IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 April 2024, accepted 17 April 2024, date of publication 25 April 2024, date of current version 3 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393558

APPLIED RESEARCH

Explainable Vision Transformers for Vein Biometric Recognition

ROCCO ALBANO[1], (Student Member, IEEE), LORENZO GIUSTI[2], EMANUELE MAIORANA[1], (Senior Member, IEEE), AND PATRIZIO CAMPISI[1], (Fellow, IEEE)

[1]Department of Industrial, Electronic and Mechanical Engineering, Roma Tre University, 00146 Rome, Italy
[2]Department of Computer, Control and Management Engineering, Sapienza University of Rome, 00185 Rome, Italy

Corresponding author: Rocco Albano (rocco.albano@uniroma3.it)

This work was supported in part by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART" - Spoke 4 - Cascade Call Project "EXperience and Privacy for Extended RealiTy" - "EXPERT").

ABSTRACT In the field of deep learning, understanding the rationale behind an automatic system's decisions is essential for building users' trust and ensuring accountability. In this regard, explainable artificial intelligence (XAI) recently emerged as a valuable tool to offer insights into a model behavior. The present study focuses on vein-based biometric recognition, investigating techniques allowing to identify which regions of a wrist-vein image are mostly exploited to carry out a verification process. Toward this aim, our research exploits vision transformers (ViTs), which rely on self-attention mechanisms to automatically detect and exploit the input parts with the content deemed most relevant for its further processing. Two distinct wrist-vein pattern datasets, namely PUT-wrist and FYO-wrist, are employed to fine-tune the considered models. Their behavior is interpreted by analyzing the attention maps generated when applying the trained networks to vein-pattern images, investigating which regions are exploited to decide a user's identity. The proposed approach testifies that the performed recognition process can improve when a ViT focuses on areas with significant vein pattern content, achieving verification performance surpassing state-of-the-art methods in open-set scenarios, while promoting transparency through explainability.

INDEX TERMS Biometric recognition, vein biometrics, wrist vein biometrics, explainable AI, vision transformers.

I. INTRODUCTION

Biometric recognition systems leverage individuals' unique physiological or behavioral attributes to perform people identification or verification, revolutionizing several applications with security-related requirements [1]. One notable sub-domain of this research field regards hand vein patterns that can be captured by exposing palms, wrists, or fingers to imaging systems relying on infrared radiation and allow the recognition of individuals based on the unique characteristics of the acquired subcutaneous traits [2], [3]. Renowned for its robustness against spoofing attacks [4], vein recognition stands out as a highly secure biometric modality [5]. Traditionally, vein recognition has been dominated by feature engineering and conventional algorithms like local binary pattern (LBP) and Gabor filters [6], [7], [8], [9]. The advent of deep learning pushed the performance of vein-based biometric recognition systems, making convolutional neural networks (CNNs) the standard choice for extracting hierarchical features from raw vein images [10], [11], [12]. However, the black-box nature of these models still represents a downside affecting their reliability. Nowadays, there is a growing demand for transparent systems that may empower users to comprehend why specific decisions are made, thereby fostering confidence in their fairness and unbiasedness [13], [14], [15].

Within the context of vein pattern recognition, the need to gain insights into the aspects considered by deep learning approaches when producing their decisions is highly relevant. Traditional approaches commonly relied on segmentation processes to generate vessel skeletons that were then employed for recognition, thus offering a solid basis for

The associate editor coordinating the review of this manuscript and approving it for publication was Carmelo Militello.

which characteristics were used for recognition. Given the lack of investigations on their interpretability, an analogous awareness still needs to be made available for deep learning approaches. Here, we analyze these aspects, prioritizing explainability as a path towards interpretability and trying to shed some light on the features that mainly contribute to producing a decision in vein-based recognition systems relying on deep learning approaches.

In more detail, we here resort to vision transformers (ViTs), a key paradigm that recently emerged within the landscape of deep learning architectures [16], to reach the desired goal. Differently from CNNs, ViTs perform vision tasks by dividing images into sequences of non-overlapping patches, which are then fed into transformer blocks [17]. In addition to being often able to learn more discriminative features compared to alternatives such as CNNs, ViTs represent an approach allowing *in-model* explainability, that is, integrating explainability mechanisms within the learning process by resorting to attention-based processes [18]. Such an approach differs from *post-model* explainability, which involves analyzing the model after it has been trained, with in-model paradigms providing end-to-end transparency, allowing one to understand not only the final decisions but also the process that led to those decisions, and typically providing greater generalization and robustness, thus enhancing trust in the model [19].

While early research investigated the potential of ViTs for vein-based biometric recognition [20], a comprehensive study elucidating robust explainability mechanisms tailored for ViTs and applied to this task is still missing in the literature. The present paper represents a step in this direction, analyzing the attention maps generated by ViTs applied to vein images when performing user verification. In detail, the state of the art on the specific modality here considered, i.e., wrist vein patterns and the use of ViTs for biometric recognition and explainability, is presented in Section II. The ViT-based framework used for wrist-vein biometric recognition is described in Section III. The setup employed in the performed experimental tests is outlined in Section IV, while the obtained results are presented in Section V. Some conclusions are eventually drawn in Section VI.

II. RELATED WORKS

As already mentioned, vein-based biometric recognition was historically approached by resorting to feature engineering and conventional machine learning algorithms, before CNNs kicked in and significantly outperformed traditional processing techniques [21], [22], [23], [24]. With specific regard to wrist-vein biometric recognition [25], early attempts focused on extracting the vessel patterns in the treated images to derive representations based on handcrafted characteristics fed to classic machine learning algorithms. For instance, support vector machines (SVMs) with LBP inputs were used in [26] to train a classifier, then used as feature extractors in a closed-set verification system, achieving an equal error rate (EER) at 1.3% when comparing samples from different acquisition sessions of the PUT wrist-vein database [27]. Such verification performance was obtained by performing tests on the same subjects employed to train the SVM classifier, with limited generalizability capacity. On the other side, open-set verification conditions were considered on [28], where the comparison between vein patterns was performed using correlations, as well as in [29], where scale-invariant feature transform (SIFT) characteristics were extracted from vein patterns and compared to estimate the similarity between samples. However, the verification rates achieved in open-set scenarios, where recognition performance is computed over subjects other than those used for training the employed solution or no training is required, are typically worse but more generalizable than those achievable in closed-set scenarios. In fact, EERs at 9.3% and 15.9% were respectively obtained in [28] and [29] on the PUT database. Handcrafted features were also employed to perform wrist-vein identification in [30], where rank-1 accuracies at 84.0% and 93.1% were respectively achieved on the PUT and FYO [31] datasets.

Deep learning approaches were instead applied to wrist-vein images in [32], where an EER=2.1% on PUT was obtained in closed-set verification when training a ResNet152 network [33], then employed to extract features used as input to a further logistic regression classifier. Closed-set conditions were also considered in [34], where a lightweight network was designed to extract discriminative wrist-vein features, achieving EERs at 1.2% on PUT and 1.84% on FYO when testing on the same subjects employed to train the used model. Open-set verification conditions have been instead considered in [35], where a siamese approach was employed to train a CNN, achieving an F1 score at 84.7%. Also, ViTs were applied to wrist-vein images, as in [20] where an accuracy at 99.5% on PUT was obtained in identification, yet not investigating any explainability aspect nor the generalizability of ViT on verification tasks.

In general, while deep learning approaches notably outperformed classic machine learning algorithms for wrist-vein biometrics in identification scenarios they are still not as efficient for verification, where networks should be employed as feature extractors. Furthermore, deep learning reliability for wrist-vein verification in open-set scenarios must be properly explored. Given that these latter conditions are closer to real-life verification, in which a solution is designed to be optimized over a certain dataset yet then applied to different subjects, we here resort to open-set experimental scenarios to conduct our analysis in the hope of deriving more general outcomes than what closed-set conditions could allow.

