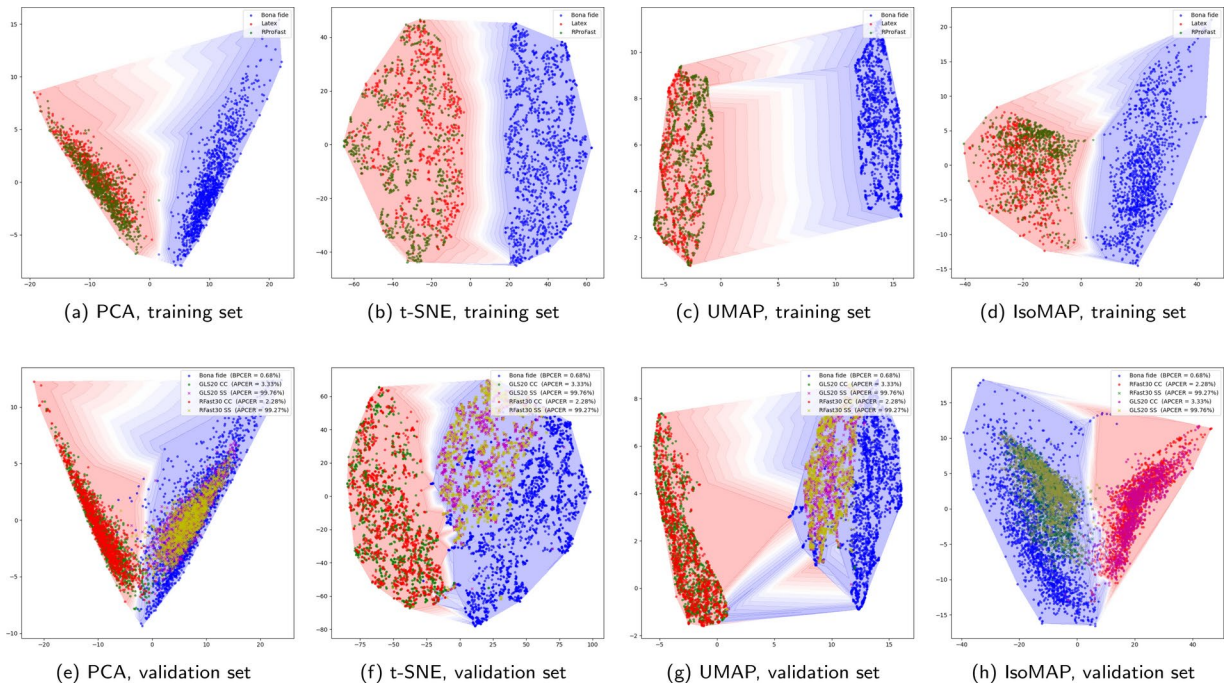


Fig. 10 Training set and validation set visualization for the JLWLivDetL PAD (Dermalog scanner)

PA test samples are positioned in these areas, these samples will be misclassified. On the other hand, since the PA class typically shows greater variability than the BF class, subsampled areas in the PA cluster are less problematic: these areas are more likely to be filled with new attack types than BF samples (as highlighted by the Megvii plots in Fig. 11 and the PADUnk Dermalog, Fig. 14).

We can, therefore, conclude that 2D reduced representations are useful for designing and interpreting PADs. In particular, because of its ability to maintain the global structure of high-dimensional data, PCA is generally more favorable

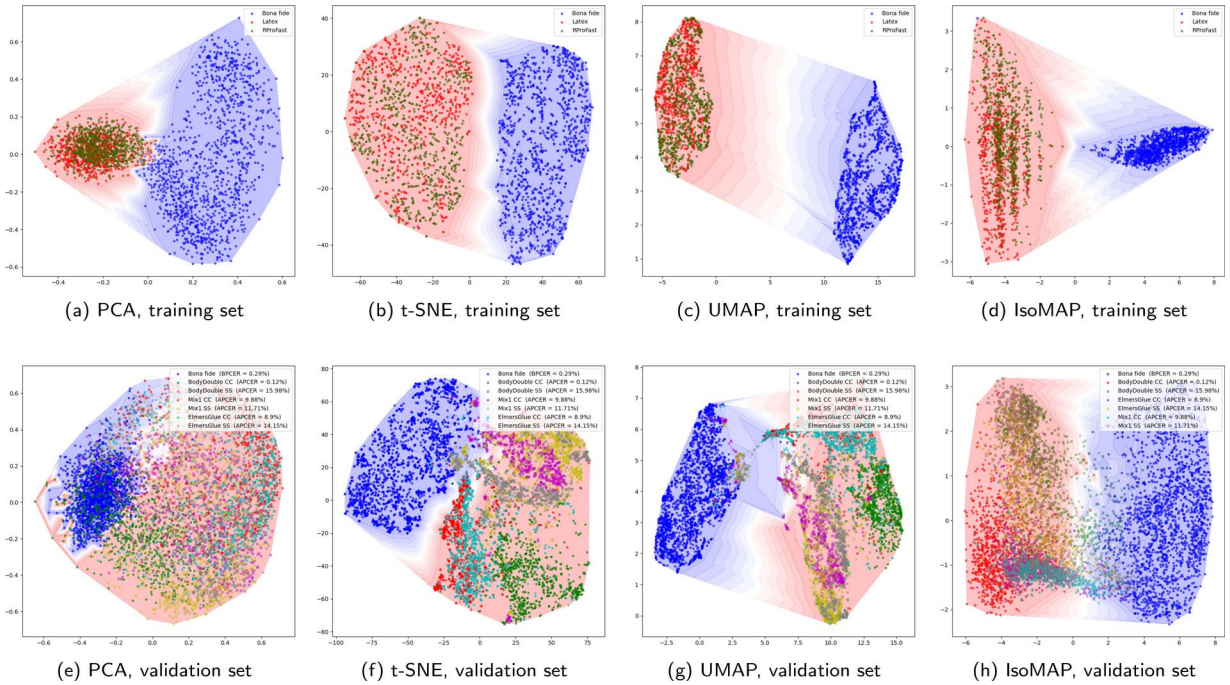


Fig. 11 Training set and validation set visualization for the Megvii PAD (Green Bit scanner)

