



UNIVERSITÀ DEGLI STUDI ROMA TRE

DEPARTMENT OF INDUSTRIAL, ELECTRONIC, AND
MECHANICAL ENGINEERING

DOCTOR OF PHILOSOPHY THESIS IN
APPLIED ELECTRONICS

**Attention-based Network and
Explainable AI for
Biometrics Recognition**

Supervisors

PROF. PATRIZIO CAMPISI

PROF. EMANUELE MAIORANA

Coordinator

PROF. MARCO CARLI

Ph.D. Candidate

ROCCO ALBANO

A.Y. 2024/2025

A dissertation thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Applied Electronics at the
Department of Industrial, Electronic, and Mechanical Engineering of
Roma Tre University.

Abstract

Biometric recognition systems have become a cornerstone of modern authentication technologies, enabling automatic identity verification based on physiological traits. While deep learning has significantly improved recognition accuracy across multiple biometric modalities, current systems still face critical challenges related to robustness, generalization, and transparency. In particular, the black-box nature of deep neural networks raises concerns regarding trustworthiness, while real-world operating conditions—such as cross-session variability, unseen subjects, and signal degradation—often lead to performance drops that are difficult to interpret.

This doctoral thesis investigates the role of attention-based deep learning architectures and intrinsic explainability mechanisms in biometric recognition, with a specific focus on iris and vein modalities. The central hypothesis of this work is that attention mechanisms, as implemented in Vision Transformer (ViT) architectures, can simultaneously enhance recognition robustness and provide meaningful insights into the decision-making processes of biometric systems.

The thesis first provides a comprehensive overview of biometric recognition pipelines, presentation attack detection (PAD), and explainable artificial intelligence (XAI), highlighting the limitations of traditional convolutional approaches and post-hoc explainability methods. Particular emphasis is placed on the distinction between intrinsic and post-model explainability, motivating the adoption of attention-based architectures as a principled compromise between performance and interpretability.

Subsequently, the effectiveness of Vision Transformers is investigated in the context of wrist-vein biometric recognition. Experiments conducted on public datasets under open-set verification scenarios demonstrate that ViT-based models are capable of extracting highly discriminative features while exhibiting superior generalization compared to convolutional neural networks. The attention maps generated by the proposed models are analyzed to assess whether the decision process aligns with physiologically significant vein patterns, thus providing intrinsic explainability. The results show that attention mechanisms consistently focus on anatomically relevant regions, supporting the interpretability and reliability of the recognition process.

The thesis further explores the detection of iris presentation attack under realistic conditions by analyzing the impact of image compression on recognition performance and explainability. By evaluating traditional and deep learning-based compression schemes, the study demonstrates how increasing signal degradation affects not only error rates, but also the spatial distribution of attention within the model. Metrics such as percent root mean square difference (PRD) and Compression Efficiency are used to link signal-level distortion to recognition failures, revealing how attention degradation correlates with incorrect classification in PAD scenarios.

Overall, this work demonstrates that explainability in biometric systems should not be regarded solely as a transparency requirement but also as a powerful analytical tool for understanding robustness and failure modes. By integrating attention-based architectures with systematic experimental evaluation under realistic conditions, this thesis contributes to the development of more reliable, interpretable, and trustworthy biometric recognition systems.

Abstract

I sistemi di riconoscimento biometrico rappresentano oggi una componente fondamentale delle moderne tecnologie di autenticazione, consentendo la verifica automatica dell'identità sulla base di caratteristiche fisiologiche. Sebbene l'impiego del deep learning abbia portato a significativi miglioramenti in termini di accuratezza, i sistemi biometrici attuali presentano ancora criticità rilevanti legate alla robustezza, alla capacità di generalizzazione e alla trasparenza dei processi decisionali. In particolare, la natura “black-box” delle reti neurali profonde solleva problematiche di affidabilità, mentre condizioni operative realistiche—come la variabilità tra sessioni, la presenza di soggetti non visti in fase di addestramento e il degrado del segnale—possono causare cali prestazionali difficili da interpretare.

Questa tesi di dottorato analizza il ruolo delle architetture deep learning basate su meccanismi di attenzione e delle tecniche di explainable artificial intelligence (XAI) intrinseche nei sistemi biometrici, con particolare riferimento alle modalità dell'iride e delle vene. L'ipotesi centrale del lavoro è che i meccanismi di attenzione, implementati all'interno delle architetture Vision Transformer (ViT), consentano di migliorare simultaneamente la robustezza del riconoscimento e la comprensione dei processi decisionali dei sistemi biometrici.

La tesi presenta inizialmente un'analisi approfondita dei sistemi biometrici, delle tecniche di presentation attack detection (PAD) e delle metodologie di explainable AI, evidenziando i limiti delle architetture convoluzionali tradizionali e degli approcci di spiegabilità post-hoc. Viene posta particolare attenzione alla distinzione tra explainability intrinseca e post-modello, motivando l'adozione di architetture basate su attenzione come soluzione equilibrata tra prestazioni e interpretabilità.

Successivamente, viene studiata l'efficacia dei Vision Transformers nel riconoscimento biometrico basato su pattern venosi del polso. Gli esperimenti condotti su dataset pubblici in scenari di verifica open-set dimostrano che i modelli ViT sono in grado di apprendere rappresentazioni altamente discriminanti, garantendo una migliore capacità di generalizzazione rispetto alle reti convoluzionali. L'analisi delle mappe di attenzione consente di verificare che le decisioni del modello si basano su regioni anatomi-

camente significative, fornendo una spiegabilità intrinseca del processo di riconoscimento.

La tesi affronta inoltre il problema della presentation attack detection dell'iride in condizioni realistiche, analizzando l'impatto della compressione delle immagini sulle prestazioni e sulla spiegabilità del sistema. Attraverso lo studio di schemi di compressione tradizionali e basati su deep learning, viene mostrato come il degrado del segnale influisca non solo sui tassi di errore, ma anche sulla distribuzione spaziale dell'attenzione del modello. L'impiego di metriche quali la Percentage Root Mean Square Difference (PRD) e la Compression Efficiency consente di collegare il livello di distorsione del segnale ai meccanismi di fallimento del sistema di PAD.

Nel complesso, questo lavoro dimostra che l'explainability nei sistemi biometrici non deve essere considerata esclusivamente come un requisito di trasparenza, ma anche come uno strumento analitico per comprendere la robustezza e i limiti dei modelli. L'integrazione di architetture basate su attenzione con una valutazione sperimentale rigorosa in condizioni realistiche contribuisce allo sviluppo di sistemi biometrici più affidabili, interpretabili e degni di fiducia.

List of Tables

3.1	Employed wrist-vein dataset characteristics.	26
4.1	Recognition performance comparison between approaches based on ViTs and ResNet152 on the PUT-Wrist dataset.	46
5.1	Iris PAD performance of the proposed methods, and comparison with literature algorithms. [1–4].	61
5.2	Iris PAD performance for ViT models fine-tuned with images JPEG-compressed at different QFs, at increasing JPEG compression levels.	63
5.3	Iris PAD performance for ViT models fine-tuned with images JPEG-AI-compressed at different TRs, at increasing JPEG AI compression levels.	64
5.4	Iris PAD performance at different compression levels for Swin models trained on uncompressed images and fine-tuned with images compressed with JPEG @ QF=10.	65
5.5	Iris PAD performance at different compression levels for Swin models trained on uncompressed images and fine-tuned with images compressed with JPEG AI @ TR=0.12.	66

List of Figures

3.1	Original wrist-vein patterns (left) and their CLAHE-enhanced representations (right) from the PUT Wrist-Vein dataset. . .	27
3.2	Illustrative representation of the Vision Transformer (ViT) [5] tailored for explainable vein biometric recognition.	30
4.1	Attention maps of ViTs trained for different numbers of epochs, for two different images of the same wrist from the PUT dataset.	44
4.2	Attention maps of ViTs trained for different numbers of epochs, for two different images of the same wrist from the FYO dataset.	44
4.3	EERs of ViT models trained on PUT wrist vein images either with or without CLAHE preprocessing.	45
4.4	Effects of CLAHE on ViTs trained on PUT.	46
5.1	Attention maps for original images: textured lenses in the first two columns on the left, no lenses in the two middle ones, soft lenses in the last two on the right.	60
5.2	Attention maps generated by ViT and Swin models for bonafide and spoofing test images from the databases considered. . .	61
5.3	Iris PAD performance for increasing JPEG compression levels. In order NDCLD15, Notredame, IIIT-WVU and Clarkson	62
5.4	Iris PAD performance for increasing JPEG AI compression levels. In order NDCLD15, Notredame, IIIT-WVU and Clarkson	62
5.5	Attention maps for increasing JPEG compression on textured iris images, with decreasing quality from left to right.	69

Contents

Abstract	ii
List of Tables	viii
List of Figures	x
Contents	xi
1 Introduction	1
1.1 Motivation	1
1.2 Transparent biometric recognition	2
1.3 Scope and objectives	3
1.4 Contributions of this thesis	4
1.5 Thesis outline	5
2 Background	9
2.1 Introduction	9
2.2 Fundamentals of biometric systems	10
2.2.1 Definition and Principles	10
2.2.2 Biometric modalities: Physiological vs Behavioural	11
2.2.3 Biometric system pipeline	11
2.2.4 Variability and Degradation in Biometric Signals	13
2.2.5 Security vulnerabilities	14
2.3 Deep Learning for Biometric recognition	14
2.3.1 Handcrafted features	15
2.3.2 Deep Learning approaches	16
2.3.3 Convolutional Neural Networks (CNNs)	16
2.4 Explainable Artificial Intelligence (XAI)	17
2.4.1 Why Explainability matters	18
2.4.2 Post-hoc XAI methods	18
2.4.3 Intrinsic explainability via Attention mechanism	18
2.4.4 Failure modes	19

2.5	Evaluation protocols	20
2.5.1	Biometric evaluation scenarios	20
2.5.2	Learning objectives and loss functions	21
2.5.3	Performance metrics and operating points	21
3	Methodology	23
3.1	Introduction	23
3.2	Datasets	25
3.2.1	Vein Datasets	25
3.2.2	Iris Datasets	27
3.3	Model architectures	29
3.3.1	ViT configuration	29
3.3.2	Swin Transformers	31
3.4	Experimental protocols	32
3.4.1	Open-set verification	33
3.4.2	PAD evaluation	33
3.4.3	Compression experiments	34
3.4.4	Similarity scoring	34
4	Explainable wrist-vein biometric Recognition	37
4.1	Introduction	37
4.2	Vision Transformer-based Explainability Framework	38
4.2.1	Patch Tokenization and Embedding	38
4.2.2	Self-Attention Mechanism	38
4.2.3	Aggregation of Attention Maps for Explainability	39
4.2.4	Interpretability of Vascular Regions	40
4.2.5	Impact of Pre-processing on Attention Explainability	41
4.2.6	Role of Explainability in Verification	42
4.3	Experiments	43
4.4	Summary and Conclusion	47
5	Iris PAD with Transformers	49
5.1	ViT for Presentation Attack Detection	52
5.1.1	Image Representation, Patch Tokenization, and Embedding	52
5.1.2	Self-Attention Mechanism for Iris PAD	55
5.1.3	Training Strategy, Data Augmentation, and Fine-Tuning Protocol	56
5.1.4	PAD Under Image Compression	57
5.1.5	Why Transformers Show Superior Robustness to Compression	58
5.2	Results	59

5.2.1	Experimental Setup and General Observations	59
5.2.2	Performance on Datasets	60
5.2.3	Influence of Patch Size and Training Duration	61
5.2.4	PAD Under Compression: Quantitative and Qualitative Insights	62
5.3	Explainability	67
5.3.1	Attention Maps as a Window Into Model Reasoning	67
5.3.2	Genuine vs Attack Patterns: What Attention Reveals	67
5.3.3	Explainability Under Compression	68
5.3.4	Explainability for Failure Analysis	69
5.3.5	Toward Transparent and Trustworthy Iris PAD Systems	70
6	Global Discussion, Conclusions, and Future Directions	71
6.1	Discussion	72
6.1.1	Comparison across biometric modalities	72
6.1.2	Role of attention mechanisms	73
6.1.3	Interpretation of attention patterns	74
6.2	Explainability and Robustness	74
6.2.1	Explainability beyond visualization	74
6.2.2	Effects of degradation and compression	75
6.2.3	Failure cases and diagnostic analysis	75
6.3	Practical Considerations	75
6.3.1	Implications for biometric system design	76
6.3.2	Operational and security aspects	76
6.4	Limitations and Future Work	76
6.4.1	Limitations of the proposed approach	76
6.4.2	Directions for future research	77
6.5	Final Conclusions	77
	References	81



Chapter 1

Introduction

1.1 Motivation

In recent years, biometric systems have played an important role in many application domains [6, 7], from digital services and consumer devices to high-security environments. The ability to recognize or verify a person’s identity through physiological or behavioral traits – such as fingerprints, face, veins, or iris patterns [8–10] - allows these systems to replace or support traditional methods based on passwords, codes, or physical tokens, offering both greater convenience and improved security. At the same time, advances in deep learning have significantly improved the performance of biometric systems, as documented in recent surveys [11–13]. Neural networks can automatically learn highly discriminative features directly from raw data. However, these models still have a clear limitation: they often behave like a “black box” [14–16]. In most cases, the system produces a decision (for example, user accepted or attack detected) without providing a clear explanation of why it made that choice.

In security-critical scenarios, this lack of transparency can be problematic. It is not enough to know that a model “works well” in terms of average accuracy. It is equally important to understand the information that the model relies on when making its decisions. Identifying the image regions that contribute the most to the predictions of the model is essential [17, 18] to verify that the system is not influenced by irrelevant details, noise, or features that are not truly meaningful from a biometric point of view.

When using deep learning architectures, it must be remembered that these networks learn on their own the features they consider useful for the task. They are not required to focus on characteristics that align with expert domain knowledge. For this reason, research has often focused mainly

on performance metrics, for example, reducing the Equal Error Rate or increasing accuracy, without examining in detail how and why the model behaves in a certain way [19]. In this situation, a neural network can achieve good performance relying on shortcuts or misleading patterns, such as reflections, image borders, sensor artifacts, or dataset-specific textures [20]. From a numerical point of view, the model may seem correct, but from a biometric or security perspective, this represents a real risk since these shortcuts may fail when conditions change or may create vulnerabilities exploitable by attackers.

For these reasons, there is a strong need to complement traditional performance metrics with an explicit evaluation of the internal behavior of the model. It is necessary to check whether the network focuses on image regions that truly correspond to biometric information (for example, wrist vein structures or iris textures) rather than on accidental or irrelevant areas. In other words, we should ask not only how accurate a model is but also what it is looking at when making its decisions.

This is where Explainable Artificial Intelligence (XAI) becomes relevant [14, 16, 21]. XAI provides tools to interpret and understand the decision-making process of machine learning models. In particular, Vision Transformers (ViT) represent a suitable architecture for this purpose. They are based on self-attention mechanisms [5, 22] that, in addition to achieving competitive performance compared to convolutional networks, naturally produce attention maps showing which areas of the image influence the final prediction. These maps offer an intrinsic and often more stable form of explainability compared to post-hoc techniques [23–25]. The goal of this thesis is to show how attention mechanisms can be used not only as a qualitative visualization tool but also as a central element in evaluating biometric models based on Transformer architectures. This work focuses especially on wrist-vein recognition and iris presentation attack detection (PAD). The main idea is that attention maps provide information that is richer and more meaningful than accuracy alone and that they can guide the design, selection, and validation of more robust and reliable biometric systems.

1.2 Transparent biometric recognition

Transparency in biometric recognition refers to the ability of a system to justify or communicate the rationale behind its decisions [26] [27]. This goes beyond mere interpretability; it encompasses the design of models whose behavior is predictable, accountable, and aligned with biometric relevance [28]. In the context of deep learning, interpretability plays a critical role

for several interconnected reasons. From a security perspective, opaque models may inadvertently exploit non-biometric cues that are correlated with training data but not related to the underlying biometric trait [20, 29]. Such dependencies can be manipulated by adversaries, increasing the vulnerability of the system to sophisticated attacks that bypass the intended security mechanisms [30, 31].

Transparency is also closely related to robustness. Systems whose internal decision processes can be inspected and understood are generally less prone to catastrophic failures when faced with previously unseen data distributions [19, 32]. By revealing how decisions are made, transparency allows practitioners to assess whether the behavior of a model is likely to remain stable under changes in acquisition conditions, sensors, or attack strategies.

Another important aspect concerns certification and regulatory compliance. Emerging regulations, such as the European Union AI Act, are expected to impose explainability requirements on high-risk applications, including biometric systems [16, 21]. In this context, transparent models provide a foundation for demonstrating compliance, as they enable developers and auditors to justify decisions and document the rationale underlying automated outcomes.

Finally, transparency contributes to scientific insight. By making the decision-making process accessible, it becomes possible to better understand which physical features are truly discriminative for a given biometric task [8, 33]. This understanding not only supports the development of more effective and principled models, but also fosters progress in the broader scientific study of biometric traits and their distinctive properties.

Transformers provide a unique opportunity to achieve transparency thanks to the inherent interpretability of their self-attention layers [22, 34]. Unlike convolution filters, attention weights explicitly quantify the contribution of each image patch to the model’s reasoning process [23–25]. This makes them a powerful tool for evaluating whether the network is responsive to stable features such as vein bifurcations or iris crypts, rather than spurious correlations caused by data set artifacts [35]. In this paper, transparency is treated not only as an ethical requirement, but also as a practical instrument for model validation, enabling a more rigorous examination of characteristic relevance, bias, and failure modes [26, 27].

1.3 Scope and objectives

The main scope of this thesis is to investigate how attention-based architectures and explainability principles can be jointly exploited to improve the development, evaluation, and reliability of biometric systems. In particular,

the work focuses on understanding how modern Transformer-based models can be leveraged not only to improve recognition performance but also to provide meaningful insights into the internal decision-making processes underlying biometric recognition and security-oriented tasks.

Within this scope, the thesis examines the role of self-attention mechanisms in biometric applications, with the objective of analyzing how Vision Transformers and Swin Transformers encode discriminative biometric information. Rather than focusing on the evolution of attention at individual network layers, the analysis emphasizes how attention contributes to the formation of robust representations and how these representations support recognition and classification decisions across different tasks and operating conditions. An additional objective of this work is to investigate the adequacy of explainability as a diagnostic and analytical tool. In this context, attention maps are studied as a means to identify anomalous behaviors, shortcut learning phenomena, and potential vulnerabilities in both biometric recognition and Presentation Attack Detection scenarios. This perspective aims to assess whether explainability can complement traditional performance metrics by revealing not only how well models perform, but also how and why specific decisions are produced.