In more detail, within the context of explainable artificial intelligence (XAI), our intent here is to shed some light on the decision-making processes of deep-learning-based wrist-vein biometric verification, thus trying to demystify the operations of approaches relying on neural networks and understand their rationale [36]. Unfortunately, XAI applications in biometric recognition are still at an infant stage [37], with most of the studies so far presented mainly focused on presentation attack detection (PAD) for facial

biometrics [38]. Among the investigations that tried to look into the aspects exploited by neural networks applied to biometric recognition during their decision-making process, the vast majority relied on post-model methods. These approaches, essentially visual toolkits, elucidate predictions of an already trained model [39]. A notable technique in this category is the gradient-weighted class activation mapping (Grad-CAM) approach, prominently featured in studies analyzing network decisions [40]. The only work devoted explicitly to the explainability of vein recognition [41], to the best of our knowledge, actually adopted such post-model paradigm, employing the local interpretable model-agnostic explanations (LIME) method [42] to visually dissect a custom CNN areas of focus on palm vein images. However, the obtained visual results are coarse and hardly highlight areas resembling the vein patterns in the images.

Within the XAI taxonomy, our approach instead fits into the functional methods according to which, to achieve better explainability, it is preferable to consider methods that jointly provide predictions and explanations. In this regard, ViTs achieved remarkable results for vision tasks. Specifically, ViTs segment images into fixed-size patches and process them through multiple layers of self-attention mechanisms, capturing long-range dependencies [16], [17]. Given their ability to capture complex patterns, ViTs have already been exploited for biometric recognition, especially for gait [43], [44]. Regarding vein biometrics, in addition to [20] where ViTs were applied to wrist-vein images, in [45] and [46] the authors have applied ViTs to finger-vein traits, while a multi-scale transformer was applied on palm-vein data in [47]. Yet, all the studies mentioned above only used ViTs to extract discriminative features from the considered images, with no reference to the associated explainability aspects and without investigating the derived attention maps.

Although there is an ongoing debate about the effectiveness of generic attention mechanisms as explainability tools [48], [49], there is also a consensus on the efficacy of ViT attention maps in providing relevant insights on models explainability for visual tasks [50], [51]. This kind of analysis proved effective in other fields such as bio-medicine [52]. For instance, ViT attention maps provided relevant insights for explainable COVID-19 screening in [18]. A comprehensive study in [53] also delves into explainable transfer learning for ViTs applied to chest X-rays. Similar works within the biometric recognition field are still rare, with a notable example presented in [54], where the authors employed ViTs to guide fingerprint embedding using minutiae matching, and emphasized the decision process of ViTs through salience maps. The present study aims at providing an analogous contribution considering wrist-vein biometrics, trying to achieve high recognition performance and provide an insightful and transparent understanding of the underlying biometric features that drive accurate verification. The present study aims at providing an analogous contribution considering wrist-vein biometrics, trying to achieve high recognition performance and provide an insightful and transparent understanding of the underlying biometric features that drive accurate verification.

III. EXPLAINABLE VEIN BIOMETRIC RECOGNITION

Initially crafted for natural language processing tasks, ViTs [16] recently emerged as a groundbreaking approach in computer vision, outperforming traditional CNNs in a variety of vision tasks, especially when dealing with large-scale data [55]. The core concept of ViTs can be mathematically distilled into the following components.

A. TOKENIZATION

Analogous to the tokenization of words in text, ViTs dissect images into fixed-size and non-overlapping patches. An image I of dimensions $H \times W \times C$ is divided into patches of size $P \times P \times C$, resulting in $N = \frac{H \cdot W}{P \cdot P}$ patches. Each patch is subsequently flattened and linearly embedded into a vector $\mathbf{u}_{ij} \in \mathbb{R}^d$. Formally, if $\mathbf{v}_{ij} \in \mathbb{R}^{P^2 \cdot C}$ denotes the vector corresponding to the patch at row i and column j of the original image, we have $\mathbf{u}_{ij} = \mathbf{v}_{ij}\mathbf{W}_e$, where $\mathbf{W}_e \in \mathbb{R}^{(P^2 \cdot C) \times d}$ is a learnable embedding matrix. Furthermore, a special <cls> token is appended to the beginning of this sequence, serving as an aggregate representation for downstream tasks, especially classification.

B. POSITIONAL EMBEDDING

By design, the transformer architecture does not provide any order or position that captures the original context provided by the spatial arrangement of the patches. To overcome this limitation, a learnable positional embedding vector $\mathbf{p}_{ij}$ is added to its corresponding token, with $\mathbf{x}_{ij} = \mathbf{u}_{ij} + \mathbf{p}_{ij}$ being the final patch embedding infused with positional information.

C. TRANSFORMER LAYERS

To capture both local and global information from an image I, a ViT receives as input the sequence $\mathbf{X} \in \mathbb{R}^{(1+N^2) \times d}$ composed of a vector $\mathbf{u}_{\text{cls}}$ that contains the embedding of the <cls> token, followed by the sequence of the patch embeddings $\mathbf{x}_{ij}$. Each layer consists of a self-attention mechanism [56] followed by a feed-forward network. For a single head of attention, the attention weights α and output o_{ij} are computed as:

$$\begin{aligned} \mathsf{Q} &= \mathbf{X}\,\mathbf{W}_Q \\ \mathsf{K} &= \mathbf{X}\,\mathbf{W}_K \\ \mathsf{V} &= \mathbf{X}\,\mathbf{W}_V \\ \alpha &= \text{softmax}\left(\frac{\mathsf{Q}\mathsf{K}^T}{\sqrt{d}}\right) \\ \mathsf{O} &= \alpha\mathsf{V}, \end{aligned}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ are learnable weight matrices. The attention maps indicate which areas of the original image are highlighted during the process and mostly used for the desired task, that is, classification during the training phase. After L ViT layers, the output corresponding to the

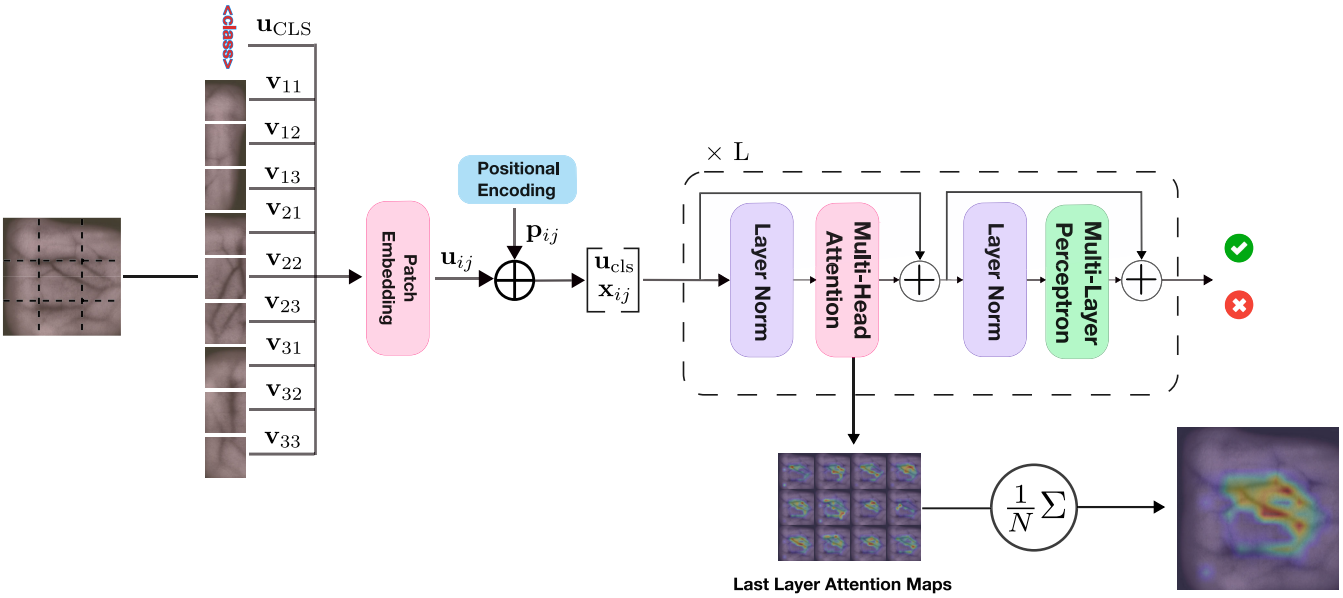


FIGURE 1. Illustrative representation of the Vision Transformer (ViT) [16] tailored for explainable vein biometric recognition.