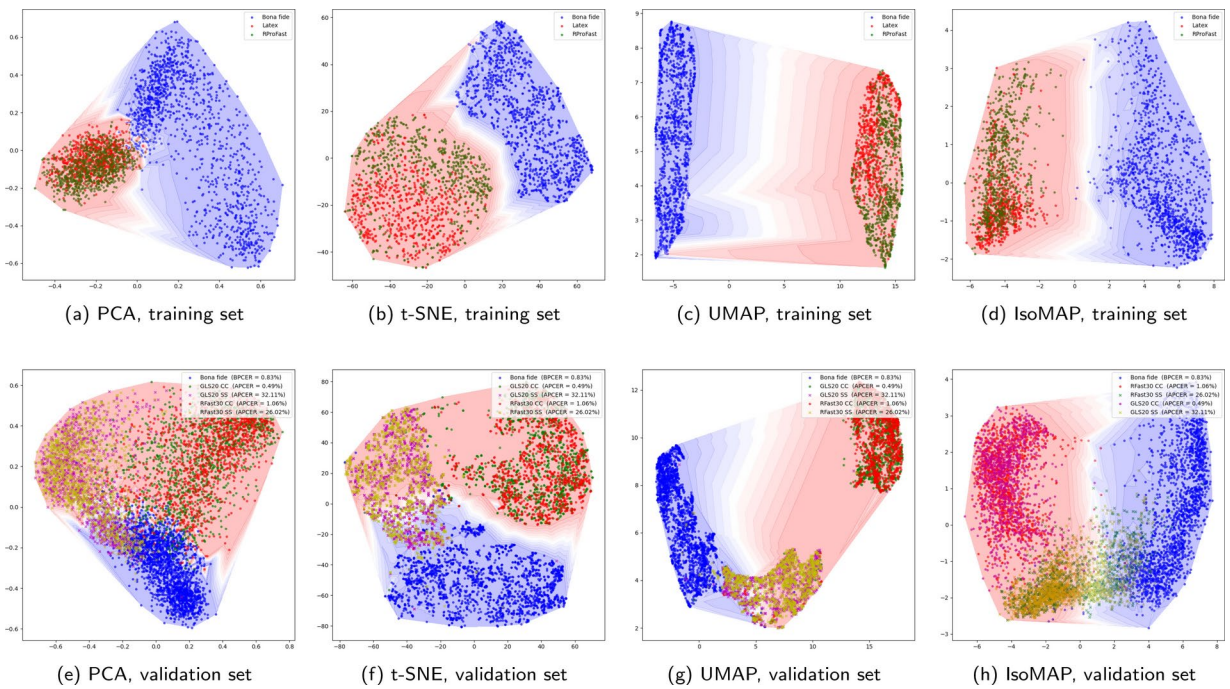


Fig. 12 Training set and validation set visualization for the Megvii PAD (Dermalog scanner)

when striving for precise measurement of distances between two samples or between a sample and a specific area within the feature space. t-SNE and UMAP, on the other hand, excel at detecting local and non-linear features, providing significant information on the distance between clusters and under-represented areas. IsoMAP serves as a middle ground, retaining a balance of global and local structures. It is most favorable when a combination of global geometry and meaningful cluster arrangement is required, but less ideal when the primary goal is to highlight fine-grained local variations or complex, non-linear cluster separations.