Finally, the thesis considers multiple biometric modalities in order to analyze how attention-based explainability behaves across heterogeneous data types and task objectives. Two representative and complementary scenarios are addressed: wrist-vein recognition, which focuses on identity verification, and iris Presentation Attack Detection, which addresses biometric security and anti-spoofing. Through this comparative analysis, the thesis highlights similarities and differences in the way attention mechanisms interact with different biometric traits, contributing to a unified and modality-aware understanding of explainable biometric systems.

In general, these objectives define a coherent framework in which attention-based models are investigated not merely as powerful classifiers, but as transparent and interpretable tools for the design and validation of robust, trustworthy, and security-oriented biometric systems.

1.4 Contributions of this thesis

The contributions of this thesis consist of a comprehensive and systematic investigation of how attention-based architectures and explainability principles can be exploited to enhance the reliability, transparency, and robustness of biometric recognition systems operating under realistic conditions.

The first contribution of this work lies in the in-depth analysis of Transformer architectures, namely Vision Transformers and Swin Transformers,

applied to two distinct yet complementary biometric tasks: wrist-vein recognition and iris presentation attack detection. Through a detailed examination of their internal representations, the thesis analyzes how self-attention mechanisms encode biometric information and how these representations evolve across network layers, revealing task-dependent and architecture-specific behaviors that are not accessible through performance metrics alone.

A second major contribution is the definition of a unified interpretability framework that integrates attention maps directly into the biometric model evaluation pipeline. Rather than relying exclusively on traditional indicators such as accuracy, EER, APCER, or BPCER, the proposed framework evaluates the consistency, relevance, and stability of the regions attended by the model. This perspective enables the identification of subtle failure modes, unintended shortcuts, and data set-dependent biases that would remain hidden when the assessment is limited to numerical performance measures. Furthermore, the thesis provides an extensive empirical analysis of the robustness of attention patterns across multiple datasets, acquisition conditions, and image compression levels. This study demonstrates that models exhibiting similar performance metrics may rely on substantially different internal reasoning processes. Such findings motivate an interpretability-driven methodology for the selection and validation of biometric models, in which architectural choices are guided not only by recognition accuracy but also by the coherence and physiological plausibility of the attended biometric structures.

Finally, this work contributes to the broader scientific understanding of biometric models based on Transformers through a systematic comparison between Vision Transformers and Swin Transformers. Differences in patch embedding strategies, hierarchical representations, and attention distributions are analyzed in relation to their impact on transparency and robustness. The results support the view that explainability should be regarded as a fundamental design component of biometric systems, rather than as an auxiliary analysis tool, particularly in real-world and security-critical deployment scenarios.

1.5 Thesis outline

This dissertation is composed of six Chapters and four Appendices. The core scientific contributions of this work, aligned with the research objectives outlined in Chapter 1, are developed from Chapter 3 to Chapter 5. These Chapters correspond to the main research activities and peer-reviewed publications produced during the Ph.D. journey, which are listed in Appendix B. Rather than representing a collection of independent studies, they form a

coherent investigation of attention-based architectures and intrinsic explainability techniques applied to biometric recognition and security-oriented tasks under realistic operating conditions. The following outline is intended to clarify the logical progression of the dissertation and the specific role of each Chapter within the overall research framework.

Chapter 2 provides the theoretical and methodological background necessary to contextualize the contributions of this thesis. Introduces the fundamental principles of biometric recognition systems, with particular emphasis on physiological modalities such as wrist-vein and iris biometrics, and discusses how acquisition devices, environmental conditions, and signal quality affect system reliability. The Chapter then reviews the main deep learning paradigms adopted in biometric recognition, covering convolutional neural networks as well as Transformer-based architectures, highlighting their respective strengths and limitations. A dedicated section is devoted to explainable artificial intelligence, where post-hoc and intrinsic interpretability methods are reviewed [36], with a focus on their applicability and limitations in biometric scenarios. Overall, this Chapter establishes the conceptual foundation required to analyze robustness, generalization, and explainability in the subsequent Chapters.

Chapter 3 introduces the experimental and methodological framework adopted throughout the thesis. The Chapter presents the datasets employed for wrist-vein recognition and iris Presentation Attack Detection (PAD), detailing their acquisition protocols, structure, and inherent challenges in terms of inter-session variability, noise, and limited data availability. It then describes the preprocessing pipelines, fine-tuning strategies, and architectural configurations adopted for Vision Transformers and Swin Transformers, as well as the procedures used to extract and analyze attention maps. By explicitly motivating each methodological choice with respect to robustness and interpretability, this Chapter establishes a clear link between the research questions and the experimental analyses carried out in the following Chapters.

In Chapter 4, attention-based Transformer architectures are investigated in the context of wrist-vein biometric recognition. This modality is particularly well suited to the analysis of explainability, as the discriminative information is spatially localized and anatomically meaningful. The Chapter analyzes how Vision Transformers and Swin Transformers encode vascular patterns through attention-based representations and examines the relationship between attention distributions and recognition performance under open-set conditions. The experimental results reveal both the strengths and limitations of existing models, such as a tendency to attend to peripheral or non-biometric regions, and demonstrate how intrinsic explainability can

be exploited to diagnose these behaviors. By linking attention patterns to recognition outcomes, this Chapter provides insights into the design of more stable, transparent, and biologically plausible biometric recognition systems.

Chapter 5 focuses on iris Presentation Attack Detection, addressing biometric security under realistic and adverse conditions. The Chapter evaluates the effectiveness of Transformer-based architectures in discriminating between bonafide and attack samples across multiple datasets and under varying levels of image compression. Through a systematic analysis of recognition performance and attention distributions, the study highlights the presence of dataset-specific shortcuts and vulnerability patterns that are not captured by standard error metrics alone. The results demonstrate how attention-based explainability can serve as an analytical tool to interpret PAD failure modes and to support the development of more robust and trustworthy biometric security systems.

Finally, Chapter 6 summarizes the main findings of the thesis, consolidates the scientific contributions, and discusses the limitations of the proposed approaches. The Chapter also outlines future research directions, with particular emphasis on robustness, generalization across acquisition conditions, and the role of explainability in regulatory and deployment-oriented contexts.



Chapter 2

Background

Biometric recognition has changed a lot over the years. It started as a small field that was used mainly in forensic applications and controlled environments, but it is now part of everyday life. We unlock our phones with our face or fingerprint, we pass through airport gates using biometric checks, and many wearable devices can measure physiological signals that can also be used for recognition. These changes have been possible due to better sensors, new machine-learning techniques, and the growing need for simple and secure authentication methods.

This chapter gives the basic concepts needed to understand the methods used in this thesis. First, Section 2.1 introduces the objectives of the chapter. Section 2.2 explains what biometric systems are, how they evolved, what modalities exist, and why variability is such an important challenge. The chapter then moves to deep learning Section 2.3, transformer-based models Section ??, explainability Section 2.4, and finally to the common evaluation protocols used in biometric research Section 2.5.

2.1 Introduction

Biometric recognition is the automatic identification or verification of a person based on a physical or behavioral characteristic. Although people often associate biometrics with recent technology, the idea is actually quite old. Early identification methods were based on manual measurements or visual comparisons. With the arrival of digital sensors and computers, automatic systems became possible, first using handcrafted features and simple classifiers. Over time, as applications moved to less controlled environments, these traditional methods started to show clear limitations.

Deep learning then became a key turning point in the field. Instead

of manually designing features, neural networks learn them directly from the data, achieving much better performance and generalization. More recently, transformer-based models have been introduced, bringing advantages in modeling global information and offering forms of explainability through attention maps.

This chapter provides the necessary background to understand these developments. Review the basic structure of biometric systems, the types of biometric traits, the main sources of variability, and the limits of older approaches. These sections prepare the ground for the deep-learning methods and explainable models analyzed in later chapters.

2.2 Fundamentals of biometric systems

Biometric systems are built to recognize individuals by analyzing characteristics that are unique or very difficult to reproduce. These traits can be physiological, such as fingerprints, iris texture, facial features, vascular patterns, or physiological signals such as PPG, ECG, or EEG. They can also be behavioral, such as gait or voice.

To work well, a biometric trait should be present in most people, show differences between individuals, and remain stable enough over time. The system must also be able to acquire the trait reliably and process it in a consistent way. Even if these principles sound simple, real-world biometric data often contain noise, distortions, and variations that make recognition difficult. Understanding how biometric systems work therefore requires looking at their history, their pipeline, and the challenges they face.

2.2.1 Definition and Principles

A biometric system identifies or verifies a person by analyzing a physical or behavioral characteristic that is naturally related to the individual. For a biometric trait to be useful in practice, several basic principles must be met. These principles guide not only the choice of the trait but also the design of the entire biometric pipeline.

Universality ensures that the trait is present in the population of interest, which means that most users can be enrolled. Distinctiveness guarantees that different people have sufficiently different patterns, allowing the system to separate identities. Permanence refers to the stability of the trait over time; if the trait changes too quickly, the system must re-enroll users frequently. Finally, measurability refers to the ease with which a sensor can capture and reliably capture a trait.

Although these concepts are often introduced early in the biometric literature, they have strong practical consequences. For example, distinctiveness affects how large the embedding space must be; permanence influences how often the system should update templates; collectability defines the sensor quality, illumination requirements, or contact constraints. In designing a real biometric application, these principles are constantly balanced against cost, usability, and security. For this reason, the study of biometric systems always begins with an understanding of these fundamental elements.

2.2.2 Biometric modalities: Physiological vs Behavioural

Physiological and behavioral modalities differ not only in the nature of the trait but also in their stability, acquisition difficulty, and vulnerability to attacks. Physiological traits, such as fingerprints, the iris, or vascular patterns, are generally stable because they depend on physical structures that do not change significantly from day to day. As a result, they tend to offer very high accuracy and are preferred in applications where security is essential. Behavioral traits, such as voice or gait, are influenced by psychological and environmental factors. They may change with emotion, fatigue, or health conditions. Even though they are less stable, they have the important advantage of being easy to capture in a non-intrusive way. In addition, they allow for continuous authentication since behavior naturally occurs during the interaction with a device.

Another difference concerns security. Physiological traits are harder to spoof because they reflect physical structures, while behavioral traits can be easier to imitate or reproduce. On the other hand, physiological traits are irreversible: if compromised, they cannot simply be replaced, while behavioral patterns can be adapted or changed.

This comparison shows why different modalities suit different applications. High-security environments tend to prefer physiological traits, while user-friendly systems or continuous authentication favor behavioral traits.

2.2.3 Biometric system pipeline

Although biometric systems may appear simple from an external perspective, each stage of their processing pipeline involves critical design choices that directly affect system performance and reliability [6,7,19]. The pipeline typically begins with the acquisition stage, which determines the quality of the data available to the system and is strongly influenced by sensor characteristics, environmental conditions, and user cooperation [6,32]. Since

imperfections introduced at this stage propagate throughout the entire system, acquisition quality plays a fundamental role in the overall robustness of biometric recognition. Preprocessing represents the first opportunity to mitigate acquisition-related variability. Typical operations include noise reduction, normalization, segmentation, and enhancement, all of which have a direct impact on the stability and discriminative power of extracted features [19, 37]. The effectiveness of preprocessing is therefore tightly coupled with the characteristics of the input data and with the assumptions made by subsequent stages of the pipeline.

Feature extraction constitutes a central component of the biometric process. Traditional biometric systems relied on handcrafted features explicitly designed by domain experts, such as texture descriptors for images or morphology-based representations for vascular patterns [9, 33, 38]. In contrast, modern biometric systems increasingly adopt the learned representations produced by deep neural networks, which automatically capture discriminative patterns directly from the data [11, 12]. The resulting biometric template must strike a balance between compactness—required for efficient storage and fast comparison—and expressiveness, which is necessary to preserve sufficient identity-related information for reliable recognition [?, 39].

During the recognition stage, biometric templates are compared through a matching process that produces a similarity score [6, 19]. An operating threshold is then selected to balance false acceptances and false rejections according to the security requirements of the target application [40, 41]. In parallel, many contemporary biometric systems incorporate template protection mechanisms—such as hashing, encryption, or revocable biometrics—to mitigate the risks of reverse engineering and identity theft [?, 39].

Although the biometric pipeline is often represented as a linear sequence of stages, its components are in practice strongly interdependent [19, 32]. Degraded acquisition quality increases the burden on preprocessing, weak feature representations require more permissive decision thresholds, and strict thresholds may lead to elevated false rejection rates. As a consequence, the design of a biometric system inevitably involves trade-offs between usability, accuracy, computational efficiency, and security [28].

From a modern perspective, these dependencies also expose critical limitations of traditional evaluation strategies. Deep learning-based models can achieve high recognition accuracy by exploiting unintended correlations or dataset-specific artifacts, without necessarily relying on stable and physiologically meaningful biometric information [20, 29]. When system behavior is assessed solely through aggregate performance metrics, such failure modes can remain undetected, particularly under changing acquisition conditions,

open-set scenarios, or adversarial settings [30, 41]. This observation motivates the need for complementary analysis tools—such as explainability methods—that allow the internal behavior of biometric models to be examined alongside their numerical performance [15, 16].

2.2.4 Variability and Degradation in Biometric Signals

Variability represents one of the most critical challenges in biometric recognition systems [19, 32]. Even when samples are acquired from the same individual using the same device, significant differences may arise due to changes in environmental conditions, user behavior, and acquisition setup. For image-based biometrics, factors such as illumination variations, shadows, motion blur, reflections, and sensor noise can introduce distortions that adversely affect recognition performance [8, 37]. Similarly, physiological signals are inherently variable, as they are influenced by stress, breathing patterns, electrode contact quality, heart rate dynamics, and motion artifacts [32, 42].

Multi-session variability further exacerbates these challenges. When biometric data are collected in different sessions that span days, weeks, or months, external and internal factors can accumulate. User posture, fatigue, emotional state, and sensor placement can vary between sessions, while physiological characteristics naturally evolve over time. These effects are particularly pronounced for physiological modalities such as photoplethysmography (PPG) and electroencephalography (EEG), whose signals exhibit intrinsic non-stationary behavior driven by underlying biological processes [19, 32].

In addition to natural variability, the degradation of data introduced by preprocessing operations and compression techniques plays a significant role in modern biometric systems. Image compression may suppress fine-grained spatial details that are essential for reliable recognition, as is the case for iris and vein patterns [33, 43], while filtering and resampling may alter the temporal structure of physiological signals. As a consequence, models trained on high-quality data can experience substantial performance degradation when exposed to compressed or otherwise degraded samples, unless the robustness to such transformations is explicitly addressed during training [44, 45]. From an evaluation standpoint, these sources of variability and degradation expose important limitations of performance-centric assessment strategies. Similar recognition accuracies may mask fundamentally different model behaviors, particularly when systems rely on spurious correlations or data set-specific artifacts that do not generalize under acquisition conditions [20, 29].

In this context, explainability techniques provide a complementary perspective by enabling the analysis of how variability and degradation affect the internal decision mechanisms of biometric models. By examining changes in attention distributions and feature representations, it becomes possible to distinguish between models that are genuinely robust and those whose apparent performance relies on fragile or non-physiological cues [15, 16].

2.2.5 Security vulnerabilities

Biometric systems face several security threats, and protecting them is as important as achieving high accuracy. Since biometric traits cannot be easily replaced once compromised, any attack that reveals or manipulates templates has long-term consequences.

Presentation attacks are particularly dangerous because they occur at the sensor level. A fake face printed on paper, a video replay, a 3D mask or textured contact lenses can be presented directly to the camera, bypassing internal security mechanisms *Czajka et al.* [46]. Physiological traits such as PPG and EEG can also be attacked by simulated or artificially generated signals. Because these attacks imitate acquisition conditions, they are often difficult to detect without additional countermeasures.

Other attacks target the communication channel, the matcher, or the stored templates. If templates are stolen, they can potentially be used to spoof the system or infer sensitive information about the user.

For these reasons, biometric systems often include Presentation Attack Detection (PAD) modules. PAD aims to distinguish genuine samples from fake ones by analyzing texture anomalies, motion patterns, physiological cues, or specific frequency signatures. A strong PAD system is essential for any biometric deployment where security is critical.

2.3 Deep Learning for Biometric recognition

Deep learning has reshaped the entire field of biometric recognition. Before its introduction, most of the systems relied on hand-made features designed manually by experts. These features worked reasonably well in controlled environments, but their performance dropped quickly when acquisition conditions changed. With the increasing use of biometrics in everyday devices, such as smartphones and wearable devices, systems needed to deal with much more variability than ever before. This shift required models capable of learning robust and flexible representations, capable of adapting to dif-

ferent sensors, different sessions, and different types of noise. Deep learning became the natural answer to these needs *Nguyen et al.* [47] .

The success of deep learning comes from its ability to extract multiple levels of abstraction from raw data. Instead of relying on a single handcrafted descriptor, deep networks learn a hierarchy of features, from simple patterns in the early layers to complex identity-related structures in the deeper layers. This approach has proven extremely effective for many biometric modalities, including face, iris, fingerprint, vein, and physiological signals such as PPG and EEG. The following subsections describe the transition from handcrafted features to deep-learning methods, focusing on the benefits, challenges, and main architectures used in biometric recognition.

2.3.1 Handcrafted features

Before deep learning, biometric systems depended almost entirely on features designed manually. These features were meant to capture what experts believed were the most discriminative characteristics of a biometric trait. For fingerprints, this included minutiae points and ridge orientations. For face recognition, common methods included Gabor filters, Local Binary Patterns (LBP), and Principal Component Analysis (PCA) approaches, such as eigenfaces. Iris recognition relied on specific phase-based encodings, while vascular biometrics often used texture descriptors such as Local Phase Quantization or filtered responses from multi-scale operators.

These methods had clear strengths. They were easy to compute, required little memory, and relatively stable when operating in controlled environments with consistent lighting, pose, and sensor characteristics. However, they also had major weaknesses. Their performance strongly depended on the acquisition conditions, and even small changes in illumination or sensor quality could affect the extracted features. The design of features was a manual and time-consuming process, and each modality required its own specialized descriptors. For example, a feature designed for fingerprint recognition would fail completely when applied to iris or vein patterns.