<cls> token, denoted as o_{cls}, is processed by a linear layer for classification as $\hat{y} = \text{softmax}(\mathbf{o}_{\text{cls}}\mathbf{W}_{\text{cls}} + \mathbf{b}_{\text{cls}})$. Cross-entropy is typically employed as a loss function over the computed probabilities to drive network training using the back-propagation algorithm.

D. ATTENTION MAPS

As previously illustrated, attention maps provide information about which regions of the input image are of particular interest to the model. These maps serve as a compass, guiding our understanding of where the model focuses when making a prediction. To construct such maps in the context of ViTs, one typically examines the attention weights, specifically from the last layer of the transformer. The rationale behind this choice is that the latter layers capture higher-level, more abstract features that directly influence the model's final decision. To provide an output that is robust to the model fluctuations, we consider a ViT equipped with multiple attention heads. Each head assigns distinct attention weights to different input regions during its operation. To derive the final attention maps, we average the attention weights across all heads. Mathematically, for a multi-head attention mechanism with h heads, the final attention map $\mathcal{A}$ can be thus formulated as:

$$\mathcal{A} = \frac{1}{h}\sum_{i=1}^{h}\mathbf{A}_i^{(L)} \tag{1}$$

where $\mathbf{A}_i^{(L)}$ is the attention weight matrix for the i-th head at the last transformer layer L. Once obtained, this averaged attention map is upsampled and overlaid on the original image to visualize the regions that the ViT paid attention to during the biometric recognition task. Such visualizations not only serve as an explanatory tool, demystifying the ViT behavior, but also play a pivotal role in diagnosing potential biases or shortcomings in the model focus and, by extension, its decision-making process. A visual depiction of the processing applied to the considered vein pattern images is shown in Figure 1. In contrast, the details of the processing performed to derive the desired attention maps are given in Algorithm 1.

It is worth noticing that the ViT attention weight matrices, and therefore the attention maps, are computed before deriving the final inner representation of the input or performing classification, differently from class-dependent post-model explainability tools such as Grad-CAM. Therefore, attention maps may provide insights on the most relevant parts of an image even when a ViT is used only as a feature extractor, as it happens in our case when verification, and not identification, is carried out in the experimental tests.

IV. EXPERIMENTAL SETUP

The datasets used in the performed tests are presented in Section IV-A, while the employed training and testing strategies are respectively outlined in Sections IV-B and IV-C.

A. DATASET

We applied ViTs to the wrist-vein patterns in two public datasets, namely PUT [27] and FYO [31].

The PUT wrist-vein database comprises 1200 images, with acquisitions taken from the right and left hands of 50 subjects. A total of 4 images were captured for each wrist during 3 acquisition sessions separated in time. Given the low correlation between each subject's right and left wrist-vein patterns, each wrist is considered a class for a total of 100 distinct classes. The samples within this dataset experience low contrast between the vein traits and the background, being, therefore, particularly hard to be processed effectively. To enhance the visibility of

Algorithm 1 Get Attention Maps

Input: Image I of dimensions $H \times W \times C$
Output: Predicted class $\hat{y}$, Attention Map $\mathcal{A}(I)$
Require: Patch size P

procedure Tokenization(I: Image)
  $\mathbf{P} =$ Divide I into a gird of N patches. ▷ $N = \frac{H \times W}{P \times P}$
  $\mathbf{v}_i = vec(\mathbf{P}_{i,j}) \quad \forall i, j$ ▷ Seq. of $\mathbb{R}^{P^2 \cdot C}$ vectors
  $\mathbf{u}_i = \mathbf{v}_i \cdot \mathbf{W}_e.$ ▷ $\mathbf{W}_e \in \mathbb{R}^{(P^2 \cdot C)} \times d$
return u. ▷ Patch Embeddings $\in \mathbb{R}^{N \times d}$
end procedure
procedure Pos. Embedding($\mathbf{u}$: patch embeddings)
  $\mathbf{x}_i = \mathbf{u}_i + \mathbf{p}_i \quad \forall i$
  $\mathbf{X}^{(0)} = [\mathbf{u}_{\text{cls}}, \mathbf{x}_1, \ldots, \mathbf{x}_N]$
return $\mathbf{X}^{(0)}$
end procedure
procedure Self Attention($\mathbf{X}^{(l)}$)
  $\mathbf{Q}^{(l)} = \mathbf{X}^{(l)} \cdot \mathbf{W}_Q^{(l)}$
  $\mathbf{K}^{(l)} = \mathbf{X}^{(l)} \cdot \mathbf{W}_K^{(l)}$
  $\mathbf{V}^{(l)} = \mathbf{X}^{(l)} \cdot \mathbf{W}_V^{(l)}$
  $\mathbf{A}^{(l)} = \text{softmax}\left(\frac{\mathbf{Q}^{(l)} \cdot (\mathbf{K}^{(l)})^\top}{\sqrt{d}}\right).$ ▷ Attention Maps
  $\mathbf{O}^{(l)} = \mathbf{A}^{(l)} \cdot \mathbf{V}^{(l)}$
return $\mathbf{O}^{(l)}, \mathbf{A}^{(l)}$
end procedure
$\mathbf{u} =$ Tokenization(I)
$\mathbf{X}^{(0)} =$ embedding($\mathbf{u}$)
for $l = 1$ to L **do**
  $\hat{\mathbf{X}}^{(l)} = \text{LN}(\mathbf{X}^{(l)})$ ▷ Layer Norm.
  $\mathbf{O}^{(l)}, \mathbf{A}^{(l)} = \text{MHA}(\hat{\mathbf{X}}^{(l)})$ ▷ Multi-head Attention.
  $\tilde{\mathbf{X}}^{(l)} = \mathbf{X}^{(l)} + \mathbf{O}^{(l)}$
  $\mathbf{X}^{(l+1)} = \tilde{\mathbf{X}}^{(l)} + \text{FF}(\text{LN}(\tilde{\mathbf{X}}^{(l)}))$ ▷ 2-Layer MLP.
end for
$\mathcal{A} = 1/h \sum_h \mathbf{A}_h^{(L)}.$
return $\mathcal{A}$

TABLE 1. Employed wrist-vein dataset characteristics.