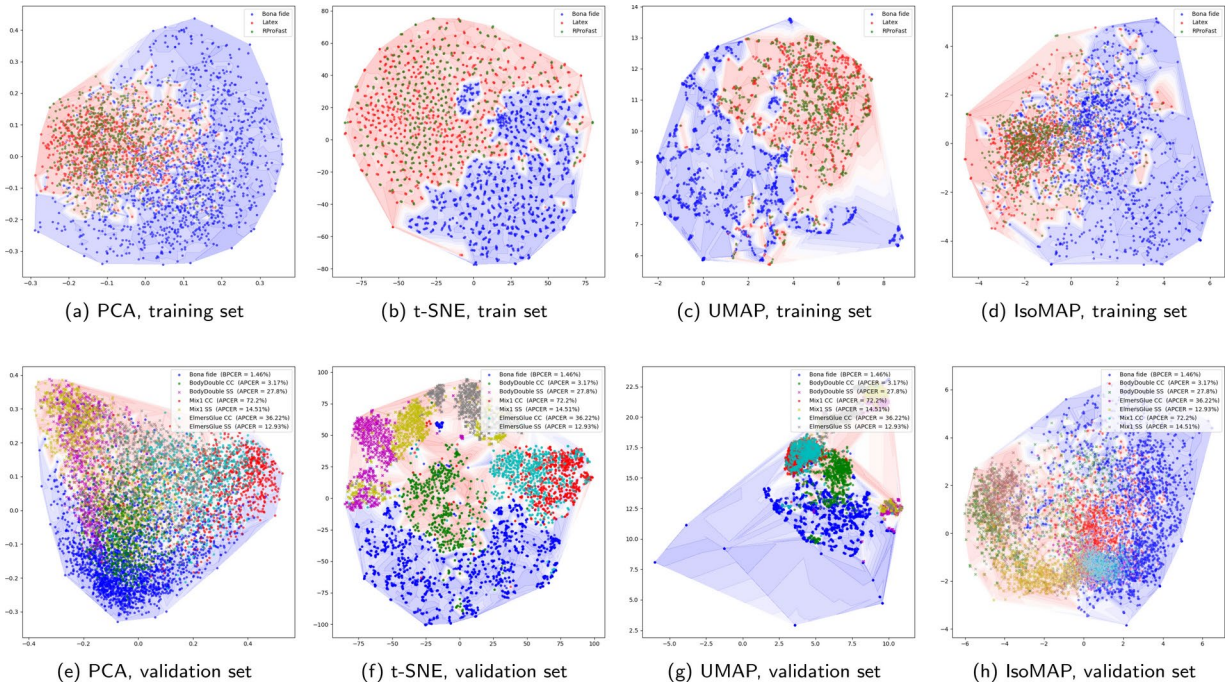


Fig. 13 Training set and validation set visualization for the PADUnk PAD (Green Bit scanner)

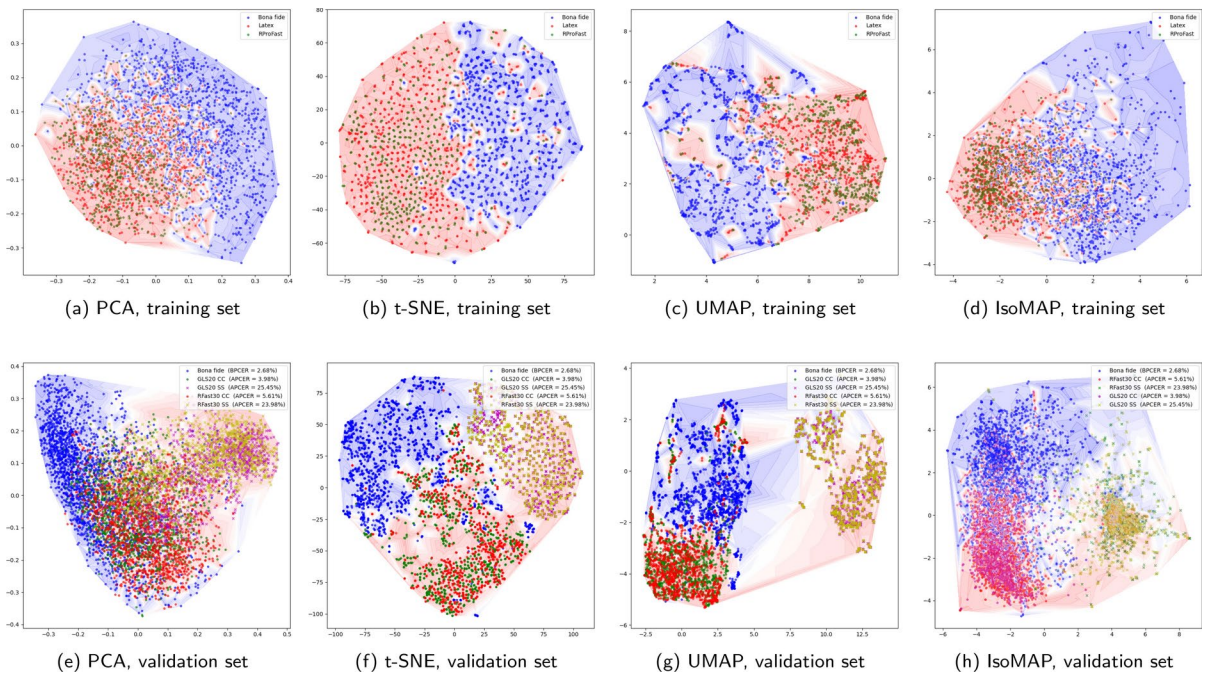


Fig. 14 Training set and validation set visualization for the PADUnk PAD (Dermalog scanner)

4.2.3 Editing impact simulation

Starting from the hypothesis that an extended learned sample space implies a better generalization ability of the PAD system, the next reported experiments aim to show how the system accuracy varies if a certain number of samples are removed from each cluster of the training set. Since we want to simulate under-represented areas, we eliminated the samples in the most populated areas.

We selected these samples using an Isolation Forest algorithm with one-hundred base estimators [44], trained and tested on the reduced training set. Isolation Forest was chosen for its ability to identify outliers in multi-dimensional datasets by exploiting the concept of isolation. This method

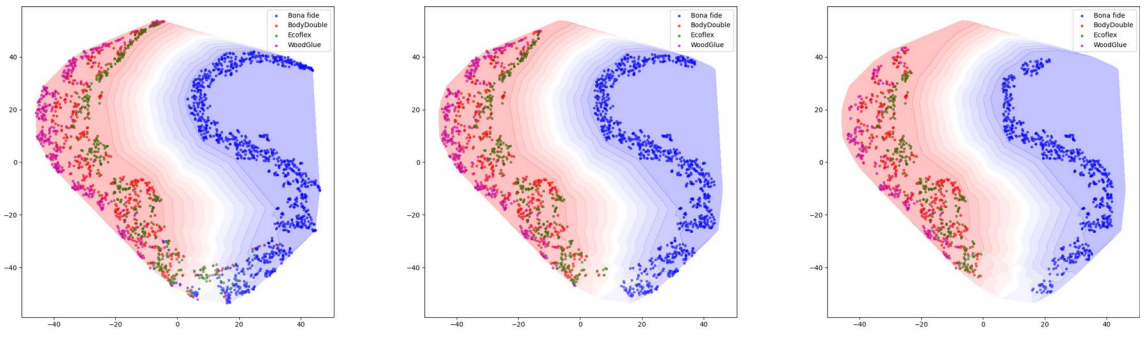
iteratively segments the feature space with random splits, and samples that are isolated most frequently are labeled as outliers. In our case, it is particularly useful because it does not make specific assumptions about the data distribution, which makes it suitable for datasets with complex distributions. Since non-linear reductions have proven to be more suitable for identifying under-represented areas, the following tests are carried out on the features reduced with t-SNE but the supplementary materials also investigate the other reductions.