A fundamental limitation was that hand-made features were based on fixed mathematical operators. They could not adapt their behavior based on the data seen during the training. As biometric applications expanded into more variable and unconstrained environments, these rigid features became increasingly inadequate. The need for representations that could automatically adapt to real-world conditions made deep learning especially attractive.

2.3.2 Deep Learning approaches

Deep learning introduced the possibility of learning features directly from data instead of designing them manually. The idea of automatically learning representations was not entirely new, but the breakthrough came when deep neural networks became trainable on large datasets thanks to GPU computing and improved training techniques. One of the turning points in the history of deep learning was the success of AlexNet in 2012, which demonstrated the power of deep architectures for image classification. This success quickly extended to biometric recognition tasks.

Deep-learning-based systems work differently from hand-made ones. Instead of applying fixed filters, they use trainable parameters that are adjusted during training. This allows the network to learn patterns that are directly useful in distinguishing individuals. For example, a deep network trained on iris images can automatically learn features related to iris texture, boundaries, or fine structures that would be difficult to design manually.

Another important contribution to deep learning is the introduction of loss functions designed specifically for biometric verification. Classical classification loss functions, such as cross-entropy, encourage the model to correctly classify each sample into one of the training classes. However, biometric systems must compare samples across subjects, including subjects not seen during training. For this reason, metric learning methods became common. Techniques such as contrastive loss, triplet loss, and margin-based softmax losses encourage the network to produce compact clusters for each individual while pushing different individuals apart. This approach significantly improves generalization in open-set scenarios.

Deep learning also brought flexibility. A network trained on one sensor can often be fine-tuned for another sensor or adapted to deal with different acquisition conditions. These capabilities solved many of the problems that limited traditional biometric systems, especially when dealing with large variability and multi-session data.

2.3.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) played a central role in the early success of deep learning for biometrics. CNNs are structured to process spatial patterns efficiently, making them ideal for image-based modalities. Their architecture is built around convolutional layers that apply small filters throughout the image. This design allows CNNs to capture local structures such as edges, textures, and repeated patterns—features that are essential for biometric recognition. One of the main strengths of CNNs is their hierarchical nature. The early layers detect simple structures such

as lines or corners, whereas the deeper layers combine these elements into more abstract patterns. This hierarchical learning process is particularly effective for biometric traits such as faces, irises, and vascular structures, which contain multiple levels of spatial detail. CNNs also benefit from weight sharing, meaning that the same filter is applied throughout the image. This reduces the number of parameters and makes the model more efficient. CNNs became the foundation for modern face recognition and systems that achieve near-human performance on large benchmarks. They also performed strongly in iris and vein recognition, where fine spatial detail is crucial. Even for non-image signals such as PPG or ECG, CNNs have been successfully adapted by treating the signal as a one-dimensional sequence and applying convolutional filters along the time axis. Despite their success, CNNs have limitations. Their receptive field grows only gradually with depth, which makes it harder for them to model global relationships unless very deep architectures are used. They also introduce inductive biases: CNNs assume that local patterns are the most important structures, which may not always hold true for certain types of biometric signals. Furthermore, CNNs do not naturally provide interpretable attention maps, making it harder to justify their decisions. These limitations motivated the shift toward transformer-based models, which can model global relationships more easily and offer more flexible architectures for both image-based and signal-based biometric tasks.

2.4 Explainable Artificial Intelligence (XAI)

As biometric systems increasingly adopt deep-learning methods, understanding how these models make decisions becomes an essential goal. Deep neural networks can achieve high accuracy, but they often operate as black boxes. In high-security applications, this lack of transparency can be problematic, especially when the system needs to explain why a certain identity was accepted or rejected. Explainable AI (XAI) focuses on providing tools and techniques to interpret the internal behavior of machine-learning models. In biometrics, explainability helps evaluate whether a model relies on the correct biometric patterns or being influenced by noise, background elements, or unintended correlations in the data. This is particularly important in forensic contexts, where decisions may have legal implications. The following subsections introduce why explainability matters, how post-hoc methods are used, how transformers provide intrinsic interpretability, and what types of failure modes XAI can help detect.

2.4.1 Why Explainability matters

Explainability is important for several reasons. First, it increases trust in the system. If an operator or user can see which parts of the input influenced a decision, they are more likely to accept the system’s result. Second, explainability can uncover biases. For example, a facial-recognition model might focus on background shadows or clothing patterns instead of facial features. Third, in error analysis, explanations help identify why the system failed—whether the cause was noise, poor illumination, unexpected variations, or limitations of the model. Explainability is also essential in forensic scenarios, where biometric decisions may be used as evidence. In such cases, it is not enough for a model to be accurate; it must also be interpretable and defensible. For this reason, attention-based systems are gaining popularity because they naturally provide information on how decisions were made.

2.4.2 Post-hoc XAI methods

Post-hoc explainability methods analyze a model after it has been trained. They do not change the structure of the model but try to interpret its decisions by computing relevance scores or visual maps. Examples include Grad-CAM, which highlights the regions of an image that contributed the most to the model’s output, and Integrated Gradients, which measures the sensitivity of the output to each input pixel. In biometrics, post-hoc methods are useful for checking the focus of a model. For example, if a face-recognition network consistently highlights hair or background areas instead of facial features, this suggests that the model is learning the wrong patterns. For physiological signals, relevance maps can help verify whether the network is focusing on meaningful parts of the waveform or on noisy segments. Although post-hoc methods are not perfect—they sometimes providing approximations rather than true internal reasoning—they remain a valuable tool in understanding model behavior.

2.4.3 Intrinsic explainability via Attention mechanism

Unlike convolutional neural networks, transformer-based models naturally produce attention maps as part of their internal computation [48]. These maps explicitly quantify the importance assigned to different parts of the input during the aggregation of information. In biometric tasks, this property enables the analysis of which regions of the face, iris, or vein pattern most strongly influence the learned representation and, consequently, the final decision of the model. Intrinsic explainability is particularly valuable

because it reflects the actual operation of the model, rather than an approximation obtained through post-hoc analysis. When attention maps highlight anatomically meaningful biometric regions—such as the iris stroma, vein bifurcations, or stable segments of ECG or PPG signals—this behavior supports the reliability and biometric relevance of the learned representations. For example, *Albano et al.* [49] demonstrated that Vision Transformers trained on wrist-vein images consistently focus on discriminative vascular patterns. In contrast, attention maps that emphasize irrelevant areas can reveal potential weaknesses, biases, or overfitting phenomena. In this paper, attention-based explainability is exploited as an analytical tool to investigate how transformer models allocate relevance across biometric inputs. By examining the spatial distribution and stability of attention maps, it becomes possible to gain insights into which structures are considered discriminative by the model and to assess whether such behavior aligns with biometric expectations. However, it is important to note that attention maps should not be interpreted as formal causal explanations of models’ decisions. Although they provide valuable insight into how information is weighted and aggregated, attention mechanisms do not guaranty a direct correspondence with causal feature importance. For this reason, attention-based explainability is employed in this work as a complementary diagnostic instrument, to be interpreted in conjunction with performance metrics and robustness analyzes, rather than as a standalone proof of correctness.

2.4.4 Failure modes

Even biometric models that achieve strong performance under standard evaluation protocols may fail when exposed to adverse or unforeseen operating conditions. Such failure modes are often difficult to detect when system behavior is assessed exclusively through aggregate performance metrics. Explainability techniques provide a complementary perspective by enabling analysis of how and why these failures occur at the level of internal model behavior. One common failure mode arises from an over-reliance on background patterns or acquisition-specific artifacts. When training datasets exhibit consistent contextual cues surrounding biometric samples, models can inadvertently learn spurious correlations rather than physiologically meaningful biometric features. This phenomenon is particularly critical in image-based modalities, where background textures, borders, or illumination patterns may dominate the model’s attention. Additional failure modes are related to data quality and variability. Blurred images, incorrect segmentation, sensor noise, or physiological distortions can alter input in

ways that disrupt feature extraction and matching. In security-oriented scenarios, such as Presentation Attack Detection, these effects may be further exacerbated by intentional manipulations or by degradations introduced through compression and preprocessing pipelines. Model-related factors also contribute to failure modes. Architectural limitations, insufficient capacity to generalize to unseen identities, or an excessive sensitivity to specific feature patterns may lead to unstable decision behavior, particularly in open-set or cross-session conditions. In such cases, models with similar recognition accuracy may exhibit fundamentally different internal reasoning processes. Explainability methods make these failure modes observable by revealing how relevance is distributed across the input space. By examining attention or relevance maps, it becomes possible to determine whether classification errors originate from noise sensitivity, poor feature localization, shortcut learning, or attack-induced artifacts. This level of analysis is essential not only for diagnosing errors, but also for guiding the design of more robust biometric systems and for validating their behavior in real-world and security-critical applications.

2.5 Evaluation protocols

The evaluation of biometric systems requires carefully defined protocols in order to ensure fair comparisons, reproducibility, and a meaningful interpretation of experimental results. Unlike conventional classification tasks, biometric evaluation must take into account different operational scenarios, population assumptions, and security requirements. Consequently, evaluation protocols are closely related to the nature of the biometric task, the structure of the data set, and the intended deployment conditions of the system. This section introduces the main evaluation paradigms adopted in biometric research and outlines the learning objectives and performance metrics used throughout this thesis.

2.5.1 Biometric evaluation scenarios

Biometric evaluation protocols are mainly characterized by the distinction between verification and identification tasks, as well as by the assumptions made about the identities observed during the testing. These aspects play a crucial role in determining both the difficulty of the task and the interpretation of the results obtained. In verification scenarios, the system performs a one-to-one comparison to assess whether a biometric sample belongs to a claimed identity. This paradigm is widely adopted in access control and personal authentication applications, where the objective is to minimize false

acceptances and false rejections under well-defined operating points. Identification, on the other hand, involves one-to-many comparisons, requiring the system to determine the identity of an individual among a set of enrolled subjects. This setting is typical of surveillance, forensic, and border-control applications and generally imposes higher computational and discriminative requirements. In addition to the task formulation, biometric evaluation protocols differ in terms of whether the identities present during the tests are known to the system. Closed-set evaluations assume that all test identities are included in the training set, often leading to optimistic performance estimates. Open-set scenarios are more challenging and realistic as the system must generalize to identities that were never observed during training. In such conditions, the discriminative quality, robustness, and stability of learned feature representations become central factors, particularly for deep learning-based biometric systems.

2.5.2 Learning objectives and loss functions

The choice of the loss function directly influences the structure of the feature space learned by a biometric model and, consequently, its behavior under different evaluation scenarios. Although standard cross-entropy loss is effective for closed-set classification, it does not explicitly enforce the properties required for biometric verification and open-set recognition, where generalization to unseen identities is essential. Metric-learning approaches address this limitation by encouraging samples from the same subject to be close in the embedding space, while pushing samples from different subjects further apart. Loss functions such as contrastive loss and triplet loss explicitly model these relationships, promoting embeddings that are suitable for similarity-based matching. Margin-based softmax losses, including ArcFace, further enhance discriminability by introducing angular margins that improve inter-class separation and intra-class compactness. These learning objectives are particularly relevant in biometric contexts, where the quality of the learned representation has a direct impact on the robustness, scalability, and interpretability of the system. Throughout this thesis, loss functions are selected and analyzed with respect to their ability to support verification, open-set evaluation, and explainability-driven analysis.

2.5.3 Performance metrics and operating points

The performance of biometric systems is typically assessed through metrics that quantify the system's ability to separate genuine attempts from impostor attempts under different operating conditions. In verification sce-

narios, commonly used metrics include the False Match Rate (FMR), False Non-Match Rate (FNMR), and the Equal Error Rate (EER), often visualized through Receiver Operating Characteristic (ROC) or Detection Error Trade-off (DET) curves. These measures allow the analysis of system behavior across a range of operating thresholds. For identification tasks, performance is commonly reported using Rank-1 accuracy and Cumulative Match Characteristic (CMC) curves, which describe the probability that the correct identity appears within the top-ranked matches. In the context of Presentation Attack Detection, evaluation relies on metrics such as the Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and their average (ACER) [46], which reflect the system’s ability to discriminate between genuine and attack samples. While these metrics provide essential quantitative indicators, they offer limited insight into the internal decision mechanisms of deep biometric models. For this reason, in this thesis, performance metrics are interpreted in conjunction with explainability analyses, allowing numerical results to be complemented with an examination of how models allocate relevance across biometric inputs. This combined evaluation perspective supports a more comprehensive assessment of robustness and reliability, particularly in security-critical and real-world deployment scenarios.



Chapter 3

Methodology

This chapter describes the methodology adopted in this thesis to study transformer-based models for vein and iris biometrics. The goal is to provide a clear and reproducible description of the datasets used, the preprocessing and augmentation applied, the model architectures and fine-tuning strategies, and the experimental protocols defined for verification, presentation attack detection (PAD), and compression analysis.

The chapter is organized as follows. Section 3.1 introduces the general methodological framework. Section 3.2 presents the datasets, distinguishing between the vein and iris data and detailing how they are prepared for training and testing. Section 3.3 describes the transformer-based architectures adopted in this work—Vision Transformers (ViT) and Swin Transformers—and explains how they are fine-tuned and how attention maps are extracted. Finally, Section 3.4 illustrates the evaluation protocols and metrics used in all experiments.

3.1 Introduction

All experiments presented in this thesis are based on deep learning models trained and evaluated on real biometric data acquired under heterogeneous and realistic conditions. The experimental framework is designed to investigate not only recognition performance, but also robustness and interpretability, considering different biometric modalities, attack scenarios, and signal degradation factors. In this work, two complementary biometric traits are analyzed. On the one hand, wrist-vein images are employed to study biometric recognition and explainability under open-set verification conditions. Vein-based biometrics are based on subcutaneous vascular patterns that are difficult to forge and exhibit a high degree of discrimina-

tive power, making them well suited for identity verification tasks [?, 9, 10]. At the same time, wrist-vein acquisition is affected by illumination variability, sensor positioning, and image quality, making this modality particularly relevant for studying model robustness and the interpretability of learned representations. However, iris images from several publicly available Presentation Attack Detection (PAD) benchmarks are used to evaluate Transformer-based models in security-critical scenarios. Iris biometrics are known for their high recognition accuracy under controlled conditions, but are also vulnerable to a wide range of presentation attacks, including printed images, textured contact lenses, and replayed digital displays [8, 30, 50]. For this reason, iris PAD constitutes a well-established and challenging testbed to assess both the discriminative capability and the reliability of modern deep learning architectures. In addition to presentation attacks, this thesis explicitly investigates the impact of signal degradation introduced by image compression. Regarding iris biometrics, the effects of JPEG compression on recognition system performance have been addressed in several research works. In particular, the impact of JPEG-2000 compression has been investigated in [51, 52], where different compression rates were applied to iris images and performance was evaluated through fractional Hamming distances between iris code templates extracted from original and decompressed samples. Subsequent studies have extended this analysis to JPEG and other lossy and lossless codecs, confirming that compression may significantly affect the fine-grained textural information required for reliable iris recognition [53–57]. These findings motivate the inclusion of compression as a controlled experimental variable in the evaluation of Transformer-based biometric models. To ensure fair and reproducible comparisons across modalities, datasets, and experimental conditions, a common processing and evaluation pipeline is adopted throughout this thesis. Raw images are first preprocessed to normalize spatial resolution, contrast, and region of interest, reducing variability unrelated to the biometric trait. During training, data augmentation techniques are applied to increase sample diversity and mitigate overfitting. Pretrained Transformer architectures are then fine-tuned on the target biometric task and employed either as feature extractors or as end-to-end classifiers, depending on the evaluation scenario. A central element of the experimental analysis is the extraction and examination of attention maps from Transformer models. These maps provide insight into which regions of the input images contribute the most strongly to the model’s decisions, enabling an interpretability-driven evaluation that complements traditional performance metrics. Finally, all experiments are conducted following standard biometric evaluation protocols, including open-set verification and PAD evaluation, and performance is measured using

widely accepted metrics defined in the biometric literature.

3.2 Datasets

The experimental analysis relies on two main families of datasets: vein datasets, used for biometric recognition and explainability studies, and iris datasets, used for Presentation Attack Detection (PAD) and compression-related experiments. Only publicly available and widely adopted databases are considered, in order to ensure reproducibility and facilitate comparisons with results reported in the existing literature.

3.2.1 Vein Datasets

Vascular biometrics has attracted increasing interest in recent years due to the intrinsic properties of vein patterns, which are internal to the human body, difficult to forge, and relatively stable over time [?,9,10]. In this thesis, wrist-vein images are used to analyze biometric recognition performance and the interpretability of attention-based models under open-set verification conditions.

Specifically, the experimental evaluation focuses on two public wrist-vein datasets: the PUT Wrist-Vein database and the FYO multi-modal vein database. Both datasets are frequently adopted in the literature on wrist-vein recognition and have also been used in recent studies involving deep learning and Transformer-based architectures, making them suitable benchmarks for comparative analysis.

The PUT Wrist-Vein dataset [58] contains 1200 near-infrared images collected from 50 subjects. For each subject, both wrists are acquired; since left and right wrist vein patterns exhibit low correlation, each wrist is treated as an independent class, resulting in 100 distinct classes. Four images per wrist are captured in each of three acquisition sessions separated in time, enabling the analysis of multi-session variability. The images are characterized by low contrast between the vascular structures and surrounding tissue, which makes recognition particularly challenging and motivates the adoption of enhancement techniques. The FYO database [59] is a multi-modal vein dataset that includes palmar, dorsal, and wrist images. In this work, only the wrist-vein subset is considered. It contains data from 160 subjects, with images acquired from both wrists during two acquisition sessions, for a total of 640 wrist images. Compared to PUT, FYO images are captured using a medical vein finder in a controlled environment and generally exhibit higher quality and clearer vascular patterns. In Tab. 3.1 you can see the details of the datasets used. The coexistence of these two datasets allows

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

Table 3.1: Employed wrist-vein dataset characteristics.

the evaluation of model behavior under different acquisition conditions and image quality levels.

Pre-processing and Augmentation

A common preprocessing pipeline is applied to wrist-vein images to ensure consistency across datasets and compatibility with Transformer-based architectures. The images are resized to match the input resolution required by the models (e.g., 224×224 pixels), while preserving the original aspect ratio by padding when necessary. Pixel intensities are then normalized according to the ImageNet statistics used during pretraining, which facilitates effective fine-tuning of pretrained models.