Characteristic	FYO	PUT-wrist
Subjects	160	50
Traits	2	2
Samples per Session	1	4
Sessions	2	3
Total Images	640	1200
Train, Validation, and Test Split		
Train Class	260	75
Train Images/Class	25	120
Validation Class	260	75
Validation Images/Class	5	12
Test Class	60	25
Test Images/Class	2	12

the traits, contrast limited adaptive histogram equalization (CLAHE) [57] is typically applied to the images of this dataset. CLAHE is a widely used image enhancement method that adjusts the contrast of an image by redistributing pixel intensities, characterized by two parameters, i.e., *clip limit* and *grid size*, here, respectively set at 5.0 and (8, 8). The results of this preprocessing to images from the PUT database can be seen in Figure 2.

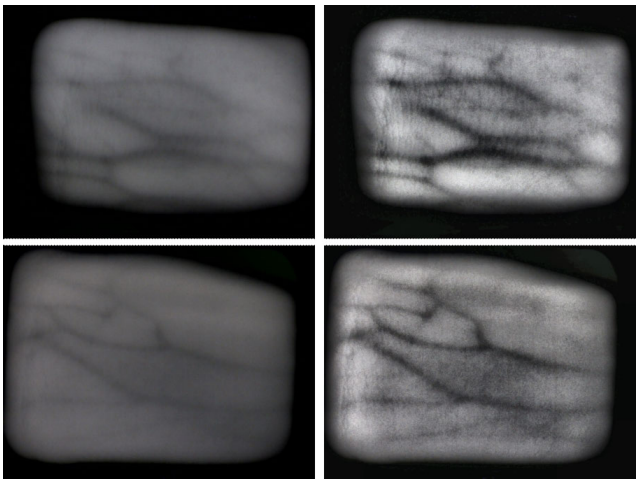


FIGURE 2. Original wrist-vein patterns (left) vs their CLAHE-enhanced representations (right) from the PUT-Wrist vein dataset.

The FYO wrist-vein database contains images taken from both hands of 160 subjects, collected using a medical vein finder in a controlled environment, with an image captured during each of two separate acquisition sessions for each participant. As for PUT, the vein patterns from the two wrists of each subject are considered as two separate classes, for a total of 320 classes. Given that images in FYO have better quality than those in PUT, applying CLAHE to the images in this dataset is typically not required. A summary of the main characteristics of the employed datasets is given in Table 1.

Given the limited amount of data in both databases, data augmentation was employed to increase the number of samples available to train the employed ViT models. In more detail, we leveraged a suite of basic image processing operations to generate augmented instances for each class, thus enhancing the diversity of the training samples. The augmentations were applied as follows: (1) *Horizontal Flip* with a probability $p = 0.5$; (2) *Rotation* with a probability $p = 0.7$ bounded by a maximum left rotation of $-10°$ and a maximum right rotation of $+10°$; (3) *Random Contrast* adjustments, initiated with a probability $p = 0.5$, with a factor range between 0.7 and 1.3. (4) *Random Illumination* variations, activated with a probability $p = 0.5$, regulated by a factor ranging from 0.7 to 1.3. The PUT dataset was augmented from 12 to 132 images per class, while the FYO dataset from 2 to 30 images per class.

B. ARCHITECTURE AND TRANSFER LEARNING

Two different ViT configurations were used in the performed tests, the first exploiting patch sizes of $P \times P = 16 \times 16$ and the second with patch sizes of $P \times P = 32 \times 32$. This choice was made to evaluate whether the employed patch size affects the model recognition performance in this task.

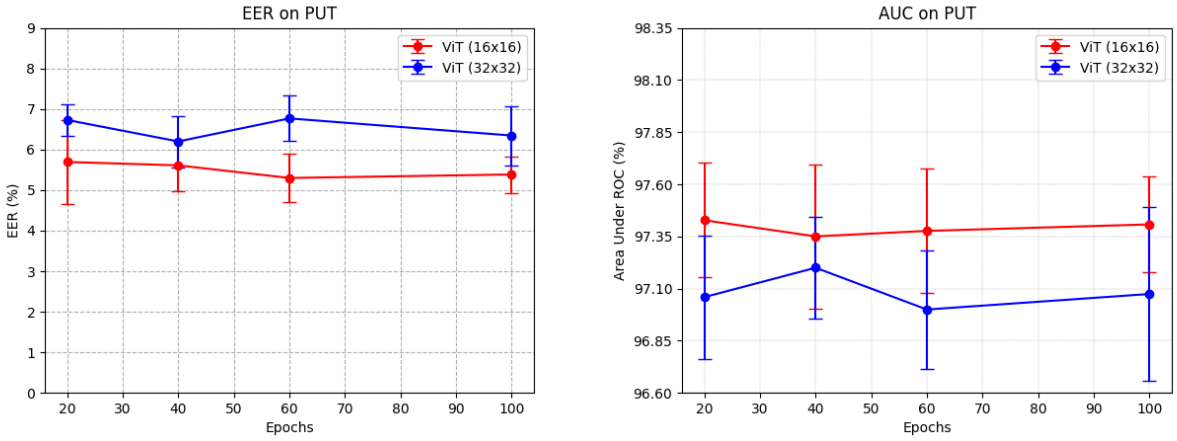


FIGURE 3. In the left graph, one can see the trend of EER obtained from the two models at different training epochs while on the right is the evolution of AUC for the PUT dataset.

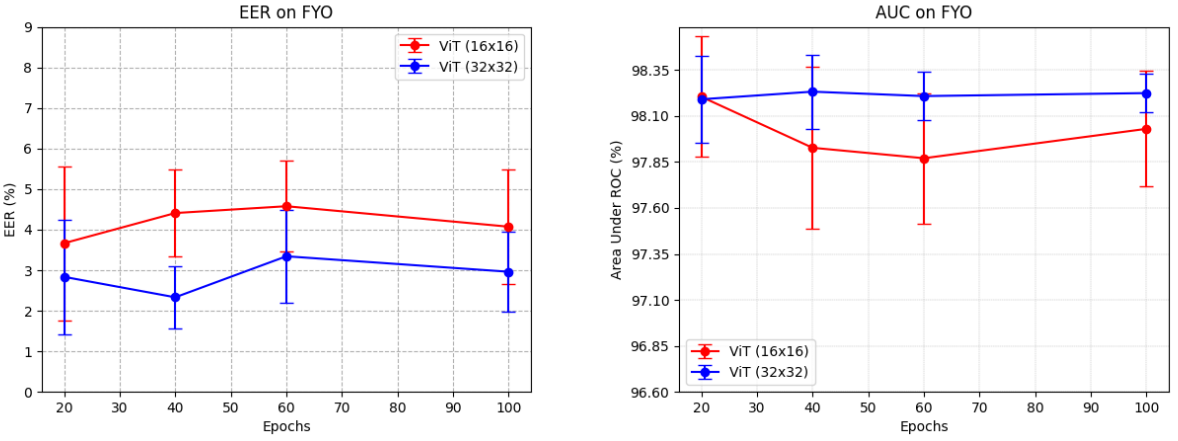


FIGURE 4. In the left graph, one can see the trend of EER obtained from the two models at different training epochs, while on the right is the evolution of AUC for the FYO dataset.

The imported architectures were pretrained on the ImageNet-1K dataset, which contains 1 million images labeled in 1,000 categories, thus providing a solid knowledge base for our model. The used ViTs are characterized by $h = 12$ heads, feeding inputs to a multi-layer perceptron (MLP), producing representations with 768 coefficients to model more complex relationships among the extracted features before performing classification. For regularization purposes, we replaced the last linear layer of the ViT model with a combination of two linear layers, interspersed with a dropout layer, to increase its generalization capability.