First, we identified the outermost points of the distribution, and we applied three different thresholds for removal based on the negative anomaly score associated with them; thus, we chose to discard 33%, 66%, and 100% of the out-of-distribution samples. Furthermore, we also evaluate filtering only BF samples and only PA samples for these thresholds. We then apply the same procedure to the most densely populated areas of the cluster, by reversing the direction of the Isolation forest measure.

The experiments were conducted by training the white-box PAD, Simple-CNN, on a specific dataset (LivDet 2019, LivDet 2021, and LivDet 2023 training sets, either GreenBit or Dermalog) and testing on another set (LivDet 2021 test sets, either GreenBit or Dermalog). For each training set, we performed three independent training runs in order to account for variability and ensure robustness in the results. For the sake of space, we selected two cases to analyze: the model trained on LivDet 2019 (referred to as GreenBit2019), which leads to good accuracies on the chosen validation set, and the model trained on LivDet 2021 (referred to as GreenBit2021), which instead leads to high error rates. The other results are reported in the supplementary materials.

Let us first focus on the results of out-of-distribution filtering. Figure 15 illustrates the effect of this filtering on the GreenBit2019 model. The simulation of under-sampled areas reveals that when performance on the validation set is strong (Table 5), filtering out-of-distribution PA samples causes the model to struggle with classifying new attack types, as evidenced by the increase in APCER. This suggests that these PA samples, despite being isolated in the feature space, may contain critical variations essential for the model to recognize never-seen-before attacks. By removing them, we simulate a scenario where key attack characteristics are missing from the training data, which significantly hampers the model's generalization ability and leads to more errors. In contrast, removing out-of-distribution BF samples results in comparable or slightly improved accuracy. We can assume that out-of-distribution BF samples introduce noise or misleading traits, and their removal could help the model differentiate between classes more effectively. In fact, BF samples, by their nature, tend to form more compact clusters than PAs due to lower intra-class variability. As a result, filtering these outlier samples might allow the model to identify new PA samples accurately without associating those traits with the BF class. Thus, a designer must ensure that the PAD system is trained on a sufficiently dense and diverse sample space, particularly for PA samples, as any gaps in representation could result in a failure to detect novel attacks. Nevertheless, in the case of the less performing GreenBit2021 model (Table 6, the impact of under-represented areas in the feature space becomes evident. Out-of-distribution filtering, in this case, leads to an increase in APCER and a decrease in BPCER, resulting in an overall degradation of performance. The higher standard deviation further underscores the instability in decision-making, pointing to poorly defined decision boundaries. Figure 16 supports this observation, by showing a more complex and less separable distribution of samples, which complicates the classification task. Therefore, ensuring sufficient coverage of the feature space, especially in less performant models, appears crucial to avoiding such performance issues and improving the ability to discriminate between classes.

Lastly, the analysis of in-distribution sample filtering (Tables 7 and 8) reveals consistent performance degradation. In-distribution samples are essential for capturing



(a) Original (b) 33% filter (c) 100% filter

Fig. 15 LivDet 2019 training set visualization with out-of-distribution samples removal (GreenBit scanner, t-SNE reduction)

Table 5 Accuracy results (mean ± std dev) of the GreenBit2019 model on the LivDet 2021 test set according to out-of-distribution training set filt

GreenBit2019	Original	Percentage of out-of-distribution samples filtered								
		33%	33% BF	33% PA	66%	66% BF	66% PA	100%	100% BF	100% PA
#removed samples	0	228	108	120	455	222	233	689	331	358
#removed BF	0	108	108	0	222	222	0	331	331	0
#removed PA	0	120	0	120	233	0	233	358	0	358
Accuracy (%)	87.68	83.08 ± 2.3	$\mathbf{93.42 \pm 2.59}$	80.8 ± 2.45	85.45 ± 1.91	$\mathbf{94.29 \pm 0.75}$	77.82 ± 2	87.1 ± 2.42	$\mathbf{94.89 \pm 0.62}$	80.44 ± 2.28
APCER (%)	14.88	21.44 ± 3.67	$\mathbf{5.05 \pm 2.14}$	25.17 ± 4.81	14.93 ± 3.61	$\mathbf{4.15 \pm 1.77}$	32.15 ± 4.19	11.63 ± 3.38	$\mathbf{2.34 \pm 0.45}$	27.95 ± 4.96
BPCER (%)	6.73	7.64 ± 0.69	10.36 ± 1.69	7.19 ± 0.87	8.92 ± 0.66	11.42 ± 2.64	$\mathbf{4.67 \pm 0.32}$	11.72 ± 2.27	13.31 ± 1.96	$\mathbf{5.26 \pm 0.93}$

The results of the filtered trainings that are better than the original training are shown in bold

the core characteristics and intra-class variations necessary for discriminating between BF and PA. If gaps exist within these core clusters, the model lacks the knowledge to generalize effectively, leading to poorly defined decision boundaries. Following a similar trend to what was observed with out-of-distribution filtering, the degradation is generally less severe when filtering out BF samples than when filtering PA samples. Moreover, despite being a less performant model, GreenBit2021 exhibits a smaller drop in performance when filtering in-distribution samples compared to GreenBit2019, but shows a higher standard deviation, indicating that the model's decision-making is less stable and more inconsistent, especially when critical samples are missing. Figures 17 and 18 further illustrate how removing in-distribution samples reduces the density within the clusters.