Due to the particularly low contrast of the PUT images, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied as an enhancement step prior to training and testing [60]. CLAHE redistributes the intensities of the pixels within local regions while limiting the amplification of the noise; the clip limit and the grid size are selected according to the settings commonly adopted in the literature (clip limit of 5.0 and grid size of 8×8). This enhancement significantly improves the visibility of vascular structures and leads to better recognition performance and more meaningful attention maps. An example of this improvement is shown in Fig. 3.1.

Given the limited number of available samples, data augmentation plays a crucial role in reducing overfitting and improving generalization. The same augmentation strategy described in [49] is adopted. Each original image may undergo random horizontal flipping, small rotations, contrast adjustments, and illumination variations, applied with predefined probabilities. Through these transformations, the PUT dataset is expanded from 12 to 132 images per class, while the FYO dataset is expanded from 2 to 30 images per

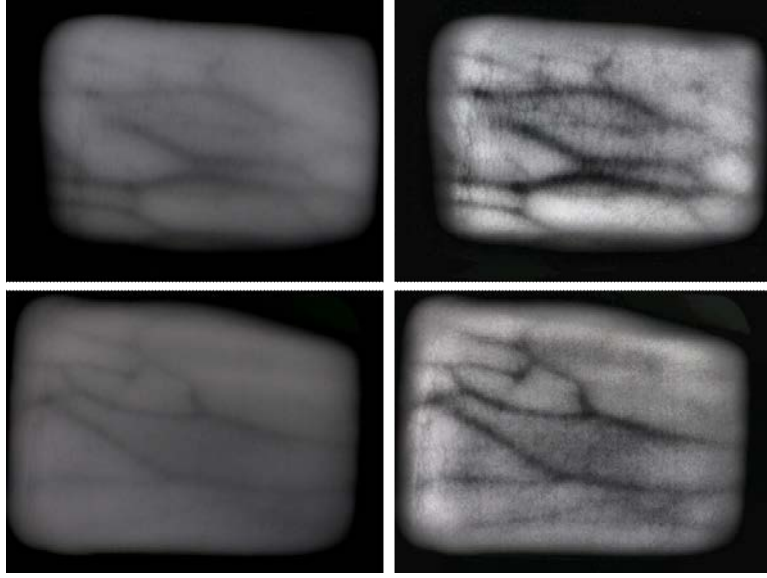


Figure 3.1: Original wrist-vein patterns (left) and their CLAHE-enhanced representations (right) from the PUT Wrist-Vein dataset.

class, resulting in a substantially richer training set for Transformer models. Although this thesis focuses on the PUT and FYO datasets, it is worth noting that several other vascular biometric datasets are widely used in the literature, including the CASIA and PolyU palm-vein databases, multiple finger-vein datasets, and more recent wrist-vein collections designed for smartphone-based acquisition [9] [?]. The existence of these datasets confirms the growing interest in vascular biometrics and provides a wider context for the methodological choices adopted in this work.

3.2.2 Iris Datasets

For iris Presentation Attack Detection (PAD) and compression-related experiments, this thesis relies on data from the NDCLD2015 dataset and from three subsets of the LivDet-Iris 2017 competition, namely Notre Dame, Clarkson, and IIITD-WVU. These databases represent the main reference benchmarks in the recent iris PAD literature and have been extensively adopted to evaluate traditional deep-learning approaches and more recent Transformer-based methods [30, 40, 50]. The NDCLD2015 data set [61] contains approximately 7 300 iris images acquired in controlled environments. The data set includes bona fide iris images without contact lenses, images acquired with soft lenses, and samples captured with textured contact lenses, which are considered presentation attacks. Following common prac-

tice in the literature, 6 000 samples are used for training, while the remaining images are reserved for testing. This data set provides a well-controlled experimental setting that is particularly suitable for evaluating baseline PAD performance and the effects of image degradation under relatively stable acquisition conditions.

The LivDet-Iris 2017 database [62] was explicitly designed to assess the robustness of the iris PAD algorithms under challenging and realistic conditions, including cross-presentation attack and cross-dataset scenarios. The official data partitions provided by the competition are adopted to define training and testing sets, ensuring consistency with prior evaluations reported in the literature.

The Notre Dame subset includes 4 800 iris images depicting both bona fide samples and images captured with textured contact lenses. The corresponding test set contains attacks generated using the same lens manufacturers present in the training set (referred to as the *known* set) and lenses not observed during training (the *unknown* set). This configuration enables a systematic evaluation of generalization with respect to unseen presentation attack instruments.

The Clarkson subset contains 4 937 images in the training set, including authentic iris samples, printed iris images, and images acquired with textured contact lenses. The testing set consists of 3 158 iris images and includes printed attacks characterized by visual features that differ from those observed during training. This data set therefore supports the analysis of PAD performance under variations in attack realization and print quality.

Finally, the IIITD-WVU subset provides one of the most challenging evaluation scenarios. The training set comprises 6 250 images collected in constrained indoor environments and includes bona fide irises, textured lenses, printouts, and lens printout attacks. The test set contains 4 209 iris images acquired from a different group of subjects, using different sensors, and under both constrained and unconstrained (outdoor) conditions. This configuration enables the evaluation of PAD models under simultaneous variations in subjects, sensors, environments, and attack types, making it particularly suitable for assessing robustness and generalization.

In general, the combination of the NDCLD2015 and LivDet-Iris 2017 subsets allows the evaluation of PAD systems in iris not only in standard within-dataset settings, but also under cross-dataset, cross-sensor and cross-presentation attack conditions, which are widely considered essential for realistic security assessments [29, 63].

Pre-processing and Augmentation

In the experiments presented in this thesis, iris images are used as provided by the original datasets, without applying additional segmentation or iris normalization beyond standard resizing and channel normalization. This choice is consistent with the experimental protocol adopted by *Albano et al.* [64], where the impact of image compression is evaluated directly on the samples used to train and testing PAD models. By avoiding task-specific preprocessing steps, the analysis focuses on the intrinsic robustness of the learned representations rather than on handcrafted normalization procedures.

All images are resized to the input resolution required by the Vision Transformer and Swin Transformer models (e.g., 224×224 pixels) and normalized using the same statistics employed during ImageNet pretraining, which facilitates effective fine-tuning of pretrained architectures. During training, data augmentation is applied through random horizontal flips, small random rotations, and variations in brightness and contrast, following the protocol described in the reference work. These transformations help models cope with minor misalignments, illumination changes, and acquisition noise, and have been shown to improve generalization performance, particularly in cross-dataset and cross-sensor PAD evaluations.

3.3 Model architectures

The core of this thesis is the analysis of transformer-based biometric architectures. Two main models are considered: Vision Transformers (ViT) and Swin Transformers. Both are initially trained in the ImageNet-1K dataset and then fine-tuned for the specific biometric tasks considered: open-set vein recognition, iris PAD, and robustness under compression.

3.3.1 ViT configuration

Vision Transformers [65] treat an image as a sequence of patches. Each image is split into fixed-size non-overlapping patches (e.g., 16×16 or 32×32 pixels), which are flattened, linearly projected to a fixed-dimensional vector, and then combined with positional encodings. The resulting sequence is fed to a stack of transformer encoder blocks composed of multi-head self-attention and feed-forward layers. A special class token aggregates information from all patches and is used to produce the final representation. Fig. 3.2 illustrates the architecture of the model. This approach is very different from CNNs, which focus mainly on local neighborhoods. ViTs instead allow

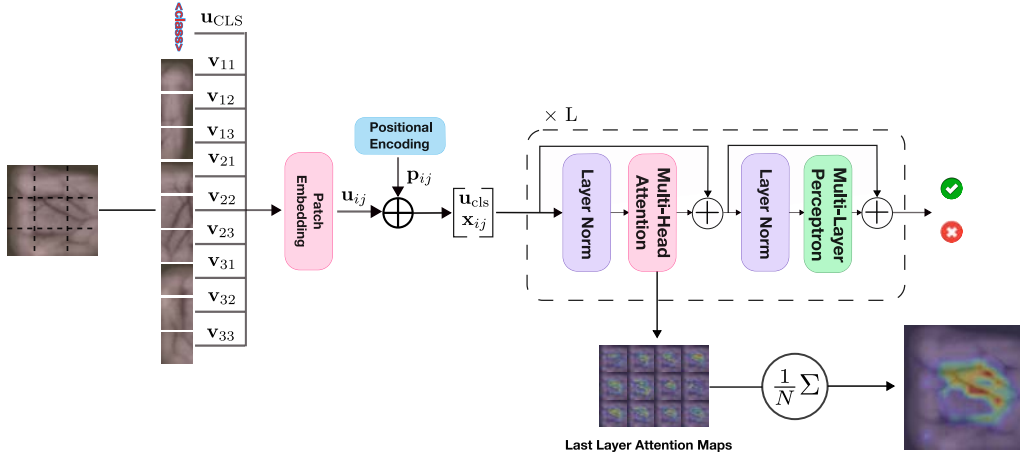


Figure 3.2: Illustrative representation of the Vision Transformer (ViT) [5] tailored for explainable vein biometric recognition.

each patch to interact with every other patch, making it possible to capture long-range relationships. This global processing is particularly beneficial for biometric modalities where identity is encoded in patterns distributed throughout the image—such as iris texture, vascular structures, or facial geometry. ViTs also provide attention maps that show the importance assigned to each patch during the forward pass. These maps are an important tool for analyzing how the model makes decisions and for checking whether the focus is on meaningful biometric regions rather than irrelevant background or artifacts. Although ViTs often require large datasets to perform at their best, transfer learning and fine-tuning strategies help adapt them even in cases where training data are limited, such as in some biometric applications. In vein recognition experiments, two ViT variants are employed, with patch sizes 16×16 and 32×32 , to study the effect of patch granularity on the ability to capture fine vascular patterns. Both models use 12 attention heads and produce a 768-dimensional embedding, which is later used for verification. The architectures are imported with weights pretrained in ImageNet-1K, providing a strong initialization that helps when training on relatively small biometric datasets.

Fine-tuning strategies

For vein recognition, the last layers of the ViT are replaced with a small projection head composed of two linear layers with a dropout layer in between, to improve regularization. The network is then fine-tuned using stochastic gradient descent (SGD) with momentum 0.9, learning rate 5×10^{-4} , batch

size 128 and early stopping based on validation loss, following the settings in the explainable ViT vein paper. Models are trained for up to 1000 epochs, with training stopped when no improvement is observed for 10 consecutive epochs. For PAD iris, ViTs are fine-tuned separately on each iris data set using cross-entropy loss, SGD optimizer, and batch size 128, again starting from ImageNet-pre-trained weights. Data augmentation is applied online at training time. The output head is adjusted to match the number of PAD classes, which can be two (genuine vs attack) or three (e.g., live, soft lens, textured lens) depending on the dataset configuration.

Attention maps extraction

Explainability is studied by analyzing the attention maps produced by the ViT. The extraction procedure follows the algorithm proposed by *Albano et al.* [49]. For a given input image, attention weight matrices are collected from the last transformer layer for all heads. These matrices are averaged across heads to obtain a single attention map, which is then up-sampled to the original image resolution and overlaid on the image. In the context of verification, the maps are computed before the final embedding or classification, so they reveal which regions of the input the model focuses on while building its internal representation, independently of the final class. This process is applied both to wrist-vein and iris images, enabling a comparison of how ViTs treat relevant structures such as vascular patterns or iris stroma in different tasks and conditions.

3.3.2 Swin Transformers

Swin Transformers *Liu et al.* [66] introduce a hierarchical vision transformer architecture that computes self-attention within local windows that are shifted between layers. This design reduces computational cost while still allowing information to flow throughout the image. Swin models build multi-scale feature maps, similar to CNN pyramids, which is advantageous for biometric traits that contain information at different spatial scales. Swin Transformers were introduced to improve the efficiency and flexibility of ViTs. They apply self-attention within small windows that move—or “shift”—around the image over layers. This mechanism captures both local details and global context while keeping computational costs manageable. For biometrics, Swin Transformers offer several advantages. The hierarchical structure, which gradually reduces spatial resolution while increasing feature depth, allows the model to learn multi-scale patterns. This is very relevant for biometric traits, where both global shape and small local textures matter. For example, in vein or iris recognition, fine-grained lines

and textures are crucial, but the overall structure also provides useful identity information. Another practical advantage is robustness to variations in resolution and image size, making Swin Transformers particularly suitable for datasets acquired with heterogeneous sensors or in real-world environments. In this paper, Swin Transformer variants trained in ImageNet-1K are adopted for iris PAD. The models are configured with standard window sizes and depth as provided by their original implementation, and the final classification head is replaced to match the number of PAD classes.

Fine-tuning strategies

Swin models are fine-tuned using cross-entropy loss and SGD with batch size 128, closely mirroring the training setup used for ViT on iris PAD. Given their hierarchical nature, all layers are usually fine-tuned but with a lower learning rate for earlier stages to preserve beneficial low-level representations. Data augmentation (flip, rotation, brightness and contrast changes) is applied in the same way as for ViT to ensure a fair comparison between architectures.

Attention maps extraction

Attention maps for Swin Transformers are obtained by collecting the attention weights inside each local window of the last layers and then reconstructing them over the whole image by aligning the windows according to their positions. Because Swin is hierarchical, attention maps from different stages can be upsampled to a common resolution and combined to provide a global view of where the model focuses at different scales. As with ViT, the maps are visualized as heatmaps overlaid on the original images, allowing a qualitative analysis of the cues used by the model to detect attacks or perform recognition.

3.4 Experimental protocols

This section describes the experimental protocols adopted to evaluate the proposed models in different and complementary scenarios, including open-set verification for vein biometrics, Presentation Attack Detection (PAD) for iris data, robustness analysis against image compression, and similarity scoring strategies. The goal is to ensure fair, reproducible, and application-oriented evaluations, in line with established practices in the biometric literature [6, 19, 41].

3.4.1 Open-set verification

For wrist-vein recognition, experiments are conducted in an open-set verification scenario, where the identities used for testing are disjoint from those used during training. This evaluation setting closely reflects real-world deployments, in which a system trained on a given population is required to generalize to previously unseen users [6, 32]. The protocol is defined as follows. A subset of classes (wrists) is used to train the model, while a different subset is reserved for testing. The trained Vision Transformer (ViT) acts as a feature extractor: for each image, the embedding associated with the class token is extracted and used as a biometric template [34]. Verification trials are then generated by pairing images from the same class (genuine pairs) and from different classes (impostor pairs). For each pair, the cosine similarity between embeddings is computed, producing a similarity score, which is a commonly adopted metric in embedding-based biometric systems [41]. From the distributions of genuine and impostor scores, Receiver Operating Characteristic (ROC) and Detection Error Trade-off (DET) curves are derived, and the Equal Error Rate (EER) is used as the primary performance indicator [19]. To reduce statistical variability, multiple folds are considered by randomly changing the partition between training and testing classes, and the results are reported as mean and standard deviation in all folds, following established evaluation practice.

3.4.2 PAD evaluation

For iris Presentation Attack Detection, the objective is to classify each input image as bona fide or attack, or into more detailed attack classes depending on the dataset. Models are trained using the official NDCLD2015 splits and the LivDet-Iris 2017 subsets, following the protocols defined by the respective benchmarks [40, 61, 62]. For NDCLD2015 and Clarkson, three-class training configurations (e.g., live, soft lens, textured lens; or live, textured lens, printouts) are adopted, as they have been shown to improve generalization, even when the final evaluation merges soft lenses with bona fide samples [30, 50]. Performance is evaluated using standard PAD metrics defined in ISO/IEC 30107 and widely adopted in the literature:

- Attack Presentation Classification Error Rate (APCER): proportion of attack samples misclassified as bona fide;
- Bona fide Presentation Classification Error Rate (BPCER): proportion of bona fide samples misclassified as attacks;

- Average Classification Error Rate (ACER): arithmetic mean of APCER and BPCER.

These metrics allow for a consistent comparison across datasets and algorithms [40, 67]. In addition, cross-attack and cross-dataset evaluations are considered, where models trained on one data set are tested on another. Such experiments are essential to assess generalization across sensors, environments, and unseen attack instruments [29, 63].

3.4.3 Compression experiments

To analyze the effect of compression on the PAD performance of the iris, both traditional JPEG compression [68] and the recent deep-learning-based JPEG-AI standard are considered. JPEG-AI [69] has been designed not only to preserve human visual quality but also to support machine vision tasks at very low bit rates, making it particularly relevant for biometric applications. In the compression experiments, iris images from the PAD datasets are compressed at multiple quality factors for JPEG and at different target bitrates for JPEG-AI. Two experimental strategies are commonly considered in the literature:

- training on original images and testing on compressed images, to evaluate robustness without retraining;
- training and testing on images compressed at the same level, to estimate the best achievable performance under a given compression regime.

Following the reference protocol adopted in [64], PAD models are trained on original images and tested on compressed samples without additional preprocessing, to isolate the impact of compression alone. Performance is evaluated using APCER, BPCER, and ACER for each codec and compression level [43, 53].

3.4.4 Similarity scoring

For vein verification, the ViT model produces a feature embedding for each wrist image. The cosine similarity between embeddings is used as the matching score, as it is well suited for normalized feature spaces [41]. A decision threshold is applied to determine whether a given pair is accepted as genuine or rejected as an impostor. By varying this threshold, the ROC and DET curves are obtained, and the EER is computed as the operating point

where the false match and false non-match rates are equal. For iris PAD, model output is interpreted as posterior probabilities over classes. Scores are derived from the attack or bona fide probabilities depending on the metric of interest. In binary configurations, decision thresholds can be adjusted to trade off APCER and BPCER; however, in most experiments, the standard maximum softmax probability is used, which is common practice in classification-based PAD systems [30].

In the compression experiments, the same scoring rules are applied consistently across codecs and compression levels, enabling a direct and fair comparison of performance under different operating conditions.