To investigate the properties of networks with different discriminative characteristics, we fine-tuned the considered ViTs on the available wrist-vein datasets varying the number of training epochs, namely 20, 40, 60, and 100. As an optimizer, we used stochastic gradient descent (SGD) [58] with momentum at 0.9, a starting learning rate of 0.05, and cosine decay without an initial warm-up. For all the experiments, the batch size was set to 128. All tests were performed on an 4 x NVIDIA® Tesla V100 GPUs with 5,120 CUDA cores and 32GB GPU memory, on a personal computing platform with an Intel® Xeon® Gold 5218 CPU @ 2.30GHz CPU using Ubuntu 18.04.6 LTS. The model was implemented in PyTorch [59] by building on top of TIMM library [60]. PyTorch, NumPy, SciPy, and Joblib are available under the BSD and Matplotlib under the PSF licenses. TIMM is available under the Apache 2.0 license.

C. TESTING

As mentioned in our tests, we considered a verification scenario in which a user asks the system to be recognized based on comparing a probe sample to the data provided during an enrolment phase. To assess the generalizability of ViTs and their suitability for real-world scenarios, open-set conditions were considered when estimating the achievable recognition performance. In more detail, we performed a 10-fold cross-validation by excluding, at each iteration, 25 classes from PUT and 60 from FYO when training the employed ViTs. After a ViT is fine-tuned, it is applied as a feature extractor to the images belonging to the classes excluded from training. The representations thus generated are then compared to compute pairwise 1-vs-1

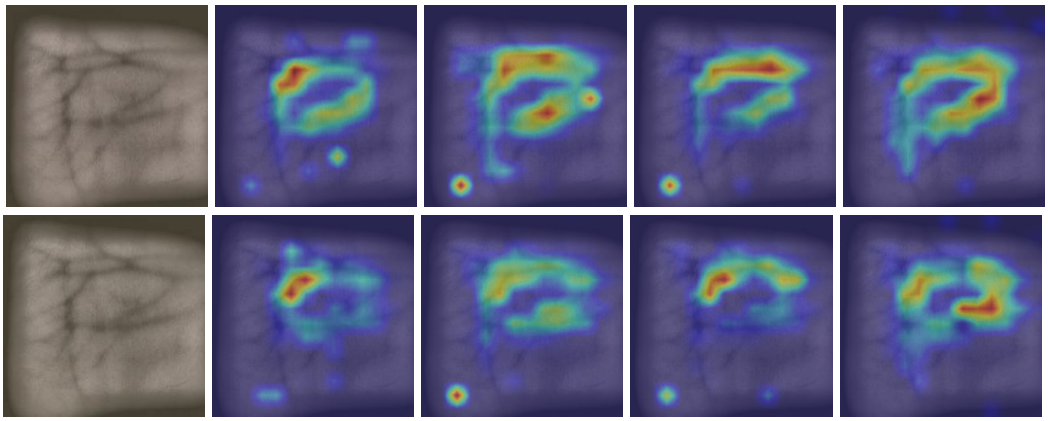


FIGURE 5. Attention maps of ViTs trained for different numbers of epochs, for two different images of the same wrist from the PUT dataset. Warmer colors signify higher attention concentration. Column (a) shows the original images, column (b) the attention maps obtained after fine-tuning for 20 epochs, column (c) 40 epochs, column (d) 60 epochs, column (e) 100 epochs.

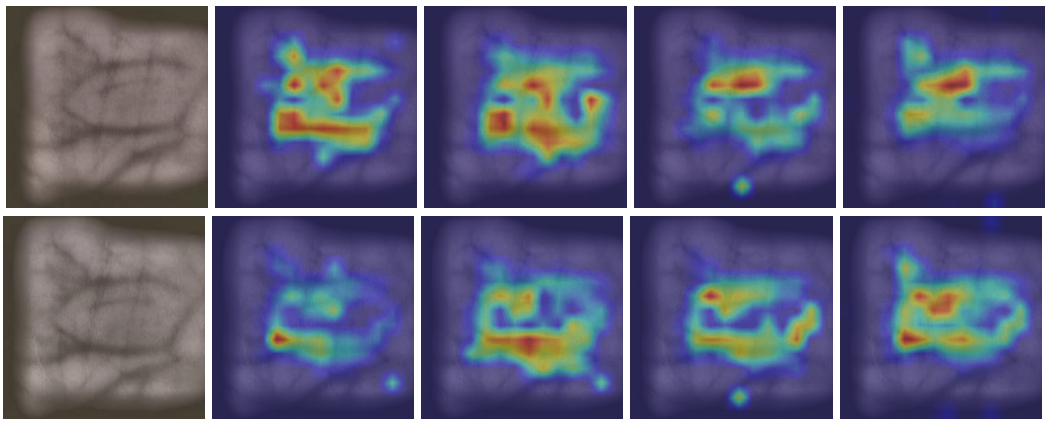


FIGURE 6. Attention maps of ViTs trained for different numbers of epochs for two images of the same wrist from the PUT dataset. Warmer colors signify higher attention concentration. Column (a) shows the original images, column (b) the attention maps obtained after fine-tuning for 20 epochs, column (c) 40 epochs, column (d) 60 epochs, column (e) 100 epochs.

similarity scores so that distributions of genuine scores are created by comparing images belonging to the same class. In contrast, the distributions of impostor scores are obtained by comparing samples belonging to different test subjects. Setting a threshold and comparing it against the obtained scores allows us to decide whether the pairwise comparison samples stem from the same user. Only the original images, not the augmented ones, are included in the test set to estimate the achievable performance.

V. EXPERIMENTAL RESULTS

As mentioned in the previous section, tests were conducted by fine-tuning the considered ViT models for increasing numbers of epochs, and adopting the trained networks as feature extractors on a disjoint set of subjects. Figures 3 and 4 show the results obtained in the performed 10-fold cross-validation tests, in terms of mean and standard deviation equal error rate (EER), with ViT models trained for different numbers of epochs. Also, the trend of the area under the curve (AUC) of the receiver operating characteristic (ROC), obtained by plotting (1-FRR) vs FAR, is reported. As for the PUT database, the results in Figure 3 are referred to images processed with CLAHE, which is instead not employed for FYO since it does not produce improvements, given that the original samples are already characterized by proper contrast and sharpness.

TABLE 2. Recognition performance comparison between approaches based on ViTs and ResNet152 on the PUT-Wrist dataset.

DATASET	MODEL	CLAHE	EER (%)	AUC (%)
PUT	ViT-16	✓	5.2 ± 0.6	97.4 ± 0.3
		X	10.2 ± 1.3	94.3 ± 0.7
	ResNet152	✓	15.5 ± 2.1	92.5 ± 2.1
		X	16.1 ± 3.5	90.7 ± 3.1
FYO	ViT-32	X	2.3 ± 0.7	98.2 ± 0.2
	ResNet152		5.0 ± 1.6	98.9 ± 0.5

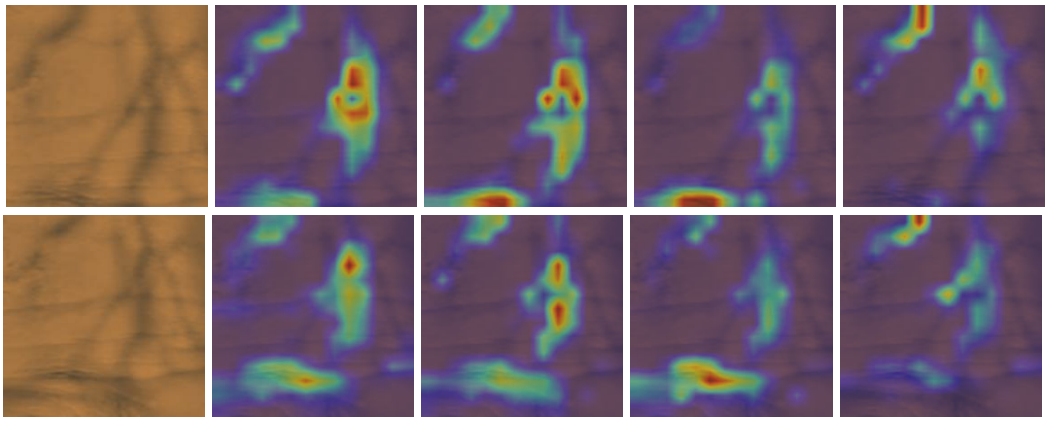


FIGURE 7. Attention maps of ViTs trained for different numbers of epochs, for two different images of the same wrist from the FYO dataset. Warmer colors signify higher attention concentration. Column (a) shows the original images, column (b) the attention maps obtained after fine-tuning for 20 epochs, column (c) 40 epochs, column (d) 60 epochs, column (e) 100 epochs.