5 Conclusions

In this paper, we conducted an extensive analysis to evaluate the ability of fingerprint presentation attack detection models to generalize to never-seen-before attacks or determine whether these models require fine-tuning with new bona-fide or presentation attack samples. Defining 'representativeness' in a geometric sense, we based the analysis on identifying gaps in the sample distribution within the feature space to address the generalization issue. Consequently, we emphasize the importance of appropriate feature space visualization techniques. Our study investigated three different two-dimensional visualization approaches: PCA, t-SNE, UMAP, and IsoMAP. The significance of precise parameterization was emphasized, particularly in non-linear approaches such as t-SNE and UMAP, to guarantee accurate and meaningful data representation.

We shed light on the strengths and limits of the different dimensionality reduction approaches, analyzing and comparing four distinct PAD models regarding interpretability. While PCA excelled at preserving the global structure of high-dimensional data, t-SNE and UMAP revealed local and non-linear properties, providing crucial insights into tightly packed data segments.

We also explored the consequences of manipulating the feature space population of a white-box model by arbitrarily removing some out-of-distribution or in-distribution samples from a specific training set and analyzing any increases or decreases in performance of the same model when retrained from scratch with such a reduced set. Our experiments have shown that removing BF samples might be generally more beneficial than removing PA if the base model already achieves good accuracy results. In particular, a PAD designer should be aware that removing out-of-distribution

Table 6 Accuracy results (mean ± std dev) of the GreenBit2021 model on the LivDet 2021 test set according to out-of-distribution training set filtering

GreenBit2021	Original	Percentage of out-of-distribution samples filtered							
		33%	33% BF	33% PA	66%	66% BF	66% PA	100%	100% BF	100% PA
#removed samples	0	370	156	214	739	304	435	1113	456	657
#removed BF	0	156	156	0	304	304	0	456	456	0
#removed PA	0	214	0	214	435	0	435	657	0	657
Accuracy (%)	77.45	67.34 ± 1.9	74.84 ± 2.09	69.18 ± 2.42	62.26 ± 2.25	71.39 ± 3.51	67.81 ± 5.32	58.7 ± 1.41	68.73 ± 2.6	68.46 ± 1.88
APCER (%)	45.24	62.71 ± 3.65	50.88 ± 3.45	51.88 ± 4.79	72.9 ± 3.33	53.27 ± 6.81	57.41 ± 9.93	76.12 ± 2.6	56.44 ± 8.33	61.85 ± 7.2
BPCER (%)	14.24	$\mathbf{11.75 \pm 1.32}$	$\mathbf{13.57 \pm 1.33}$	$\mathbf{12.96 \pm 0.5}$	$\mathbf{12.7 \pm 0.46}$	15.71 ± 1.2	$\mathbf{13.81 \pm 2.09}$	$\mathbf{13.34 \pm 1.35}$	16.28 ± 1.44	$\mathbf{11.92 \pm 1.77}$

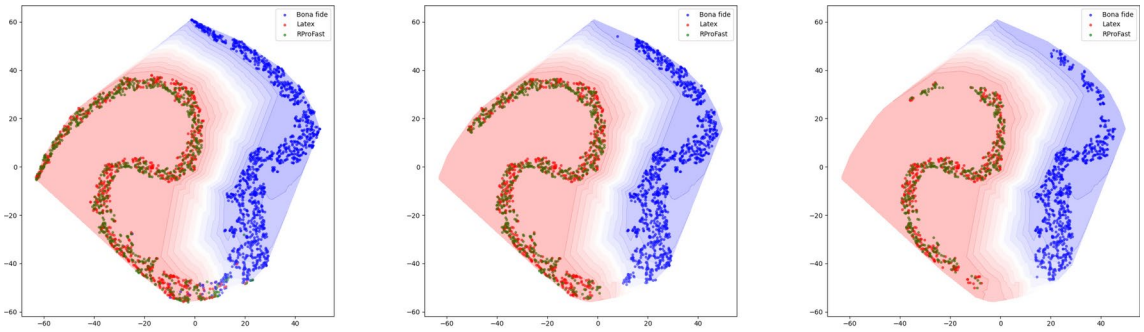
The results of the filtered trainings that are better than the original training are shown in bold

samples can reduce noise in the data, improving the model's ability to distinguish between classes, especially when intra-class features are well-defined. However, excessive filtering can lead to a risk of underfitting, where the model loses critical information needed to recognize samples belonging to less represented classes. This suggests that a proper balance between representative samples and outliers is crucial to maintain the model's ability to generalize.

To sum up, this first exploration allowed us to employ feature reduction techniques and derive some practical guidelines for assisting the interpretation of the results during the design process. We aim to go in-depth in order to suggest the designer what kind of PAI it needs to make the PAD more robust and if this need can be made achievable by the use of synthetic fake images.

6 Supplementary information

This article includes a supplementary file containing the feature space and accuracy results (reported as mean ± standard deviation) of the GreenBit2019, GreenBit2021, GreenBit2023, Dermalog2021, and Dermalog2023 models on the LivDet 2021 test set. These results are presented according to both out-of-distribution and in-distribution training set filtering conditions.