Chapter 4

Explainable wrist-vein biometric Recognition

4.1 Introduction

The biometric recognition of the Wrist-vein represents a highly secure modality for the verification of identity due to its intrinsic resistance to spoofing, the stability of the subcutaneous vascular patterns and the difficulty of externally replicating the structures of the vein [70], [6]. Traditional vein-based systems relied on handcrafted features, such as local binary patterns (LBP), maximum curvature extraction, or SIFT descriptors [71], [72], which provided interpretable representations but limited discriminative capability. With the advent of deep learning, convolutional neural networks (CNNs) significantly improved performance [73]; however, they also introduced opacity in the decision-making process, raising concerns regarding transparency, robustness, and the ability to guaranty that model decisions rely on genuine vascular information. Vision Transformers (ViTs) have recently emerged as a powerful paradigm in computer vision, replacing convolutional priors with global self-attention mechanisms and offering intrinsic explainability through attention maps. Unlike CNNs, ViTs operate on sequences of image patches [65] and learn long-range dependencies directly from the input, enabling them to highlight the image regions that most influence their final decision. This property is particularly advantageous in biometric recognition, where understanding why a model identifies a user—or fails to do so—is essential for evaluating trustworthiness and diagnosing potential vulnerabilities. In this Chapter, we investigate the suitability of ViT architectures for wrist-vein biometric verification and evaluate their explainability by an-

alyzing the attention maps [49] produced during the recognition process. Using the PUT and FYO wrist-vein datasets [58, 59], we demonstrate that ViTs not only achieve competitive verification performance but also reveal which vascular regions contribute most to identity discrimination. These visual explanations provide unprecedented insights into model behavior and highlight the dependence of recognition accuracy on image quality, pre-processing (e.g., CLAHE enhancement [60]), and training strategies. The material presented in this Chapter is based on the study reported in Albano et al.

4.2 Vision Transformer–based Explainability Framework

The explainability component of the proposed wrist-vein recognition system is intrinsically embedded in the Vision Transformer (ViT) architecture. Unlike convolutional models, which require post-hoc methods [74] to expose their internal reasoning, ViTs rely on self-attention mechanisms that naturally provide a transparent view of the decision-making process. This Section outlines the framework adopted to extract, process, and interpret attention maps, enabling a detailed understanding of how discriminative vascular features are leveraged during biometric verification.

4.2.1 Patch Tokenization and Embedding

In the ViT pipeline, each wrist-vein image is partitioned into a grid of non-overlapping patches of predefined size (e.g., 16×16 or 32×32 pixels). These patches are flattened and projected into a latent space through a learned linear embedding, generating a sequence of fixed-length tokens. A special classification token (jcls_i) [65] is pretended to this sequence, serving as a global representation that aggregates information across patches. This token generation process plays a central role in explainability: because each token corresponds to a specific region of the image, the attention distribution across tokens provides a spatially meaningful indication of which areas contribute most to the recognition process.

4.2.2 Self-Attention Mechanism

Each transformer layer computes self-attention, which models the relationships among all image patches. Attention weights quantify the contribution of each patch to the transformation of another, effectively highlighting the

regions deemed most relevant for the downstream verification task. For each layer l and each attention head i , the attention weights are computed as [22]:

$$A_i^{(l)} = \text{softmax}\left(\frac{Q^{(l)}(K^{(l)})^T}{\sqrt{d}}\right) \quad (4.1)$$

where Q , K and V denote the query, key and value projections of the input token sequence. These attention matrices encode the internal reasoning process of the model: patches assigned high attention weights exert the strongest influence on the representation of the network and final prediction.

4.2.3 Aggregation of Attention Maps for Explainability

To obtain a stable and interpretable explanation of the behavior of the model, attention maps are extracted from the last layer of the transformer, which captures the discriminative features of the highest-level. Since each layer contains multiple attention heads, their outputs are averaged to produce a single robust map [49]:

$$A = \frac{1}{h} \sum_{i=1}^h A_i^{(L)} \quad (4.2)$$

where h is the number of heads and L is the final layer. This averaging step reduces noise and head-specific variability, resulting in a unified attention representation that clearly highlights the vascular regions driving the decision.

The final attention map is then up sampled and superimposed on the original image, producing a heat map where warmer colors correspond to a higher contribution in the verification process. This visualization provides an intuitive and anatomically meaningful explanation of the focus of the model.

Algorithm 1 Get Attention Maps**Input:** Patch size P **Output:** Predicted class $\hat{y}$, Attention map $\mathcal{A}(I)$

```

1: procedure TOKENIZATION( $I$ )
2:    $\mathbf{P} \leftarrow \text{Split}(I)$  ( $\triangleright$ )  $N=HW_{\overline{P^2}}$  patches of size  $P \times P$ 
3:    $\mathbf{v}_i \leftarrow \text{vec}(\mathbf{P}_i)$  ( $\triangleright$ )  $\mathbf{v}_i \in \mathbb{R}^{P^2C}$ 
4:    $\mathbf{u}_i \leftarrow \mathbf{v}_i \mathbf{W}_e$  ( $\triangleright$ )  $\mathbf{W}_e \in \mathbb{R}^{(P^2C) \times d}$ 
5:   return  $\mathbf{u}$  ( $\triangleright$ )  $\mathbf{u} \in \mathbb{R}^{N \times d}$ 
6: end procedure
7: procedure POSEMBEDDING( $\mathbf{u}$ )
8:    $\mathbf{x}_i \leftarrow \mathbf{u}_i + \mathbf{p}_i \quad \forall i$ 
9:    $\mathbf{X}^{(0)} \leftarrow [\mathbf{u}_{\text{cls}}, \mathbf{x}_1, \dots, \mathbf{x}_N]$ 
10:  return  $\mathbf{X}^{(0)}$ 
11: end procedure
12: procedure SELFATTENTION( $\mathbf{X}^{(l)}$ )
13:   $\mathbf{Q}^{(l)} \leftarrow \mathbf{X}^{(l)} \mathbf{W}_Q^{(l)}$ 
14:   $\mathbf{K}^{(l)} \leftarrow \mathbf{X}^{(l)} \mathbf{W}_K^{(l)}$ 
15:   $\mathbf{V}^{(l)} \leftarrow \mathbf{X}^{(l)} \mathbf{W}_V^{(l)}$ 
16:   $\mathbf{A}^{(l)} \leftarrow \text{softmax}\left(\frac{\mathbf{Q}^{(l)}(\mathbf{K}^{(l)})^\top}{\sqrt{d}}\right)$ 
17:   $\mathbf{O}^{(l)} \leftarrow \mathbf{A}^{(l)} \mathbf{V}^{(l)}$ 
18:  return  $\mathbf{O}^{(l)}, \mathbf{A}^{(l)}$ 
19: end procedure
20:  $\mathbf{u} \leftarrow \text{TOKENIZATION}(I)$ 
21:  $\mathbf{X}^{(0)} \leftarrow \text{POSEMBEDDING}(\mathbf{u})$ 
22: for  $l = 0$  to  $L - 1$  do
23:    $\hat{\mathbf{X}}^{(l)} \leftarrow \text{LN}(\mathbf{X}^{(l)})$ 
24:    $\mathbf{O}^{(l)}, \mathbf{A}^{(l)} \leftarrow \text{SELFATTENTION}(\hat{\mathbf{X}}^{(l)})$ 
25:    $\tilde{\mathbf{X}}^{(l)} \leftarrow \mathbf{X}^{(l)} + \mathbf{O}^{(l)}$ 
26:    $\mathbf{X}^{(l+1)} \leftarrow \tilde{\mathbf{X}}^{(l)} + \text{FF}(\text{LN}(\tilde{\mathbf{X}}^{(l)}))$ 
27: end for
28:  $\mathcal{A}(I) \leftarrow \frac{1}{H} \sum_{h=1}^H \mathbf{A}_h^{(L)}$ 
29: return  $\mathcal{A}(I)$ 

```

4.2.4 Interpretability of Vascular Regions

One of the most relevant aspects of the proposed explainability framework lies in its ability to provide a direct correspondence between the regions highlighted by the attention mechanism and the anatomical vascular patterns present in wrist-vein images. In traditional deep neural networks,

the identification of discriminative features often requires post-hoc interpretability methods such as saliency maps, LayerCAM, or gradient-based techniques [74, 75]. However, these methods often suffer from instability and may not capture the true internal reasoning of the model [18, 76]. In contrast, the Vision Transformer inherently exposes its decision-making process through attention weights that are structurally tied to the input patches and are therefore spatially interpretable. When analyzing the attention maps obtained from the last transformer layer, it becomes evident that the model consistently focuses on vascular-rich areas—particularly the branches of the main veins, the bifurcations, and the segments of the high-contrast vessels. These regions are known to contain highly discriminatory biometric information and their prominence in the maps confirms that the model has successfully learned to identify stable and meaningful features. In addition, as training progresses through successive epochs, the attention maps become more localized and sharper, indicating a refinement in the internal representation of the model. An additional insight arises from cases where the model incorrectly identifies a subject or produces uncertain similarity scores. In such situations, the attention maps tend to appear dispersed or fragmented, reflecting a lack of confidence or reliance on non-informative patterns. This phenomenon illustrates how explainability can also serve as a diagnostic tool: by observing which regions influence the decision, one can infer whether the model’s behavior is reliable, whether the input is of sufficient quality, or whether additional preprocessing steps may be needed. Thus, interpretability not only provides human-readable visualization, but actively contributes to evaluating and improving the robustness of the recognition pipeline.

4.2.5 Impact of Pre-processing on Attention Explainability

Image preprocessing plays a fundamental role in not only enhancing recognition accuracy but also shaping the quality and usefulness of the attention maps. The PUT data set represents a clear example of this dependency. Due to its inherently low contrast and illumination variability, raw images often exhibit faint vascular structures, making it challenging for the model to identify meaningful features. When training a ViT directly on non-enhanced PUT images, attention maps tend to highlight broad, noisy, or peripheral regions instead of focusing on actual veins. These maps show a degree of uncertainty that corresponds to a measurable drop in verification performance. By applying CLAHE (Contrast Limited Adaptive Histogram Equalization), vessel patterns become more pronounced and ho-

mogeneous across subjects and sessions. This enhancement not only supports feature extraction, but also guides the attention mechanism towards consistent and anatomically relevant areas. Interestingly, models trained on CLAHE-enhanced data show robust interpretability even when evaluated on non-enhanced images. In this case, the attention maps continue to emphasize the vascular structures, demonstrating that preprocessing contributes to shaping the internal representation of the model in a way that generalizes beyond the training domain. This relationship highlights an important methodological insight: attention maps can be used as an evaluative tool for preprocessing pipelines. Instead of relying solely on accuracy metrics, one can assess preprocessing choices by examining how clearly and consistently the attention mechanism localizes vein structures. Such an analysis provides a more holistic understanding of model behavior, revealing not only whether the network performs well, but also whether it relies on the appropriate cues.

4.2.6 Role of Explainability in Verification

Explainability plays a dual and complementary role within the proposed wrist-vein recognition system. On the one hand, it provides a transparent window into the inner workings of the Vision Transformer, enabling practitioners to verify whether the model bases its decisions on valid biometric cues. However, it functions as a practical tool for evaluating the reliability of individual predictions, the effectiveness of preprocessing steps, and the overall robustness of the recognition pipeline. In biometric systems, especially those deployed in sensitive or security-critical scenarios, understanding why a model accepts or rejects a sample is as important as achieving high accuracy. Vision Transformers are particularly advantageous in this regard because their attention mechanism inherently localizes the regions of interest that contribute to the decision. By inspecting the attention maps associated with a verification attempt, one can determine whether the model correctly focused on vein structures or whether it may have been misled by noise, background textures, or acquisition artifacts. This capacity is especially valuable when analyzing mismatches or borderline similarity scores. In such cases, traditional metrics provide little insight into the cause of the error, whereas attention maps reveal whether the model struggled due to insufficient image quality, subtle anatomical differences, or reliance on unstable cues. As a result, explainability enhances not only interpretive transparency, but also practical diagnostic capability. Furthermore, explainability supports the design of more robust biometric systems. By evaluating the consistency of attention patterns across subjects, sessions,

and datasets, developers can identify failure modes, assess the generalizability of the model, and refine preprocessing steps or training strategies accordingly. Ultimately, the integration of explainability transforms Vision Transformer from a black-box classifier into a tool whose internal reasoning can be scrutinized, validated, and optimized—thus advancing the broader goal of trustworthy biometric recognition.

4.3 Experiments

The experimental analysis conducted in this chapter has two main goals. First, we evaluated the verification performance achievable by Vision Transformer models on wrist-vein images under realistic open-set conditions, considering different training configurations and preprocessing strategies. Second, we investigate how these factors influence the attention maps, and consequently the explainability of the recognition process. The proposed framework is assessed on two public wrist-vein databases, PUT and FYO. Each data set contains near-infrared images of both wrists acquired in multiple sessions, with the left and right wrists of the same subject treated as distinct classes. In PUT, 50 subjects are acquired in three sessions, with four images per wrist and session, resulting in 100 vein classes and 1200 images. Due to the low contrast between the vessels and the background, images from PUT are typically enhanced using CLAHE, which increases the visibility of venous structures. FYO contains images from 160 subjects acquired in two sessions with a medical vein finder; higher image quality makes contrast enhancement less critical. For both datasets, classes are partitioned into training, validation, and test splits according to a subject-disjoint protocol that is later exploited to simulate open-set verification scenarios. Given the limited number of samples per class, data augmentation is applied to increase intra-class variability while preserving the underlying vascular pattern. The adopted pipeline includes random horizontal flips, small in-plane rotations, and controlled variations in contrast and illumination. Each operation is activated with a fixed probability, so several augmented versions of each training image are generated. As a result, the number of images per class is increased from 12 to 132 for PUT and from 2 to 30 for FYO, providing a richer set of examples to fine-tune the ViT models.

Two ViT configurations are considered, differing in patch size. The first uses patches of 16×16 pixels, while the second relies on 32×32 patches. Both architectures are initialized from ImageNet-1K [77] pretrained weights and share the same internal structure, with 12 attention heads and a final MLP producing 768-dimensional representations. To improve generaliza-

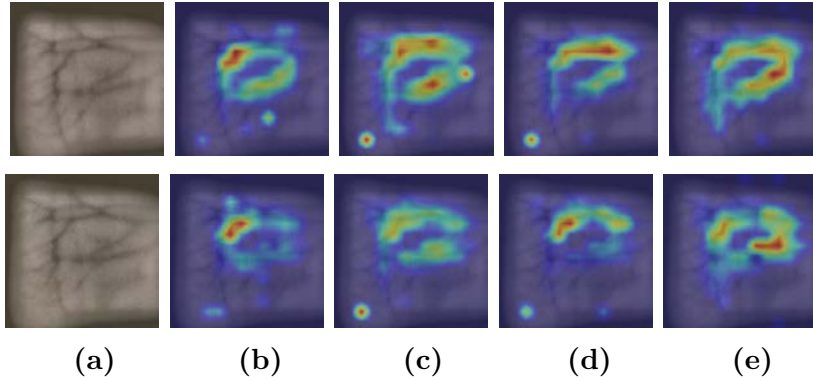


Figure 4.1: Attention maps of ViTs trained for different numbers of epochs, for two different images of the same wrist from the PUT dataset.

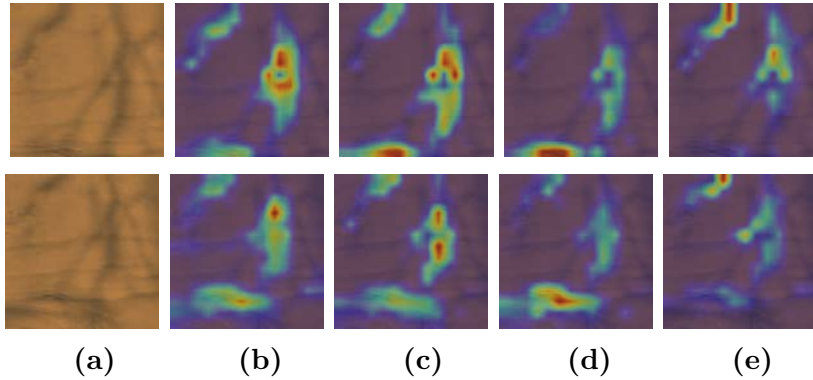


Figure 4.2: Attention maps of ViTs trained for different numbers of epochs, for two different images of the same wrist from the FYO dataset. .

tion, the original classification head is replaced with a small two-layer head that includes dropout. The models are trained with stochastic gradient descent with momentum 0.9, cosine learning-rate decay, and a batch size of 128. Fine-tuning is performed for different numbers of epochs (20, 40, 60, and 100), allowing us to analyze how the evolution of the training process affects both recognition performance and the quality of attention maps, Fig. 4.1 and Fig. 4.2. In , warm colors signify higher concentration on attention. 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, and column (e) 100 epochs. All experiments are implemented in PyTorch using the TIMM library and executed on a multi-GPUs server equipped with four NVIDIA Tesla V100 GPUs. The verification protocol is designed to mimic realistic open-set conditions. At each fold of a 10-fold cross-validation, a subset of classes is excluded from the fine-tuning stage

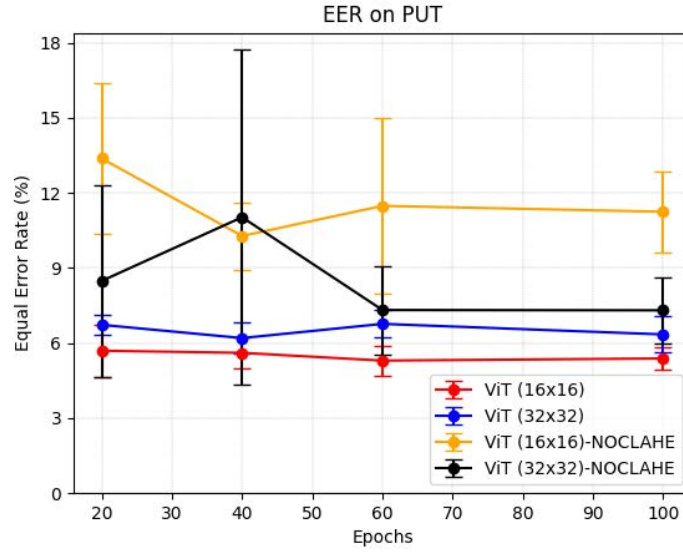


Figure 4.3: EERs of ViT models trained on PUT wrist vein images either with or without CLAHE preprocessing.

and reserved for testing (25 classes for PUT and 60 for FYO). After training, the ViT is used as a feature extractor for all images belonging to the held-out classes. Pairwise similarity scores are then computed between feature vectors, generating genuine score distributions (comparisons between images of the same class) and impostor distributions (comparisons between different classes). By sweeping a decision threshold over these scores, we estimate the Equal Error Rate (EER) and the Area Under the ROC Curve (AUC), which provide a compact summary of the verification performance. Only original, non-augmented images are included in the test set to avoid artificially inflating the results. Quantitatively, the best configuration for PUT is obtained with a ViT using 16×16 patches, fine-tuned for 100 epochs on CLAHE-preprocessed images, as shown in Fig. 4.3.