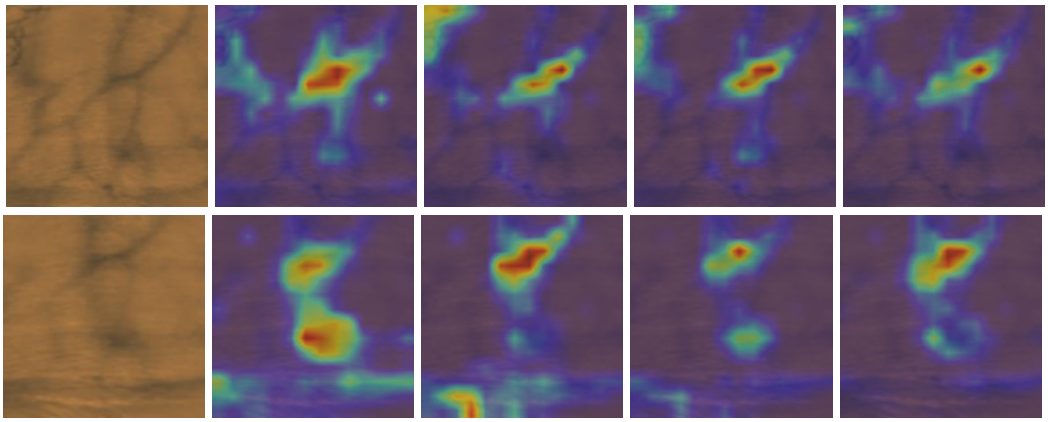


FIGURE 8. Attention maps of ViTs trained for different numbers of epochs, for two different images of the same wrist from the FYO dataset. Warmer colors signify higher attention concentration. Column (a) shows the original images, column (b) the attention maps obtained after fine-tuning for 20 epochs, column (c) 40 epochs, column (d) 60 epochs, column (e) 100 epochs.

The ViT configuration resulting in the best values of average EER and AUC for the PUT database relies on 16×16 patches and fine-tuned for 100 epochs. As for the FYO dataset, the ViT configuration leading to the lowest average EER and highest average AUC uses a patch size of 32×32 and fine-tuned for 40 epochs. The obtained results, therefore, testify that the selected patch size may significantly impact the achievable recognition rate, especially for the FYO dataset, whose better image quality allows it to perform better than PUT. In more detail, the best average EERs obtained in open-set verification is 5.2% for PUT and 2.3% for FYO.

Figures 5 and 6 show the attention maps associated with distinct images of two subjects from the PUT database, created by ViTs fine-tuned for different numbers of epochs. It is worth remarking that such maps refer to subjects used during performance evaluation and, therefore, are not included in the training dataset employed to fine-tune the ViTs. As can be seen, regions with higher relevance (warmer colors) significantly overlay with areas containing wrist vessels, thus testifying the effectiveness of the considered solutions and also their generalizability, since they are able to localize relevant vessel regions also on images from subjects other than those seen during fine-tuning. Furthermore, it can also be appreciated that ViTs trained for more significant numbers of epochs are more effective since they reduce the spurious regions on which attention is placed.

The attention maps associated with distinct images of two subjects from the FYO database, created by ViTs fine-tuned for different numbers of epochs, are instead shown in Figures 7 and 8. As for the images from PUT, the attention maps properly focus on areas containing relevant vessel contents, which are, therefore, those mostly considered to decide on the presented subjects.

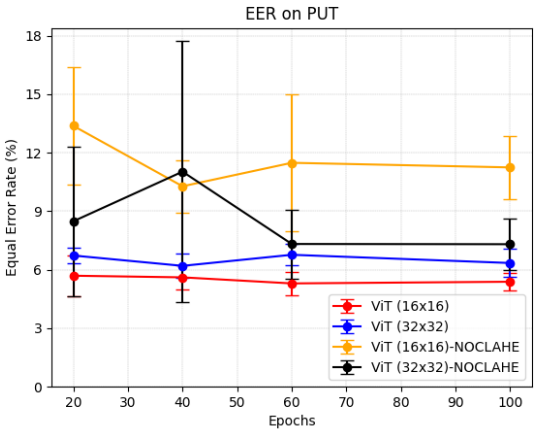
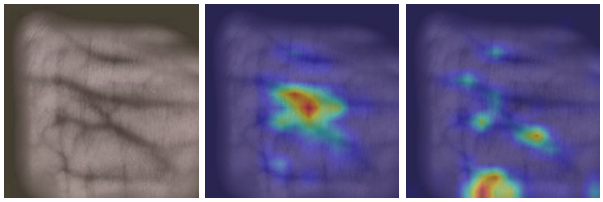


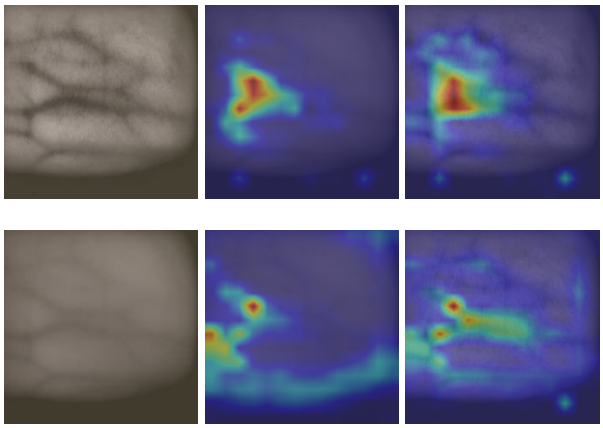
FIGURE 9. EERs of ViT models trained on PUT wrist vein images either with or without CLAHE preprocessing.



(a) (b) (c)

FIGURE 10. Attention maps for images with or without CLAHE. Column (a) shows an image from PUT preprocessed with CLAHE, column (b) the corresponding attention map created by a ViT-16 model trained on CLAHE images, column (c) the attention map created on the original image by a ViT-16 model trained on original images.

Additional tests were performed to assess the effectiveness and the effects of using CLAHE on images from the PUT database. Figure 9 reports the performance achievable on PUT when training ViTs on images either preprocessed or not with CLAHE. The best average EER achieved when avoiding CLAHE is 7.3%, with a notable worsening from the 5.2% obtained with CLAHE. Moreover, Figure 10 offers a visual comparison of the effects of CLAHE on the images processed by ViTs. The maps there reported show that, due to the limited contrast in the original images, networks trained on data not preprocessed with CLAHE might find it hard to pay attention to areas with relevant vessel content, with consequences on the achievable recognition performance. Fine-tuning ViTs on images with enhanced contrast may also be beneficial when applying the trained network on not-enhanced images. Specifically, models fine-tuned on images preprocessed with CLAHE produce more informative attention maps when applied to both the original and enhanced images of subjects, as shown in Figure 11. Conversely, leveraging the original images for fine-tuning does not allow ViT to focus appropriately on the most relevant areas of the inputs when enhanced images are fed to the trained network. Resorting to attention maps for explainability also offers interesting insights into the effects of CLAHE on PUT wrist-vein images.