(a) Original (b) 33% filter (c) 100% filter

Fig. 16 LivDet 2021 training set visualization with out-of-distribution samples removal (Green Bit scanner, t-SNE reduction)

Table 7 Accuracy results (mean ± std dev) of the GreenBit2019 model on the LivDet 2021 test set according to in-distribution training set filtering

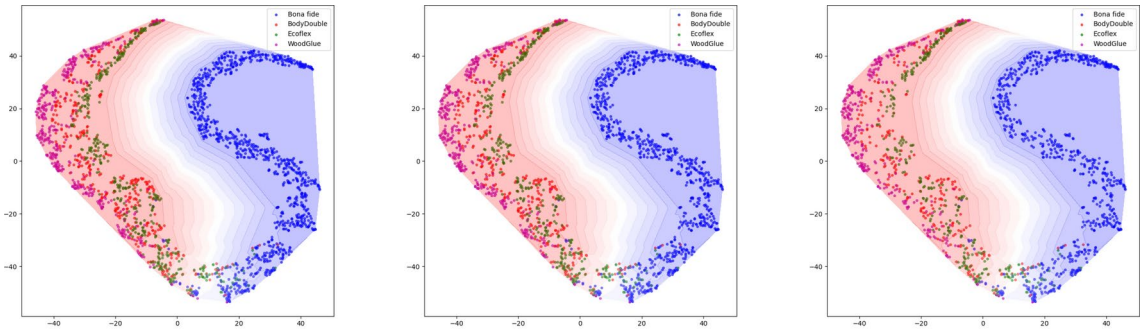
GreenBit2019	Original	Percentage of in-distribution samples filtered								
		33%	33% BF	33% PA	66%	66% BF	66% PA	100%	100% BF	100% PA
#removed samples	0	228	108	120	455	222	233	689	331	358
#removed BF	0	108	108	0	222	222	0	331	331	0
#removed PA	0	120	0	120	233	0	233	358	0	358
Accuracy (%)	87.68	80.47 ± 1.1	84.53 ± 2.79	81.53 ± 2.06	82.65 ± 2.5	85.17 ± 1.7	81.46 ± 3.58	83.39 ± 1.18	87.1 ± 3.53	81.22 ± 1.73
APCER (%)	14.88	25.02 ± 2.86	18.9 ± 5.45	26.07 ± 2.49	23.83 ± 3.02	19.1 ± 4.94	25.71 ± 4.7	23.15 ± 4.55	12.56 ± 2.62	26.93 ± 3.74
BPCER (%)	6.73	$\mathbf{5.58 \pm 1.1}$	$\mathbf{6.43 \pm 1.39}$	$\mathbf{5.91 \pm 1.39}$	$\mathbf{6.2 \pm 1.36}$	$5.85 \pm \mathbf{0.89}$	$\mathbf{5.79 \pm 2.08}$	$\mathbf{5.8 \pm 1.03}$	8.02 ± 1.4	$\mathbf{5.49 \pm 1.88}$

The results of the filtered trainings that are better than the original training are shown in bold

Table 8 Accuracy results (mean ± std dev) of the GreenBit2021 model on the LivDet 2021 test set according to in-distribution training set filtering

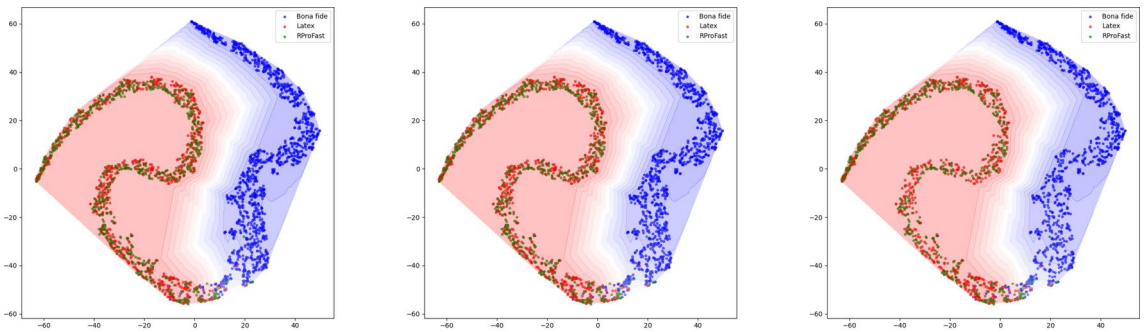
GreenBit2019	Original	Percentage of in-distribution samples filtered								
		33%	33% BF	33% PA	66%	66% BF	66% PA	100%	100% BF	100% PA
#removed samples	0	370	156	214	739	304	435	1113	456	657
#removed BF	0	156	156	0	304	304	0	456	456	0
#removed PA	0	214	0	214	435	0	435	657	0	657
Accuracy (%)	77.45	75.74 ± 3.16	$\mathbf{77.49 \pm 2.9}$	75.87 ± 2.64	75.53 ± 2.01	73.26 ± 5.16	74.58 ± 3.09	73.91 ± 2.98	75.21 ± 2.62	$\mathbf{78.61 \pm 4.82}$
APCER (%)	45.24	46.44 ± 5.32	$\mathbf{42.05 \pm 5.15}$	47.07 ± 7.28	48.8 ± 4.92	$\mathbf{43.85 \pm 9.98}$	49.56 ± 6	48.34 ± 7.08	45.39 ± 5.21	$\mathbf{41.02 \pm 14.65}$
BPCER (%)	14.24	$\mathbf{13.42 \pm 1.23}$	14.31 ± 1	$\mathbf{13.56 \pm 2.58}$	14.71 ± 0.82	19.18 ± 1.77	$\mathbf{13.78 \pm 1.17}$	16.68 ± 1.72	16.68 ± 0.75	15.51 ± 3.1

The results of the filtered trainings that are better than the original training are shown in bold



(a) Original (b) 33% filter (c) 100% filter

Fig. 17 LivDet 2019 training set visualization with in-distribution samples removal (Green Bit scanner, t-SNE reduction)



(a) Original (b) 33% filter (c) 100% filter

Fig. 18 LivDet 2021 training set visualization with in-of-distribution samples removal (Green Bit scanner, t-SNE reduction)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00138-025-01666-z.