A qualitative inspection of the attention maps further complements these findings. For both PUT and FYO, attention maps associated with well-trained models show a strong concentration on the main vascular branches, particularly around bifurcations and high-contrast segments, while the background and non-informative regions receive negligible attention. As the number of training epochs increases, attention becomes more localized and less noisy, suggesting that the model progressively refines its internal representation and learns to ignore spurious artifacts. This effect is especially visible for PUT, where early-epoch models sometimes distribute attention over large regions of the wrist or the image borders, whereas later-epoch

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

Table 4.1: Recognition performance comparison between approaches based on ViTs and ResNet152 on the PUT-Wrist dataset.

models sharply highlight the vein skeleton. The experiments on CLAHE also reveal a strong link between preprocessing and explainability. When ViTs are trained without CLAHE on low-contrast PUT images, attention maps often fail to clearly distinguish vessel regions, and the corresponding EER is considerably higher. In contrast, models trained on CLAHE-enhanced images produce attention maps that align closely with the visible vein patterns, even when tested on non-enhanced inputs, as shown in Fig. 4.4. This behavior demonstrates that explainability can be used as a diagnostic tool: attention maps not only justify the final scores, but also expose whether the model is able to consistently focus on meaningful biometric structures under different acquisition or preprocessing conditions.

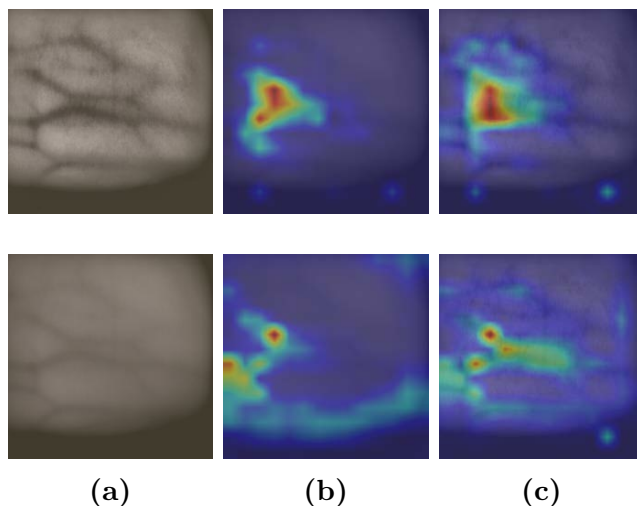


Figure 4.4: Effects of CLAHE on ViTs trained on PUT verification

Finally, to assess the benefits of ViTs, we compare their performance with a strong convolutional baseline based on ResNet152 [78], fine-tuned under the same open-set protocol. In PUT, the ViT trained with CLAHE sub-

stantially outperforms ResNet152, reducing the EER from approximately 15–16% to approximately 5% and increasing the AUC by several percentage points. In FYO, ViT and ResNet152 obtain comparable AUC values, but ViT achieves a lower EER, highlighting more favorable operating points in verification scenarios, as shown in Tab. 4.1. In general, these results confirm that transformers can extract highly discriminative vein features and, at the same time, expose their decision process through attention maps, overcoming some of the limitations of CNN-based approaches.

4.4 Summary and Conclusion

This chapter has investigated the use of Vision Transformers for explainable wrist-vein biometric recognition. Starting from the general framework introduced in Chapter 3, we focused here on a specific case study in which self-attention mechanisms are exploited not only to improve verification accuracy but also to reveal the internal reasoning of the model through attention maps. From a methodological point of view, we adopted ViTs pretrained on ImageNet and fine-tuned them on the PUT and FYO wrist-vein datasets, using data augmentation and transfer learning to cope with the limited number of samples per subject. The models were trained according to an open-set verification protocol, where a subset of users is excluded from training and used exclusively for testing. In this setting, the ViT is employed as a feature extractor and verification scores are computed by comparing feature vectors associated with enrollment and probe images. The explainability framework is based on aggregating the attention matrices of the last transformer layer, producing attention maps that can be directly overlaid on the input images to highlight the regions that most influence the decision. The experimental results demonstrate that ViTs can achieve competitive or superior performance with respect to a strong CNN baseline, especially in the more challenging PUT dataset, where the combination of CLAHE preprocessing, an appropriate patch size, and sufficient fine-tuning leads to significant improvements in EER and AUC. At the same time, associated attention maps consistently emphasize venous structures, confirming that the network does not rely on background artifacts or data set-specific biases. Qualitative analysis in different epochs and preprocessing strategies shows that better verification performance is systematically associated with sharper, more localized attention patterns focused on vascular regions. In general, the findings presented in this chapter support two main conclusions. First, Vision Transformers represent an effective solution for wrist-vein biometric verification in open-set scenarios, capable of extracting discriminative representations from relatively small

datasets. Second, the attention maps naturally produced by these models provide a powerful tool for explainability: they make it possible to verify that decisions are grounded on meaningful biometric cues, to assess the impact of preprocessing and training choices, and to diagnose potential failure modes. In the broader context of this thesis, this case study illustrates how attention-based architectures can be used to design biometric systems that are not only accurate, but also transparent and trustworthy—an aspect that will be further explored in the following chapters for other biometric traits and application scenarios.



Chapter 5

Iris PAD with Transformers

The iris has long been considered one of the most reliable biometric traits due to its high degree of uniqueness, stability over time, and resistance to environmental noise [8, 33]. However, the widespread adoption of iris recognition systems in consumer devices, border control infrastructures, and large-scale identity programs has greatly increased the incentive for malicious actors to attempt Presentation Attacks (PA). These attacks consist of presenting an artificial representation of the iris—such as high-quality printed photos, glossy prints, digital replays on mobile screens, textured contact lenses, or prosthetic eyes—to the sensor with the goal of impersonating a legitimate user [30, 31]. As these attacks become increasingly sophisticated, traditional biometric systems alone are no longer sufficient to guaranty security [40], and Presentation Attack Detection (PAD) modules have become essential. PAD for iris images is particularly challenging for several reasons. First, the iris is a highly structured texture, and many attack instruments are capable of reproducing this structure with high fidelity. Second, genuine iris samples exhibit a rich set of micro-textural cues, complex variations of reflectance, stroma fibers, radial and angular patterns, and subtle photometric properties that are not trivially captured by simple descriptors [33]. Third, variations in acquisition devices—including near-infrared (NIR) sensors, visible-light cameras, smartphone cameras, and cross-spectral systems—introduce strong domain shifts that complicate generalization [50, 63]. Each sensor operates with different optics, lighting conditions, noise characteristics, and preprocessing pipelines, meaning that a PAD method trained on one data set may fail on another if not designed with sufficient robustness. For example, the influence of several factors on PAD of the iris has been evaluated by considering the role of image segmentation [79], demographic aspects of the subjects involved [80], and gender-related biases [81].

Deep learning has significantly improved PAD performance compared to handcrafted approaches [11, 82]. Convolutional neural networks (CNNs) have shown the ability to detect subtle spoof signals, including printer dot patterns, interference patterns effects in digital displays, and lens boundaries in textured contact lens attacks [83, 84]. Nevertheless, CNN architectures exhibit several inherent limitations when applied to iris Presentation Attack Detection. A first critical aspect concerns the lack of interpretability: although convolutional filters operate at multiple spatial scales, their individual contribution to the final classification decision is not directly accessible [15, 16]. As a result, it becomes difficult to identify which specific regions of the iris have influenced the decision, limiting the ability to verify whether the model relies on genuine biometric information rather than on spurious cues. Another important limitation is related to sensitivity to compression and pre-processing operations. CNN-based models tend to rely heavily on high-frequency image components, which are often attenuated or removed by common compression standards such as JPEG and JPEG2000 [43, 85]. This dependence makes convolutional architectures particularly vulnerable to image quality degradation, leading to a significant drop in PAD performance when images are subjected to compression or aggressive enhancement procedures [44, 64]. CNNs also suffer from a pronounced localization bias. During training, these models tend to focus on image regions that maximize discriminative power, even when such regions are not biologically meaningful or directly related to the texture of the iris [20, 29]. This behavior can cause the network to exploit dataset-specific artifacts or unintended correlations, increasing the risk of overfitting and introducing vulnerabilities that may be exploited by sophisticated presentation attacks. Finally, convolutional architectures show limited capability in modeling global inconsistencies. Many presentation attacks introduce anomalies that manifest at a global level, such as unnatural contrast uniformity or large-scale texture regularities in printed iris images. However, because CNNs primarily extract features through localized receptive fields, they struggle to capture these global discrepancies, which can reduce their effectiveness in distinguishing bona fide samples from attacks. These limitations become critical when biometric systems operate in bandwidth-limited or unconstrained environments, where iris images undergo lossy compression before transmission or storage [?]. Compression may remove or distort micro-textures essential for PAD, leading to a significant drop in performance. Understanding how PAD behaves under compression and how to design models resilient to such distortions is therefore a crucial open problem.

Vision Transformers (ViTs) introduce a paradigm shift in iris Presenta-

tion Attack Detection [34, 64, 82] by departing from the localized processing paradigm of convolutional architectures and instead modeling image content through global self-attention mechanisms [22]. By operating on sequences of image patches, ViTs are able to capture long-range dependencies and globally coherent texture patterns, which are particularly relevant in the analysis of iris structures and presentation attacks. This global modeling capability reduces the dependence on localized high-frequency details, thereby increasing robustness to image compression and other quality-degrading operations [43]. At the same time, the attention mechanism provides a transparent view of the decision-making process, allowing the extraction of attention maps that reveal which regions of the iris contribute the most to the classification outcome [23–25]. As a result, ViTs exhibit a higher degree of generalization across datasets acquired with different sensors or involving various attack instruments, addressing some of the main limitations observed in CNN-based PAD systems [50, 63].

Within this context, the two studies presented in this thesis—one focused on analyzing iris PAD under JPEG, JPEG2000, and JPEG-AI compression, and the other dedicated to the development and evaluation of a ViT-based PAD architecture—constitute one of the most comprehensive investigations to date on the role of Transformer models in iris anti-spoofing. These works provide a systematic evaluation of PAD robustness across multiple compression formats and quality levels, offering a controlled analysis of how different codecs affect both recognition performance and security [?, 85, 86]. Furthermore, they include a detailed comparison between Vision Transformers, convolutional neural networks, and other deep learning models under identical distortion conditions, highlighting the distinctive advantages of attention-based architectures. Beyond performance metrics, the studies also incorporate an explanation analysis that reveals how self-attention mechanisms respond to spoof-related cues, offering insight into the internal reasoning of the model. The results demonstrate that modern codecs, such as JPEG-AI, are able to preserve discriminative information more effectively than classical compression standards, even at high compression rates [69]. Crucially, the analysis shows that Vision Transformers maintain stable and meaningful attention patterns even under severe image degradation, reinforcing their suitability for robust and reliable iris Presentation Attack Detection [23]. This chapter elaborates these findings and integrates them into a coherent narrative that highlights both methodological innovation and practical relevance. It also emphasizes how explainability plays a central role in operational security: a system that merely outputs a decision is insufficient for high-risk environments, while a transparent Transformer-based PAD module can justify its classification by pointing to

specific iris regions that contain false evidence [26, 27].

5.1 ViT for Presentation Attack Detection

The adoption of Vision Transformers for the detection of iris presentation attacks represents a substantial shift in the way biometric security systems interpret and classify visual information. Traditional CNN-based PAD systems are based mainly on locally processed texture cues. Although effective in controlled conditions, such networks remain sensitive to variations introduced by acquisition devices, illumination changes, and image compression pipelines. Vision Transformers (ViTs), on the other hand, process images using a fundamentally different principle: global self-attention, which allows them to establish long-range relationships between distant parts of the iris, integrate information from multiple spatial regions simultaneously, and adaptively focus on the areas containing the strongest evidence of spoofing. This Section provides an extensive, multi-layered description of the ViT-based PAD framework, covering image tokenization, patch embeddings, self-attention behavior, training strategies, and an in-depth analysis of why Transformers exhibit strong resilience to compression and cross-dataset variation. The discussion is grounded in the experimental observations and methodological contributions of the two studies conducted on iris PAD with Transformers and under image compression.

5.1.1 Image Representation, Patch Tokenization, and Embedding

An iris image contains highly structured information distributed in fine repetitive patterns such as crypts, freckles, radial flints, contraction flints, pigment spots, and textural granularity. Capturing these cues is essential for distinguishing genuine samples from presentation attacks. However, CNN filters often emphasize very local information, which can lead to poor generalization when textural details are partially corrupted or altered by compression algorithms. The Vision Transformer treats the image not as a continuous texture but as a sequence of discrete patches, each representing a portion of the iris. This design was originally inspired by natural language processing, where inputs are converted into token words, forming a sequence. In the visual domain, patches become “visual tokens, allowing the network to handle images similarly to sentences.

Patch Extraction

Given an input image resized to a fixed spatial resolution, such as 224×224 pixels, the proposed framework partitions the image into a set of non-overlapping patches that constitute the basic tokens processed by the Vision Transformer. When a patch size of 16×16 pixels is adopted, the image is decomposed into 196 patches, while a larger patch size of 32×32 pixels results in a more compact representation composed of 49 patches. This design choice directly affects the granularity with which visual information is encoded and subsequently analyzed by the self-attention mechanism. Patch size plays a crucial role in shaping how the model perceives and interprets iris patterns. Smaller patches, such as those of size 16×16 , preserve a higher level of fine-grained detail, enabling the network to capture subtle local artifacts that are often indicative of presentation attacks. These include, for example, contact lens boundaries, printer dot patterns, or small-scale texture inconsistencies that may not be uniformly distributed across the iris. By maintaining a dense spatial resolution, the model is better equipped to isolate and respond to these localized cues, which can be critical to accurate PAD decisions. In contrast, larger patches, such as 32×32 pixels, encourage the model to focus on broader structures and more global relationships within the iris image. This coarser representation emphasizes semantic and large-scale properties, which can be particularly informative when presentation attack cues are not purely textural but instead relate to global appearance characteristics, such as specular reflections, illumination artifacts, or abnormal reflectance patterns. By reducing the number of tokens, larger patches also promote a more holistic view of the iris, allowing self-attention to model long-range dependencies more explicitly. Overall, the choice of patch size represents a trade-off between preserving fine-scale detail and emphasizing global structure. As a consequence, it has a profound impact on both the discriminative capability of the model and its robustness in PAD scenarios, influencing the types of cues that are prioritized during the recognition process.

Patch Embedding and Positional Encoding

Each flattened patch is transformed into a feature vector through a learned linear projection. Because Transformers lack built-in spatial hierarchies, positional encodings are added to preserve the spatial organization of patches *Vaswani et al.* [22]. These encodings play a critical role in iris Presentation Attack Detection, as they allow the model to retain information about the relative and absolute positions of different regions within the iris image. In PAD scenarios, spatial relationships between iris regions carry a significant

semantic meaning. The ability to distinguish between central and peripheral regions, as well as to model their interactions, is essential for understanding whether observed patterns are consistent with a genuine biological structure or indicative of an attack. Moreover, several spoof signatures tend to manifest in predictable spatial locations. Examples include the presence of contact lens edges near the limbic boundary or printing defects and texture discontinuities close to the pupil area. Preserving positional information enables the model to associate such artifacts with their expected anatomical context. Positional encodings are also crucial for identifying global inconsistencies across distant regions of the iris. Presentation attacks often introduce anomalies that are not confined to a single local area but instead affect the overall spatial coherence of the image. By maintaining explicit positional cues, the self-attention mechanism can more effectively compare distant patches and detect irregularities that emerge at a global scale. As a result, positional encodings substantially enhance the ability of the Transformer to model spatial structure and to discriminate between bona fide samples and presentation attacks.

CLS Token Aggregation

Similarly to standard Vision Transformer classification tasks, a learnable CLS token is added to the sequence of patch tokens. Through successive self-attention operations across the transformer layers, this token progressively aggregates information from all patches and ultimately serves as the compact representation provided as input to the PAD classifier. The adoption of the CLS token introduces two key properties that are particularly relevant in the context of iris Presentation Attack Detection. First, it enables a global integration of spoof-related indicators. Rather than relying exclusively on localized texture cues, the final representation encoded by the CLS token reflects how such local patterns interact with the rest of the iris. This global perspective is essential in PAD, as presentation attacks often manifest through complex combinations of local artifacts and broader structural inconsistencies that cannot be reliably captured in isolation. Second, the CLS token plays a central role in explainability. The attention weights associated with the CLS token provide a direct indication of which patches contribute the most significantly to the final classification decision. By analyzing these attention distributions, it becomes possible to identify the iris regions that the model considers most informative when distinguishing between bonafide samples and attacks. This property offers valuable insight into the internal reasoning of the model and supports a transparent evaluation of whether decisions are based on meaningful bio-

metric cues. Overall, the combination of patch tokenization and CLS-based aggregation establishes the foundation for a globally informed PAD system capable not only of isolating complex patterns related to different attack instruments, but also of interpreting them in a coherent and explainable manner.

5.1.2 Self-Attention Mechanism for Iris PAD

The self-attention mechanism constitutes the defining element that makes Vision Transformers fundamentally different from convolutional neural networks, and it is particularly well suited for iris Presentation Attack Detection. Instead of relying on fixed local receptive fields, the Transformer constructs a representation in which each image patch is evaluated in relation to every other patch. This global interaction allows the model to assess whether different regions of the iris exhibit the coherence and structural continuity that naturally characterize bona fide samples. Genuine irises tend to have a globally consistent textural organization: radial stromal fibers, smooth pigment transitions, and photometric properties shaped by subsurface scattering. When self-attention processes such images, it typically amplifies these stable relations, producing attention patterns that are anatomically meaningful and spatially concentrated. Presentation attacks, on the contrary, exhibit internal inconsistencies that emerge precisely when distant regions are compared. Printed irises often contain dot patterns or repetitive textures that remain locally plausible but globally incompatible with real biological structure. Display attacks introduce moiré interference, backlight gradients, and pixel-grid artifacts that disrupt the natural flow of the iris texture. Textured contact lenses introduce a mismatch between the printed lens pattern and the genuine underlying iris, creating discontinuities that cannot be reconciled through global attention. Self-attention highlights these abnormalities because the relationships among patches do not conform to the internal model learned about the physiology of the iris. This behavior can be summarized in terms of three high-level tendencies that characterize the response of the model to genuine and spoofed iris samples. The genuine iris typically exhibits a globally coherent and self-consistent texture, reflecting the underlying biological structure of the iris and the natural correlations that exist in different regions. In contrast, presentation attacks tend to introduce structural or photometric deviations that disrupt this coherence through artificial textures, unnatural reflectance patterns, or inconsistencies arising from the attack medium. During the classification process, the model learns to associate such deviations with evidence of spoofing, leveraging these irregularities as discriminative cues to

distinguish attacks from bonafide samples. What distinguishes Transformers from CNNs is not merely the ability to detect local irregularities, but the ability to evaluate whether all image regions “agree” with each other. A patch that appears suspicious does not mislead the Transformer, unless its inconsistency reverberates across the entire representation. This relational criterion explains why Vision Transformers often outperform convolutional models on PAD tasks, especially when the differences between bonafide and attack samples are subtle or spatially distributed.