(a) (b) (c)

FIGURE 11. Effects of CLAHE on ViTs trained on PUT. Column (a) shows an image from PUT with (top) and without (bottom) CLAHE preprocessing. Column (b) shows the attention maps obtained applying the trained models on the original image, while column (c) shows the attention maps obtained from the CLAHE-preprocessed images. For columns (b) and (c), the first row refers to ViTs trained on the CLAHE-preprocessed PUT dataset, the second row to ViTs trained without CLAHE.

Eventually, for comparison purposes, we report in Table 2 the recognition performance achievable in open-set verification with the proposed approach based on ViT, and with a state of the art approach relying on ResNet152 [33], as proposed in [32]. As for tests with ViTs, a 10-fold cross validation was performed by fine-tuning, at each iteration, a pretrained ResNet152 on a subset of the considered vein databases, and then using the trained model as feature extractor for the disjoint dataset upon which the achievable verification rates are estimated. The models are trained for 1000 training epochs using early stopping with a patience of 10 epochs. As optimizer, we used SGD with a moment of 0.9 and a learning rate of 0.0005, with batch size at 128. It can be seen that ViTs outperform ResNet152 in both the considered datasets, testifying the effectiveness of the proposed solution in extracting discriminative characteristics from vein patterns. It is worth mentioning that far better results were reported in [32] using ResNet158, yet closed-set verification conditions were there used, testing the models on the same subjects exploited for training. Results quite similar to those in [32] are obtained with ResNet152 in our tests under the same closed-set conditions, yet open-set represents a much more challenging scenario.

VI. CONCLUSION

In this paper, we demonstrated the effectiveness of using ViTs for vein biometric recognition and exploited the attention maps produced by ViTs when processing their inputs to argue for the explainability of the obtained results. We thus provide insights into the employed models' inner behavior when making their decisions.

Tests on the PUT and FYO wrist-vein datasets testified that the considered approach based on ViT can outperform state-of-the-art alternatives for open-set verification, with the best

EERs obtained on the two datasets respectively at 5.2% and 2.3%. More interestingly, we have demonstrated for the first time in literature that such models' decisions are taken mostly by focusing on image regions containing significant venous contents.

For the PUT database, we also proved the effectiveness of preprocessing the available wrist-vein images with CLAHE to enhance their quality and, consequently, the achievable performance by also providing examples of the differences in processing original or enhanced images through ViT attention maps.

In conclusion, with this work, we evaluated the explainability of wrist-vein biometric recognition by resorting to the attention maps produced when employing ViTs for image processing. The results provide interesting insights into the decision-making process for vascular biometric recognition, thus fostering transparency and trust in the performed methods and promoting responsible AI deployment. Conducting tests in open-set verification scenarios further increases the robustness and generalizability of the obtained outcomes.

REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004.

[2] Y. Ding, D. Zhuang, and K. Wang, "A study of hand vein recognition method," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, vol. 4, Aug. 2005, pp. 2106–2110.

[3] A. Uhl, C. Busch, S. Marcel, and R. Veldhuis, *Handbook of Vascular Biometrics*. Switzerland: Springer, 2020.

[4] L. Wang and G. Leedham, "Near- and far- infrared imaging for vein pattern biometrics," in *Proc. IEEE Int. Conf. Video Signal Based Surveill.*, Nov. 2006, p. 52.

[5] D. Mulyono and H. Shi Jinn, "A study of finger vein biometric for personal identification," in *Proc. Int. Symp. Biometrics Secur. Technol.*, Apr. 2008, pp. 1–8.

[6] N. Miura, A. Nagasaka, and T. Miyatake, "Extraction of finger-vein patterns using maximum curvature points in image profiles," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 8, pp. 1185–1194, Aug. 2007.

[7] J. Liu and Y. Zhang, "Palm-dorsa vein recognition based on two-dimensional Fisher linear discriminant," in *Proc. Int. Conf. Image Anal. Signal Process.*, Oct. 2011, pp. 550–552.

[8] J. Wang, H. Li, G. Wang, M. Li, and D. Li, "Vein recognition based on (2D) 2FPCA," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 6, no. 4, pp. 323–332, 2013.

[9] H. T. Van, C. M. Duong, G. Van Vu, and T. H. Le, "Palm vein recognition using enhanced symmetry local binary pattern and SIFT features," in *Proc. 19th Int. Symp. Commun. Inf. Technol. (ISCIT)*, Sep. 2019, pp. 311–316.

[10] J. M. Song, W. Kim, and K. R. Park, "Finger-vein recognition based on deep DenseNet using composite image," *IEEE Access*, vol. 7, pp. 66845–66863, 2019.

[11] Z. K. J. Jasim, A. H. Mohammed, L. Elwiya, B. D. Al-Jabbari, and H. Alhaji, "Human identification with finger vein image using deep learning," in *Proc. 5th Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, 2021, pp. 672–677.

[12] R. Hernández-García, E. H. Salazar-Jurado, R. J. Barrientos, F. M. Castro, J. Ramos-Cózar, and N. Guil, "From synthetic data to real palm vein identification: A fine-tuning approach," in *Proc. IEEE 13th Int. Conf. Pattern Recognit. Syst. (ICPRS)*, Jul. 2023, pp. 1–7.

[13] D. Castelvecchi, "Can we open the black box of AI?" *Nature*, vol. 538, no. 7623, pp. 20–23, Oct. 2016.

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

[15] K. Santosh and C. Wall, "Trustworthy and EXplainable AI for biometrics," in *AI, Ethical Issues and Explainability—Applied Biometrics*. Singapore: Springer, 2022, pp. 29–46.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[17] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[18] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, "XViT-COS: Explainable vision transformer based COVID-19 screening using radiography," *IEEE J. Translational Eng. Health Med.*, vol. 10, pp. 1–10, 2022.

[19] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable ai methods—A brief overview," in *Proc. Int. Workshop Extending Explainable AI Beyond Deep Models Classifiers*, 2022, pp. 13–38.

[20] R. Garcia-Martin and R. Sanchez-Reillo, "Vision transformers for vein biometric recognition," *IEEE Access*, vol. 11, pp. 22060–22080, 2023.

[21] W. Liu, W. Li, L. Sun, L. Zhang, and P. Chen, "Finger vein recognition based on deep learning," in *Proc. 12th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Jun. 2017, pp. 205–210.

[22] W. Kim, J. M. Song, and K. R. Park, "Multimodal biometric recognition based on convolutional neural network by the fusion of finger-vein and finger shape using near-infrared (NIR) camera sensor," *Sensors*, vol. 18, no. 7, p. 2296, Jul. 2018.

[23] K. Shaheed, H. Liu, G. Yang, I. Qureshi, J. Gou, and Y. Yin, "A systematic review of finger vein recognition techniques," *Information*, vol. 9, no. 9, p. 213, Aug. 2018.

[24] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–34, May 2018.

[25] F. Marattukalam, D. Cole, P. Gulati, and W. H. Abdulla, "On wrist vein recognition for human biometrics," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 66–73.

[26] A. Das, U. Pal, M. A. Ferrer Ballester, and M. Blumenstein, "A new wrist vein biometric system," in *Proc. IEEE Symp. Comput. Intell. Biometrics Identity Manage. (CIBIM)*, Dec. 2014, pp. 68–75.

[27] R. Kabacinski and M. Kowalski, "Vein pattern database and benchmark results," *Electron. Lett.*, vol. 47, no. 20, pp. 1127–1128, 2011.