Author contributions S.C., R.C., G.O., M.M. and G.L.M. conceived and planned the experiments. S.C., R.C. carried out the experiments. S.C., R.C., G.O. and M.M.. contributed to the interpretation of the results. All the authors wrote the manuscript. All authors reviewed the manuscript. G.L.M. was in charge of overall direction.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest We declare that there are no conflict of interest related to this research.

Ethical approval Furthermore, this study does not involve research with human participants and/or animals. Editorial Policies for: Springer journals and proceedings: https://www.springer.com/gp/editorial-policies Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies*Scientific Reports*: https://www.nature.com/srep/journal-policies/editorial-policies BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies.

References

1. Roy, A., Memon, N., Ross, A.: Masterprint: exploring the vulnerability of partial fingerprint-based authentication systems. IEEE Trans. Inf. Forensics Secur. **12**, 2013–2025 (2017). https://doi.org/10.1109/TIFS.2017.2691658
2. Marcel, S., Fierrez, J., Evans, N.: Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment Advances in Computer Vision and Pattern Recognition. Springer, Singapore (2023). https://doi.org/10.1007/978-981-19-5288-3
3. Engelsma, J.J., Jain, A.K.: Generalizing fingerprint spoof detector: learning a one-class classifier. In: 2019 International Conference on Biometrics (ICB), pp. 1–8 (2019). https://doi.org/10.1109/ICB45273.2019.8987319

4. Rattani, A., Scheirer, W.J., Ross, A.: Open set fingerprint spoof detection across novel fabrication materials. IEEE Trans. Inf. Forensics Secur. **10**, 2447–2460 (2015). https://doi.org/10.1109/TIFS.2015.2464772
5. Chugh, T., Jain, A.K.: Fingerprint presentation attack detection: generalization and efficiency. In: 2019 International Conference on Biometrics (ICB), pp. 1–8 (2019). https://doi.org/10.1109/ICB45273.2019.8987374
6. Kanj, S., Abdallah, F., Denœux, T., Tout, K.: Editing training data for multi-label classification with the k-nearest neighbor rule. Pattern Anal. Appl. **19**, 145–161 (2016). https://doi.org/10.1007/s10044-015-0452-8
7. Sorzano, C.O.S., Vargas, J., Montano, A.P.: A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877 (2014). arXiv:abs/1403.2877
8. Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality reduction: a comparative review. J. Mach. Learn. Res. **10**, 66–71 (2009)
9. Kobak, D., Berens, P.: The art of using t-sne for single-cell transcriptomics. Nat. Commun. **10**, 5416 (2019). https://doi.org/10.1038/s41467-019-13056-x
10. Chugh, T., Jain, A.K.: Fingerprint spoof detector generalization. IEEE Trans. Inf. Forensics Secur. **16**, 42–55 (2021). https://doi.org/10.1109/TIFS.2020.2990789
11. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
12. Sousedik, C., Busch, C.: Presentation attack detection methods for fingerprint recognition systems: a survey. IET Biom. **3**, 219–233 (2014)
13. Micheletto, M., et al.: Review of the fingerprint liveness detection (livdet) competition series: from 2009 to 2021. In: Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment, pp. 57–76 (2023)
14. Singh, J.M., Madhun, A., Li, G., Ramachandra, R., Yildirim Yayilgan, S., Bajwa, I.S., Sanfilippo, F. (Eds.): A survey on unknown presentation attack detection for fingerprint. In: Yildirim Yayilgan, S., Bajwa, I.S., Sanfilippo, F. (Eds.) Intelligent Technologies and Applications, pp. 189–202. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71711-7_15
15. Shaheed, K., et al.: Deep learning techniques for biometric security: a systematic review of presentation attack detection systems. Eng. Appl. Artif. Intell. **129**, 107569 (2024)
16. Joshi, I., et al.: Synthetic data in human analysis: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **46**, 4957–4976 (2024). https://doi.org/10.1109/TPAMI.2024.3362821
17. Grosz, S.A., Jain, A.K.: Spoofgan: synthetic fingerprint spoof images. IEEE Trans. Inf. Forensics Secur. **18**, 730–743 (2023). https://doi.org/10.1109/TIFS.2022.3227762
18. Brasil Vieira Wyzykowski, A., Jain, A.K.: Synthetic latent fingerprint generator. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 971–980 (2023)
19. Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. **2**, 559–572 (1901). https://doi.org/10.1080/14786440109462720
20. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
21. Jain, A.K., Deb, D., Engelsma, J.J.: Biometrics: trust, but verify. IEEE Trans. Biom. Behav. Identity Sci. **4**, 303–323 (2022). https://doi.org/10.1109/TBIOM.2021.3115465
22. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**, 66 (2019)
23. Ismail, A.A., Feizi, S., Bravo, H.C.: Improving deep learning interpretability by saliency guided training. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (2024)
24. Ghorbani, A., Berenbaum, D., Ivgi, M., Dafna, Y., Zou, J.Y.: Beyond importance scores: interpreting tabular ml by visualizing feature semantics. Information (2022). https://doi.org/10.3390/info13010015
25. Sequeira, A.F., Silva, W., Pinto, J.R., Gonçalves, T., Cardoso, J.S.: Interpretable biometrics: Should we rethink how presentation attack detection is evaluated? In: 2020 8th International Workshop on Biometrics and Forensics (IWBF), pp. 1–6 (2020). https://doi.org/10.1109/IWBF49977.2020.9107949
26. Sumithra, V., Surendran, S.: A review of various linear and non linear dimensionality reduction techniques. Int. J. Comput. Sci. Inf. Technol. **6**, 2354–2360 (2015)
27. Cunningham, J.P., Ghahramani, Z.: Linear dimensionality reduction: survey, insights, and generalizations. J. Mach. Learn. Res. **16**, 2859–2900 (2015)
28. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000). https://doi.org/10.1126/science.290.5500.2319
29. Sharma, R.P., Dey, S.: A comparative study of handcrafted local texture descriptors for fingerprint liveness detection under real world scenarios. Multimed. Tools Appl. **80**, 9993–10012 (2021). https://doi.org/10.1007/s11042-020-10136-9
30. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: uniform manifold approximation and projection. J. Open Source Softw. **3**, 861 (2018). https://doi.org/10.21105/joss.00861
31. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000). https://doi.org/10.1126/science.290.5500.2319
32. Torgerson, W.S.: Multidimensional scaling: I. Theory and method. Psychometrika **17**, 401–419 (1952)
33. Wattenberg, M., Viégas, F., Johnson, I.: How to Use t-sne Effectively. Distill (2016)
34. Zhang, Y., et al.: A score-level fusion of fingerprint matching with fingerprint liveness detection. IEEE Access **8**, 183391–183400 (2020). https://doi.org/10.1109/ACCESS.2020.3027846
35. Zhang, Y., et al.: Slim-rescnn: a deep residual convolutional neural network for fingerprint liveness detection. IEEE Access **7**, 91476–91487 (2019). https://doi.org/10.1109/ACCESS.2019.2927357
36. Deng, J., et al.: Imagenet: A Large-scale Hierarchical Image Database, pp. 248–255 (2009)
37. González-Soler, L.J., et al.: Local feature encoding for unknown presentation attack detection: an analysis of different local feature descriptors. IET Biom. **10**, 374–391 (2021). https://doi.org/10.1049/bme2.12023
38. Orrù, G., et al.: Livdet in action—fingerprint liveness detection competition 2019. In: 2019 International Conference on Biometrics (ICB), pp. 1–6 (2019)
39. Casula, R., et al.: Livdet 2021 fingerprint liveness detection competition—into the unknown. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–6 (2021). https://doi.org/10.1109/IJCB52358.2021.9484399
40. Micheletto, M., et al.: Livdet2023—fingerprint liveness detection competition: advancing generalization. In: 2023 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8 (2023)
41. Casula, R., et al.: Are spoofs from latent fingerprints a real threat for the best state of art liveness detectors. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 3412–3418 (2021). https://doi.org/10.1109/ICPR48806.2021.9413301
42. 30107-3:2023(E): I. Information technology " Biometric presentation attack detection " Part 3: Testing and reporting. Standard, International Organization for Standardization (2023)