5.1.3 Training Strategy, Data Augmentation, and Fine-Tuning Protocol

Training a Vision Transformer for PAD of the iris requires careful design of the learning pipeline. Iris PAD datasets are typically small, attacks vary between sensors and acquisition environments, and the cues that distinguish real from fake samples may be extremely subtle. For this reason, naïve fine-tuning is insufficient: the training process must enforce robustness, encourage invariance to image degradations, and stabilize the formation of meaningful attention patterns. The model is initialized from large-scale ImageNet pre-training, which provides strong priors for visual structure. During fine-tuning, the network is exposed to a balanced mixture of bonafide and attack samples so that it does not learn shortcuts tied to specific class frequencies. The key challenge is to ensure that the model encounters sufficient variability to generalize beyond the conditions represented in a single data set. To this end, data augmentation plays a central role. Controlled variations in illumination, contrast, and spatial geometry emulate differences between acquisition devices. Mild blur and noise perturbations simulate lower-quality sensors. Most importantly, the use of synthetic compression forces the Transformer to learn discriminative cues that remain stable even when the high-frequency content of the image is altered. The optimization procedure consolidates these representations through a progressive refinement of attention dynamics. During training, attention tends to be diffuse and sensitive to irrelevant patterns; as learning progresses, attention gradually stabilizes over the iris regions where genuine structural and photometric cues reside. The complete training flow typically follows a structured optimization strategy designed to ensure stable convergence and robust generalization. An initial learning rate warm-up phase is employed to mitigate unstable gradients during the early stages of training, when model parameters are still far from an optimal configuration. This phase helps prevent abrupt updates that could negatively affect learning dynamics. As training progresses, a cosine annealing schedule is adopted

to gradually reduce the learning rate, allowing the network to refine its feature representations in a smooth and controlled manner. In addition, weight regularization and dropout are incorporated throughout the training process to limit overfitting. These mechanisms encourage the model to learn more generalizable patterns by constraining parameter growth and reducing excessive reliance on specific activations, thus improving robustness across different datasets and acquisition conditions. This combination of data diversity, optimization stability, and attention refinement ensures that the final model acquires a representation of the iris that remains reliable even under conditions far from the original training distribution.

5.1.4 PAD Under Image Compression

Evaluating PAD under image compression is essential because biometric systems rarely operate on pristine data. In real deployments—such as border control systems, mobile identity verification, or large-scale cloud-based matching—iris images are often compressed to meet bandwidth or storage requirements. However, compression is not just a technical convenience; it directly interferes with the visual evidence used to discriminate bona fide from spoof samples. Since many PAD cues reside in the fine micro-textural structure of the iris, any degradation of these details may affect the model’s ability to detect attacks.

Different compression standards affect the iris in distinct ways:

- JPEG removes high-frequency texture components and introduces block artifacts, severely distorting micro-patterns used to detect prints and contact lenses.
- JPEG 2000 avoids blocking, but smooths the image through wavelet reconstruction, diminishing the subtle irregularities that indicate spoofing.
- JPEG-AI, leveraging deep learning-based reconstruction, preserves both micro-textures and global structures more effectively, retaining discriminative cues far better than traditional codecs.

The experimental observations from your compression study show that the severity of PAD degradation is tightly coupled with the codec’s ability to preserve discriminative information. CNN-based PAD models typically collapse under strong compression because their filters rely on high-frequency textures that are destroyed first. Transformers, on the contrary, retain a surprising degree of stability. Even when the local texture is partially lost, the global structural relationships between the iris regions persist, allowing the

ViT to identify attack patterns that survive compression. Under JPEG-AI, this stability becomes particularly notable: the reconstructed images preserve enough discriminative structure that the Transformer’s performance remains close to that achieved on uncompressed data.

This demonstrates not only that compression introduces non-trivial challenges for PAD, but also that ViTs provide a natural architectural advantage in dealing with these challenges, especially when paired with modern learned codecs.

5.1.5 Why Transformers Show Superior Robustness to Compression

The remarkable robustness of Vision Transformers under compression arises from the fact that their representational strategy differs fundamentally from that of convolutional networks. CNNs extract features through spatially localized filters that respond strongly to fine detail. When compression eliminates or distorts high-frequency information, these filters lose access to the very signals upon which they depend. As a result, CNN-based PAD systems degrade sharply as the bitrate decreases. Transformers, on the other hand, derive meaning from relational information—how different regions of the iris compare and interact with each other. Even if compression corrupts a subset of patches, the global structure of the iris remains coherent enough for the self-attention mechanism to function effectively. Instead of relying on isolated textures, the Transformer assesses the compatibility of patches, and this compatibility is much more resilient to local degradation than pixel-level detail. This robustness can be understood through three complementary aspects that characterize the behavior of Vision Transformers under image degradation. First, the model relies on relational reasoning, basing its decisions on the coherence between different patches rather than on isolated local frequency components. As a consequence, the Transformer is inherently more tolerant to the loss of fine micro-texture details, which are often the most affected by compression, because it evaluates how patterns relate to one another across the entire iris. Second, the use of multi-head attention enables a form of specialization across different attention heads. Each head can focus on a distinct dimension of the problem, such as illumination consistency, structural geometry, or residual texture information. This specialization introduces redundancy in the representation, ensuring that when compression degrades certain cues, other complementary cues can still be exploited to support reliable PAD decisions. Finally, the robustness of the Transformer is reinforced by the structural stability of the iris at a global level. Even under aggressive compression, the overall organization

of the iris—encompassing radial flow, large-scale texture arrangements, and reflectance patterns—remains detectable. Using this preserved global structure, the Vision Transformer is able to maintain effective PAD performance in scenarios where convolutional neural networks, which rely more heavily on localized detail, tend to fail. In essence, the Transformer succeeds not because it ignores fine detail, but because it supplements it with a holistic understanding of the iris. Compression may erase local evidence, but it rarely destroys the broader irregularities that betray a presentation attack. The Transformer is therefore better positioned to exploit what remains, making it an inherently more reliable architecture for real-world PAD systems that cannot rely on perfectly preserved imagery.

5.2 Results

The experimental analysis conducted in the context of this work provides a comprehensive understanding of how Vision Transformers behave in iris Presentation Attack Detection under varying acquisition conditions, datasets, and image quality levels. The results highlight not only the discriminative power of ViT-based architectures but also their robustness to compression and cross-sensor variations, and ultimately how their internal mechanisms contribute to improved reliability. The section presents these findings in a structured yet discursive manner, integrating observations from multiple experiments and providing a unified interpretation of the empirical evidence.

5.2.1 Experimental Setup and General Observations

The evaluation involved multiple publicly available PAD datasets, each characterized by distinct attack types, acquisition devices, and image statistics. This diversity creates a challenging experimental scenario that reflects realistic operational conditions. The datasets—such as Clarkson, IIITD-WVU, and NDCLD’15—include bonafide samples alongside a variety of spoof attempts, including printed photographs, replay attacks using digital displays, and textured contact lenses. This broad corpus ensures that the model is not tailored to a narrow class of attacks but learns discriminative cues that generalize across spoofing modalities. Fig. 5.1 reports the attention maps obtained from the original images, providing insight into how the model focuses on different iris regions when textured contact lenses, no contact lenses, or soft contact lenses are present. Transformers were trained under several configurations that varied in terms of patch size, training duration, and data preprocessing pipeline. Although the exact optimal parameters differed slightly from one dataset to another, the overall trend showed

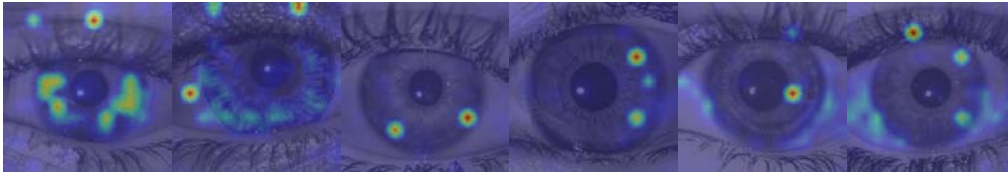


Figure 5.1: Attention maps for original images: textured lenses in the first two columns on the left, no lenses in the two middle ones, soft lenses in the last two on the right.

that Vision Transformers quickly surpassed convolutional baselines as soon as they were allowed to leverage their global reasoning capabilities. In early epochs, performance metrics were only moderately superior to CNNs; however, once the self-attention patterns stabilized, typically around mid-to-late training, the ViT began to consistently outperform convolutional models by significant margins. This dynamic supports the hypothesis that the Transformer discovers spoof cues that CNNs are structurally incapable of representing.

5.2.2 Performance on Datasets

In datasets such as NDCLD15 that contain challenging textured contact lens attacks—the Transformer demonstrated a strong advantage. Contact lenses are often difficult to detect because they introduce subtle alterations in the iris pattern rather than the explicit artifacts present in printed or replay attacks. CNN models trained on NDCLD15 frequently focused on spurious local structures or noise patterns, leading to inconsistent generalization between subjects and sensors. The ViT, by contrast, learned to identify global inconsistencies in the lens boundary region, illumination anomalies introduced by lens reflectance, and slight geometric discontinuities that arise when an artificial pattern overlays the natural iris, as shown in Fig. 5.2. These cues remain discernible even when the textural detail is limited.

A similar trend was observed in the Clarkson dataset, where the ViT displayed a marked ability to separate bonafide samples from spoof samples despite variations in acquisition environments. Digital display attacks—which often exhibit faint, subvisible grid structures, or faint aliasing interference—were consistently identified by the Transformer, even at lower resolution or under compression.

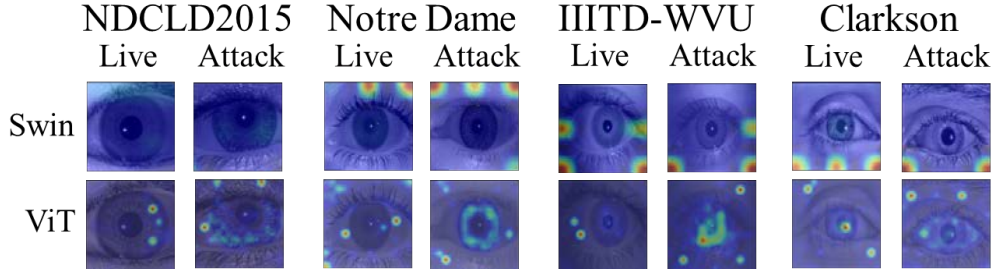


Figure 5.2: Attention maps generated by ViT and Swin models for bonafide and spoofing test images from the databases considered.

Database	Metric	D-NetPAD	PBS	A-PBS	d-CBAM	DFCANet	ViT	Swin
NDCLD2015	BPCER	-	0.00%	0.06%	0.60%	21.35%	0.00%	0.00%
	APCER	-	1.09%	0.08%	0.80%	7.16%	0.00%	0.00%
	ACER	-	0.54%	0.07%	0.70%	13.13%	0.00%	0.00%
Notre Dame	BPCER	3.32%	1.06%	0.00%	4.37%	-	0.61%	0.00%
	APCER	10.38%	8.89%	7.88%	11.67%	-	17.38%	6.42%
	ACER	6.81%	4.97%	3.94%	8.02%	-	9.00%	3.21%
IIITD-WVU	BPCER	10.12%	8.26%	4.13%	-	6.75%	21.51%	11.41%
	APCER	36.41%	5.76%	8.86%	-	16.80%	24.07%	13.39%
	ACER	23.27%	7.01%	6.50%	-	11.78%	22.79%	12.40%
Clarkson	BPCER	0.94%	0.00%	0.81%	0.87%	5.08%	0.35%	0.40%
	APCER	5.78%	8.97%	6.16%	5.79%	1.81%	9.10%	4.02%
	ACER	3.36%	4.48%	3.48%	3.33%	3.44%	4.72%	2.21%

Table 5.1: Iris PAD performance of the proposed methods, and comparison with literature algorithms. [1–4].

5.2.3 Influence of Patch Size and Training Duration

One of the most insightful observations concerns the impact of patch size. Smaller patches (e.g. 16×16) tend to preserve more micro-texture details, enabling the model to detect fine-grained evidence associated with spoof attempts. Larger patches (e.g. 32×32) produce a more abstract representation, emphasizing structural coherence rather than micro-level discrepancies. Although small patches often performed slightly better on datasets with strong textural information, larger patches proved advantageous under compression, where local detail is partially lost. The complementary behavior of these two settings suggests that a hybrid or dynamic patch-size strategy may further improve the robustness of the PAD. The number of training epochs also influenced the final performance. During initial training, attention distributions were unstable and often incorporated irrelevant regions such as eyelids or background artifacts. As training progressed, attention maps became more anatomically meaningful, converging on areas of the iris where bona fide vs attack discrepancies naturally appear. This stabilization corresponded with significant improvements in verification metrics,

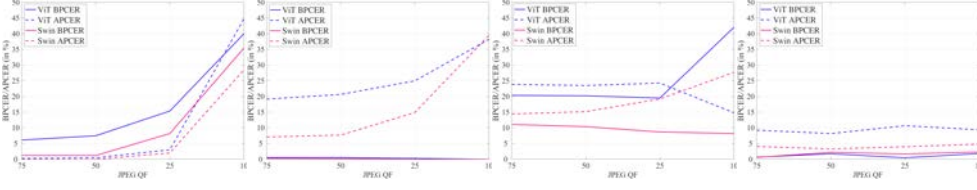


Figure 5.3: Iris PAD performance for increasing JPEG compression levels. In order NDCLD15, Notredame, IIIT-WVU and Clarkson

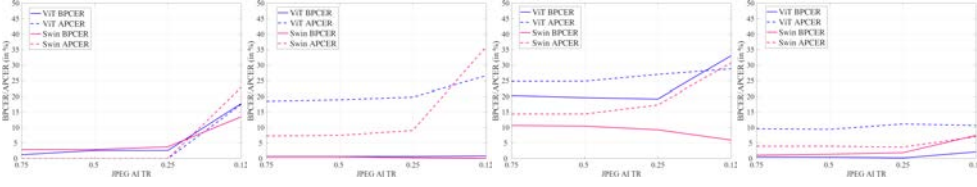


Figure 5.4: Iris PAD performance for increasing JPEG AI compression levels. In order NDCLD15, Notredame, IIIT-WVU and Clarkson

reflecting a correlation between representational maturity and classification accuracy.

5.2.4 PAD Under Compression: Quantitative and Qualitative Insights

The compression experiments reveal one of the most significant contributions of your study: systematic comparison of PAD performance under JPEG, JPEG2000, and JPEG-AI compression. Although CNN models showed dramatic performance degradation as compression increased—especially under JPEG—Vision Transformers retained much of their discriminative ability. This robustness emerges from two synergistic effects. First, many spoof signatures remain visible after compression, although they appear weaker or more distorted. Transformers, which rely heavily on relational cues, can still identify inconsistencies between iris regions, even when local texture is partially erased. Second, JPEG-AI compression preserves discriminative information significantly better than classical codecs. Under JPEG-AI, the PAD performance with a ViT remained near the levels obtained with uncompressed images over a wide range of bitrates, demonstrating that advanced learned compression formats align naturally with transformer-based perception.

Qualitatively, attention maps under compression provide additional ev-

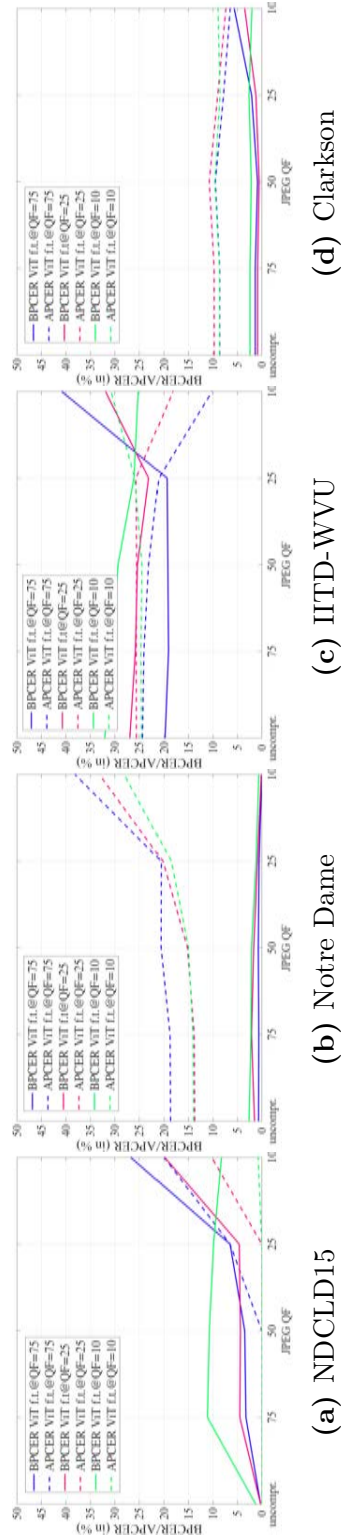


Table 5.2: Iris PAD performance for ViT models fine-tuned with images JPEG-compressed at different QFs, at increasing JPEG compression levels.