[28] O. Nikisins, T. Eglitis, A. Anjos, and S. Marcel, "Fast cross-correlation based wrist vein recognition algorithm with rotation and translation compensation," in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Jun. 2018, pp. 1–7.

[29] R. Garcia-Martin and R. Sanchez-Reillo, "Wrist vascular biometric recognition using a portable contactless system," *Sensors*, vol. 20, no. 5, p. 1469, Mar. 2020.

[30] F. O. Babalola, Ö. Toygar, and Y. Bitirim, "Wrist vein recognition by fusion of multiple handcrafted methods," in *Proc. 3rd Int. Congr. Human-Comput. Interact., Optim. Robotic Appl. (HORA)*, Jun. 2021, pp. 1–5.

[31] Ö. Toygar, F. O. Babalola, and Y. Bitirim, "FYO: A novel multimodal vein database with palmar, dorsal and wrist biometrics," *IEEE Access*, vol. 8, pp. 82461–82470, 2020.

[32] R. Garcia-Martin and R. Sanchez-Reillo, "Deep learning for vein biometric recognition on a smartphone," *IEEE Access*, vol. 9, pp. 98812–98832, 2021.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

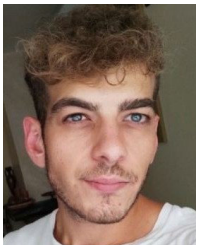
[34] T.-V. Nguyen, S.-J. Horng, D.-T. Vu, H. Chen, and T. Li, "LAWNet: A lightweight attention-based deep learning model for wrist vein verification in smartphones using RGB images," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.

[35] F. Marattukalam, W. Abdulla, D. Cole, and P. Gulati, "Deep learning-based wrist vascular biometric recognition," *Sensors*, vol. 23, no. 6, p. 3132, Mar. 2023.

[36] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[37] P. C. Neto, T. Gonçalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, and J. S. Cardoso, "Explainable biometrics in the age of deep learning," 2022, *arXiv:2208.09500*.
[38] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 1–37, Jan. 2018.
[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[40] 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 Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[41] Y.-Y. Chen, S.-Y. Jhong, C.-H. Hsia, and K.-L. Hua, "Explainable AI: A multispectral palm-vein identification system with new augmentation features," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 3s, pp. 1–21, Oct. 2021.
[42] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," 2016, *arXiv:1602.04938*.
[43] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu, "Gait-ViT: Gait recognition with vision transformer," *Sensors*, vol. 22, no. 19, p. 7362, Sep. 2022.
[44] P. Delgado-Santos, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez, "Exploring transformers for behavioural biometrics: A case study in gait recognition," *Pattern Recognit.*, vol. 143, Nov. 2023, Art. no. 109798.
[45] Z. Zhao, H. Zhang, Z. Chen, and J. Yang, "TransFinger: Transformer based finger tri-modal biometrics," in *Proc. Chin. Conf. Biometric Recognit.*, 2022, pp. 114–124.
[46] X. Li and B.-B. Zhang, "FV-ViT: Vision transformer for finger vein recognition," *IEEE Access*, vol. 11, pp. 75451–75461, 2023.
[47] H. Qin, C. Gong, Y. Li, X. Gao, and M. A. El-Yacoubi, "Label enhancement-based multiscale transformer for palm-vein recognition," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–17, 2023.
[48] S. Jain and B. C. Wallace, "Attention is not explanation," 2019, *arXiv:1902.10186*.
[49] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," 2019, *arXiv:1908.04626*.
[50] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M. B. Amin, and B. Kang, "Survey on explainable AI: From approaches, limitations and applications aspects," *Human-Centric Intell. Syst.*, vol. 3, no. 3, pp. 161–188, Aug. 2023.
[51] J. An and I. Joe, "Attention map-guided visual explanations for deep neural networks," *Appl. Sci.*, vol. 12, no. 8, p. 3846, Apr. 2022.
[52] B. M. de Vries, G. J. C. Zwezerijnen, G. L. Burchell, F. H. P. van Velden, C. W. Menke-van der Houven van Oordt, and R. Boellaard, "Explainable artificial intelligence (XAI) in radiology and nuclear medicine: A literature review," *Frontiers Med.*, vol. 10, May 2023.
[53] M. Usman, T. Zia, and A. Tariq, "Analyzing transfer learning of vision transformers for interpreting chest radiography," *J. Digit. Imag.*, vol. 35, no. 6, pp. 1445–1462, Dec. 2022.
[54] S. A. Grosz, J. J. Engelsma, R. Ranjan, N. Ramakrishnan, M. Aggarwal, G. G. Medioni, and A. K. Jain, "Minutiae-guided fingerprint embeddings via vision transformers," 2022, *arXiv:2210.13994*.
[55] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[57] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI Signal Processing-Systems for Signal, Image, Video Technol.*, vol. 38, no. 1, pp. 35–44, Aug. 2004.
[58] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.
[59] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017.
[60] R. Wightman. (2019). *PyTorch Image Models*. GitHub Repository. [Online]. Available: https://github.com/huggingface/pytorch-image-models

ROCCO ALBANO (Student Member, IEEE) received the bachelor's degree in electrical engineering and the master's degree in information and communication technology engineering from Roma Tre University, Italy, where he is currently pursuing the Ph.D. degree with specializing in neural networks focusing on attention mechanisms and explainable AI for biometric recognition. He combines a solid academic foundation with a passion for advancing artificial intelligence technologies. His work exemplifies a commitment to pushing the boundaries of AI in transparent and interpretable ways. His current research interests include improving biometric recognition capabilities through innovative applications of neural networks and attention mechanisms.



LORENZO GIUSTI received the bachelor's degree in computer science and engineering from Roma Tre University, specializing in quantum computing, and the master's degree in data science, with a focus on deep learning, from the Sapienza University of Rome, Italy, where he is currently pursuing the Ph.D. degree in data science, specialized in geometric and topological deep learning. His research journey includes a significant stint as a Visiting Ph.D. Student with the University of Cambridge and a Research Scientist Intern with NASA's Jet Propulsion Laboratory, where he led a project on Martian terrain modeling using spacecraft imagery and neural radiance fields. At CERN, he innovated in anomaly detection for particle accelerators.

EMANUELE MAIORANA (Senior Member, IEEE) received the Ph.D. degree in biomedical, electromagnetism, and telecommunication engineering, with European Doctorate Label, from Roma Tre University, Rome, Italy, in 2009. He is currently an Assistant Professor with the Department of Industrial, Electronics and Mechanical Engineering, Roma Tre University. His research interests include digital signal and image processing, with specific emphasis on biometric recognition. He was the General Chair of the 9th IEEE International Workshop on Biometrics and Forensics (IWBF) 2021 and the General Chair of the 16th IEEE International Workshop on Information Forensics and Security (WIFS) 2024. He is an Associate Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

PATRIZIO CAMPISI (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Roma Tre University, Rome, Italy. He is currently a Full Professor with the Department of Industrial, Electronics and Mechanical Engineering, Roma Tre University. His current research interests include biometrics and secure multimedia communications. He is the Vice President of Publications for the IEEE Biometrics Council. He was the IEEE SPS Director for Student Services, from 2015 to 2017, and the Chair of the IEEE Technical Committee on Information Forensics and Security, from 2017 to 2018. He was the General Chair of the 26th European Signal Processing Conference EUSIPCO 2018, Italy, the 7th IEEE Workshop on Information Forensics and Security (WIFS) 2015, Italy, and the 12th ACM Workshop on Multimedia and Security 2010, Italy. He was an Associate Editor and a Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and an Associate Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He was the Editor-in-Chief of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, from 2018 to 2021.

Open Access funding provided by 'Università degli Studi Roma Tre' within the CRUI CARE Agreement