43. Kobak, D., Linderman, G.C.: Initialization is critical for preserving global data structure in both t-sne and umap. Nat. Biotechnol. **39**, 156–157 (2021). https://doi.org/10.1038/s41587-020-00809-z
44. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422 (2008). https://doi.org/10.1109/ICDM.2008.17

Simone Carta received his BSc Degree in Electrical, Electronic, and Computer Engineering from the University of Cagliari in 2020. He also received, with honors, hisMScDegree in Computer Engineering, Cybersecurity, and Artificial Intelligence from the University of Cagliari in 2023, discussing a thesis entitled "Increasing the Generalization Ability of FingerprintPresentation Attack Detection Systems by an Explainability-based Framework". Since 2023, he is currently a PhD student at the PRA Lab.

Roberto Casula obtained a PhD degree in Electronical andComputer Engineering from the University of Cagliari (Italy) in 2023, discussing a thesis called "The Art of Fingerprint Spoofing". Since November 2015 he has been collaborating with the PRA Labin the field of fingerprint spoofing and fingerprint liveness detection. He is currently an Assistant Professor with the Pattern Recognition and Applications Laboratory (PRA Lab), at the Departmentof Electrical and Electronic Engineering (DIEE). His research interests include fingerprint spoofing, fingerprint liveness detection, deepfake detection and analysis, and crowd detection and analysis.

Giulia Orrù received he Ph.D. degree in Electronical and Computer Engineering from the University of Cagliari in 2021. She is currently Assistant Professor of Computer Engineering atthe University of Cagliari, Italy. In 2014 she joined the research group on Pattern Recognition and Applications Laboratory (PRA lab) at the Dept. of Electrical and Electronic Engineering (DIEE),working on pattern recognition and its applications, specifically on biometric recognition, presentation attack detection systems and adaptive biometric systems.

Marco Micheletto received his Ph.D. degree in Electronical and Computer Engineering from the University of Cagliari, Italy, in 2023. He is currently an Assistant Professor with the PatternRecognition and Applications Laboratory (PRA Lab), at the Department of Electrical and Electronic Engineering (DIEE). His research interests include integration of fingerprint comparison systemswith presentation attack detectors, fingerprint liveness detection, electroencephalography signal processing for biometric and behavioral purposes, and the analysis and detection of deepfaketechnologies.

Gian Luca Marcialis received hisPh.D. degree in Electronic Engineering and Computer Science from the University of Cagliari, Italy, in 2004. He is currently Associate Professor at University of Cagliari, and Research director of theBiometric Unit of the Pattern Recognition and Applications Laboratory - PRA Lab at the Department of Electrical and Electronic Engineering. His research interests are in the fields of biometrics,namely, fingerprint presentation attack detection, fingerprint classification, multiple classifiers for biometric identification, self update-based biometric systems, facial deepfake detection andanomaly group behavior in crowds.