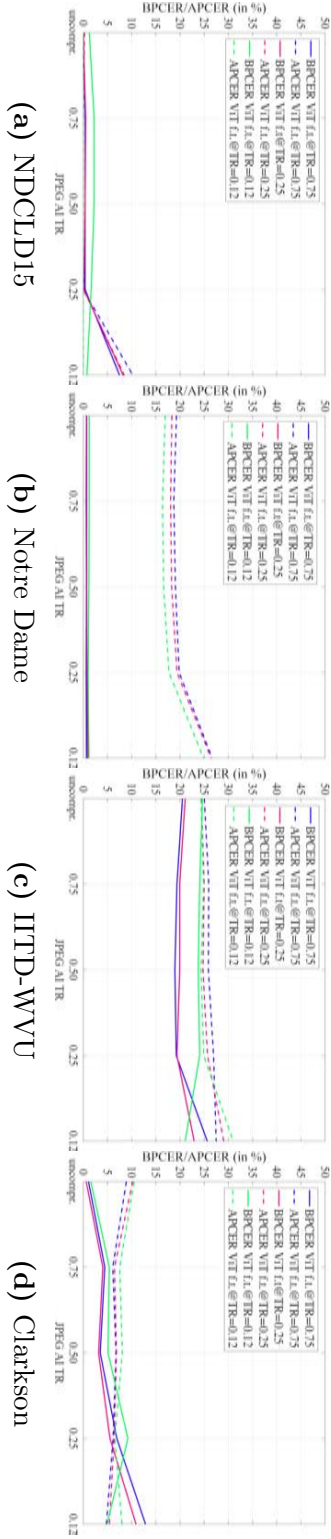


Table 5.3: Iris PAD performance for ViT models fine-tuned with images JPEG-AI-compressed at different TRs, at increasing JPEG AI compression levels.

Metric	QF	NDCLD15		Notre Dame		IITD-WVU		Clarkson	
		Original	Fine-tuned	Original	Fine-tuned	Original	Fine-tuned	Original	Fine-tuned
BPCR	uncompressed	0.00%	0.00%	0.00%	0.00%	11.41%	28.04%	0.40%	1.91%
	75	1.22%	0.37%	0.31%	0.00%	11.17%	28.04%	0.65%	2.86%
	50	1.29%	0.37%	0.23%	0.00%	10.43%	26.06%	2.23%	4.95%
	25	8.22%	0.37%	0.00%	0.00%	8.76%	26.60%	1.78%	2.72%
	10	35.41%	0.12%	0.00%	0.00%	8.22%	24.19%	2.31%	1.18%
APCER	uncompressed	0.00%	0.00%	6.42%	9.12%	13.39%	7.22%	4.02%	2.91%
	75	0.00%	0.00%	7.12%	11.11%	14.43%	8.80%	4.12%	3.64%
	50	0.00%	0.00%	7.74%	11.33%	15.21%	8.80%	3.38%	3.20%
	25	2.05%	0.00%	14.95%	14.15%	19.25%	8.86%	4.38%	3.90%
	10	28.51%	0.00%	39.58%	26.94%	27.73%	11.85%	4.81%	4.10%

Table 5.4: Iris PAD performance at different compression levels for Swin models trained on uncompressed images and fine-tuned with images compressed with JPEG @ QF=10.

Table 5.5: Iris PAD performance at different compression levels for Swin models trained on uncompressed images and fine-tuned with images compressed with JPEG AI @ TR=0.12.

Metric	TR	NDCLD15		Notre Dame		ITD-WVU		Clarkson	
		Original	Fine-tuned	Original	Fine-tuned	Original	Fine-tuned	Original	Fine-tuned
		uncompressed	uncompressed	uncompressed	uncompressed	uncompressed	uncompressed	uncompressed	uncompressed
BP CER	0.75	2.90%	0.62%	0.66%	0.70%	10.78%	30.74%	1.18%	10.90%
	0.50	2.90%	0.52%	0.66%	1.11%	10.53%	30.74%	1.46%	10.90%
	0.25	3.83%	0.52%	0.32%	1.94%	9.34%	31.41%	1.97%	9.17%
	0.12	13.51%	0.00%	0.12%	0.00%	6.07%	15.90%	7.56%	6.89%
AP CER	uncompressed	0.00%	0.00%	6.42%	3.24%	13.39%	5.67%	4.02%	1.20%
	0.75	0.00%	0.00%	7.33%	4.23%	14.41%	8.23%	4.08%	1.91%
	0.50	0.00%	0.00%	7.33%	4.38%	14.41%	8.15%	4.08%	1.91%
	0.25	0.00%	0.00%	9.12%	4.83%	17.34%	8.04%	3.75%	2.31%
	0.12	22.91%	0.00%	35.91%	23.70%	30.83%	18.10%	7.12%	0.59%

idence of this resilience. Under mild JPEG compression, the Transformer maintained meaningful attention patterns that focused on iris structures. Under severe compression, although the visual appearance degraded markedly, attention still gravitated toward regions containing traceable spoof cues, whereas CNNs lost consistency and became dependent on artifacts introduced by the codec.

5.3 Explainability

Explainability is a central component of this work, not only because it provides valuable insights into how Vision Transformers operate, but also because it strengthens the security of PAD systems. In biometric applications, transparency is essential: if a system rejects an identity claim or detects a spoof attempt, operators must have the ability to understand why. Explainability therefore acts as a bridge between raw performance metrics and operational trust.

5.3.1 Attention Maps as a Window Into Model Reasoning

Vision Transformers offer an intrinsic form of explainability through their attention mechanisms. By examining how the model distributes attention across iris regions, it becomes possible to understand the cues that influenced the final decision. In genuine samples, attention maps consistently concentrate on regions of the iris where the structural and photometric properties align with physiological norms. The radial flow of the texture, the contrast between the iris and the pupil, and the intricate patterns of the stroma become focal points in the model’s reasoning. In presentation attacks, the behavior changes substantially. Printed iris produce abrupt textures transitions that differ from biological structures; display attacks introduce periodic artifacts; and textured lenses distort the natural continuity of the iris. ViT reacts to these anomalies by disproportionately focusing on affected regions, effectively signaling the underlying irregularities.

5.3.2 Genuine vs Attack Patterns: What Attention Reveals

The contrast between attention patterns associated with bonafide samples and those generated by presentation attacks reveals several stable and recurring tendencies. Genuine iris images typically produce continuous and

anatomically plausible attention distributions, reflecting the smooth and coherent structure of real iris tissue. In contrast, printed attacks tend to elicit heightened attention around regions characterized by unnatural repetition or the presence of discrete edges, which arise from the printing process and are inconsistent with genuine iris texture. Replay attacks exhibit a different behavior, often shifting the model’s attention to areas affected by illumination inconsistencies, screen refresh artifacts, or reflections induced by display. These cues, while subtle, introduce irregularities that the attention mechanism learns to associate with spoofing. Similarly, textured contact lenses cause attention to cluster around ring boundaries or regions with atypical reflectance patterns, where the artificial overlay disrupts the natural appearance of the iris. Together, these observations highlight how attention-based analysis enables a fine-grained interpretation of model behavior across different attack types, reinforcing the value of explainability in iris Presentation Attack Detection. These tendencies reflect meaningful distinctions that align with domain-specific knowledge. Importantly, the explainability offered by ViTs validates that the model is not exploiting irrelevant background artifacts—as CNNs sometimes do—but is genuinely responding to features related to the structure of the iris.

5.3.3 Explainability Under Compression

Compression introduces distortions that may obscure the visual cues used for PAD. However, attention maps reveal that Transformers retain a considerable ability to identify spoof signatures even when image quality is substantially degraded. Under mild JPEG compression, attention remains focused on the iris region, albeit with slightly reduced sharpness, Fig. 5.5. Under severe JPEG or JPEG2000 compression, attention grows more diffuse but still highlights regions that contain remnants of spoof information, such as partial interference patterns or lens boundaries. Under JPEG-AI compression, attention maps remain remarkably stable. The codec preserves subtle structures essential for PAD, allowing the ViT to maintain a nearly unchanged attention distribution relative to the original images. This phenomenon illustrates how explainability can serve as a diagnostic tool: by analyzing attention responses under different compression pipelines, it becomes possible to predict how PAD performance will degrade and to identify preprocessing configurations that maximize robustness.

Consistently with these observations, models that were explicitly fine-tuned on compressed datasets achieve improved PAD performance under image quality degradation. The fine-tuning strategy was specifically adopted to investigate whether adapting the model to compression artifacts during

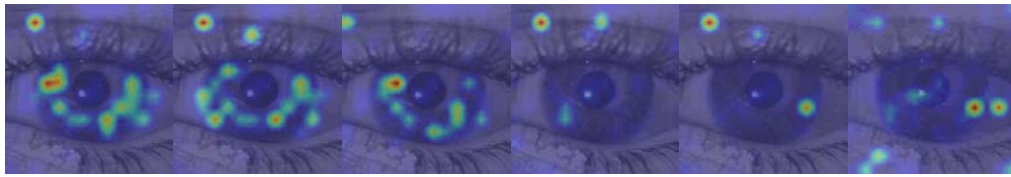


Figure 5.5: Attention maps for increasing JPEG compression on textured iris images, with decreasing quality from left to right.

training could mitigate the performance loss observed when operating on compressed images. The experimental results confirm that this approach is beneficial. When evaluated on compressed data, fine-tuned models consistently outperform the corresponding models, for both ViT Fig. 5.2 and 5.3 and Swin Tab. 5.4 and 5.5 model. trained solely on uncompressed images, achieving lower error rates across the considered compression levels. This indicates that fine-tuning enables the network to better accommodate the distortions introduced by lossy compression within the same experimental setting. The advantage of fine-tuning becomes more pronounced as the severity of compression increases. Although models trained only on original images exhibit a notable degradation in performance at lower quality factors, fine-tuned models show a more controlled reduction in accuracy, demonstrating increased robustness to compression-induced distortions within the evaluated datasets. It is worth noting that these improvements are observed without altering the model architecture or the evaluation protocol and can therefore be directly attributed to the inclusion of compressed samples during the fine-tuning phase. This confirms the effectiveness of compression-aware fine-tuning in improving PAD performance under degraded image quality, within the scope of the conducted experiments.

5.3.4 Explainability for Failure Analysis

One of the most powerful contributions of attention-based explainability lies in its ability to reveal why the model fails. In cases of misclassification—incorrect acceptance of an attack or false rejection of a genuine sample—attention maps show where the model focused and whether that focus was justified. False positives may arise when eyelid occlusions or sensor noise erroneously attract attention. False negatives, on the contrary, may occur when compression removes too much discriminative detail or when the iris is partially occluded. By visually inspecting these patterns, system designers can diagnose whether a specific preprocessing strategy, sensor configuration, or attack class introduces vulnerabilities. This enables informed decisions on how to adapt the PAD pipeline, such as modifying illumination,

adjusting segmentation, or even redesigning the acquisition workflow.

5.3.5 Toward Transparent and Trustworthy Iris PAD Systems

The integration of explainability into PAD represents a shift toward systems that are not only accurate but also accountable. In real-world deployments, transparency improves the credibility of the system, supports audit practices, and allows human operators to verify whether decisions are based on meaningful biometric evidence. Vision Transformers naturally support this level of transparency, offering a fine-grained and interpretable view of their internal decision process. Explainability therefore plays a dual role: it confirms that the Transformer relies on legitimate cues, and it exposes potential weaknesses that can be addressed to strengthen the security of the biometric system. For these reasons, the explainability component is not just a visual add-on but a fundamental aspect of designing practical, trustworthy, next-generation PAD solutions.



Chapter 6

Global Discussion, Conclusions, and Future Directions

The previous chapters of this thesis have presented a detailed experimental investigation of attention-based deep learning models applied to biometric recognition and security tasks. In particular, Chapter 4 focused on explainable wrist-vein biometric verification, while Chapter 5 addressed iris Presentation Attack Detection using Transformer-based architectures. Each of these chapters was designed as a self-contained case study, providing methodological details, experimental results, and modality-specific interpretations.

However, the contribution of this work extends beyond the individual analysis of these two biometric traits. The overarching objective of the thesis is not merely to demonstrate that Vision Transformers perform well in wrist-vein recognition or iris PAD, but to explore how attention mechanisms, explainability, and robustness interact across different biometric modalities and operational scenarios. For this reason, a dedicated discussion chapter is required to integrate the findings, identify common patterns, and extract higher-level insights that cannot be fully captured within modality-specific analyses.

This chapter serves as a unifying component of the thesis. Rather than introducing new experiments or additional performance evaluations, it provides a conceptual synthesis of the results presented earlier. By comparing wrist-vein verification and iris PAD, the chapter highlights how attention-based models behave under different biometric conditions, how explainability emerges as a consistent and informative property, and how robustness

to factors such as compression, session variability, and open-set scenarios manifests across tasks.

A key motivation for this global discussion is the observation that attention maps convey information that goes beyond qualitative visualization. Across both biometric modalities, the stability, localization, and semantic alignment of attention patterns are closely linked to the reliability of the model. When attention consistently focuses on physiologically meaningful regions, the corresponding models exhibit improved generalization and resilience to data degradation. In contrast, unstable or artifact-driven attention often anticipates performance degradation, even when standard evaluation metrics remain acceptable. This recurring relationship suggests that explainability can be interpreted as an indicator of robustness, rather than as a purely interpretative tool.

In addition, this chapter provides an appropriate context for assessing the practical relevance of the proposed approaches. Real-world biometric systems operate under constraints that differ markedly from controlled experimental conditions, including heterogeneous sensors, limited bandwidth, and evolving attack strategies. Therefore, the discussion emphasizes how the findings of this thesis translate into deployment considerations, system design choices, and security assessment practices.

Finally, the chapter outlines the limitations of the current work and identifies promising directions for future research. By explicitly acknowledging open challenges—such as the quantitative evaluation of explainability, computational efficiency, and multi-biometric fusion—the thesis positions itself as a foundation for continued investigation rather than a closed solution. In this sense, the present chapter bridges the detailed experimental contributions of the thesis with its broader scientific and practical implications.

6.1 Discussion

This section discusses the main observations emerging from the experimental results of Chapters 4 and 5. The goal is not to reiterate numerical performance metrics but to analyze how and why the proposed models behave as observed, with particular attention to common trends across biometric modalities.

6.1.1 Comparison across biometric modalities

Wrist-vein verification and iris Presentation Attack Detection represent two fundamentally different biometric problems. Wrist-vein recognition is a ver-

ification task based on subcutaneous vascular patterns, typically acquired under near-infrared illumination, and is affected by session variability, physiological changes, and acquisition noise. Iris PAD, on the other hand, is a security task aimed at distinguishing bona fide samples from spoofing attempts, where the discriminative cues are often subtle, heterogeneous, and strongly influenced by attack instruments and acquisition conditions.

Despite these differences, the experimental results reveal striking similarities in how Vision Transformers process the two modalities. In both cases, the models benefit from a global representation of the biometric signal, rather than relying solely on local texture descriptors. For wrist-vein images, attention tends to focus on vascular structures, junctions, and regions of high physiological relevance, while suppressing background areas and sensor-related artifacts. For iris images, attention focuses on regions where textural coherence, reflectance properties, and geometric consistency are most informative for distinguishing genuine samples from attacks.

This convergence suggests that attention mechanisms capture a higher-level notion of biometric consistency that is not tied to a specific modality. Instead of learning handcrafted, modality-specific features, the Transformer learns to evaluate whether the observed signal exhibits the structural and statistical properties expected of a genuine biometric trait. When these properties are violated—due to spoofing, noise, or degradation—the attention distribution changes accordingly, allowing both discrimination and interpretation.

6.1.2 Role of attention mechanisms

Across both biometric tasks, attention mechanisms play a central role in enabling robust and interpretable representations. Unlike convolutional networks, which progressively aggregate local features through fixed receptive fields, Vision Transformers establish direct relationships between distant regions of the input. This allows the model to reason about the global structure from the earliest layers.

In wrist-vein recognition, this capability is particularly relevant for modeling long-range vascular patterns that span the entire region of interest. In PAD iris, it enables the detection of inconsistencies that emerge only when multiple regions are jointly considered, such as unnatural periodicity, abnormal reflectance, or mismatches between local texture and global geometry. In both cases, attention provides a flexible mechanism for prioritizing informative regions while down-weighting irrelevant or misleading areas.

Another important aspect is that attention mechanisms reduce the model’s reliance on fragile local signals. This property becomes evident when con-

sidering scenarios that involve compression, noise, or domain shifts. By emphasizing relational information rather than isolated pixel-level details, Transformers are able to maintain stable performance under conditions that significantly degrade convolutional baselines.

6.1.3 Interpretation of attention patterns

The analysis of attention maps across experiments reveals consistent qualitative behaviors that align with domain knowledge. In both wrist-vein- and iris-based tasks, well-performing models exhibit attention patterns that are stable between samples, sessions, and acquisition conditions. These patterns tend to focus on regions that are physiologically or physically meaningful, providing an intuitive explanation of the model’s decisions.

In contrast, unstable or dispersed attention distributions often precede performance degradation. When attention shifts toward background regions, sensor artifacts, or compression-induced noise, the corresponding predictions become less reliable. This observation reinforces the idea that attention maps are not merely visualization tools but valuable indicators of model behavior. The consistency and semantic alignment of attention patterns provide insight into whether a model has learned meaningful biometric representations or is instead exploiting spurious correlations.

6.2 Explainability and Robustness

This section examines the relationship between explainability and robustness, arguing that the two concepts are deeply interconnected in the context of biometric systems.

6.2.1 Explainability beyond visualization

Explainability is often treated as an auxiliary component, added after model training to provide qualitative insights. However, in this thesis, explainability emerges as an intrinsic property of attention-based models. The attention maps produced by Vision Transformers are not post-hoc approximations, but direct by-products of the internal computation of the model.

This distinction is crucial. Because attention weights actively influence the model’s predictions, their analysis provides genuine insight into the decision-making process. Across experiments, models that produce semantically meaningful and stable attention maps are also those that achieve better generalization and robustness. This suggests that explainability can

be interpreted as a proxy for model reliability, rather than as a purely descriptive feature.

6.2.2 Effects of degradation and compression

The impact of image degradation and compression further highlights the connection between explainability and robustness. In both wrist-vein and iris experiments, performance degradation is accompanied by changes in attention distribution. Under mild degradation, attention remains focused on relevant biometric regions and performance is largely preserved. As degradation increases, attention becomes more diffuse or shifts toward artifact-prone areas, signaling a loss of discriminative information.

This behavior is particularly evident in the iris PAD experiments under different compression standards. While traditional compression algorithms significantly disrupt local textures, attention-based models are able to preserve meaningful relational cues, especially when modern learned compression techniques are used. The stability of attention maps under compression provides an intuitive explanation for the observed robustness of Vision Transformers compared to convolutional baselines.

6.2.3 Failure cases and diagnostic analysis

Attention maps also prove valuable for analyzing failure cases. False rejections or false acceptances can often be traced back to atypical attention patterns, such as excessive focus on occluded regions, background noise, or compression artifacts. By examining these patterns, it becomes possible to identify weaknesses in preprocessing pipelines, segmentation strategies, or training data composition.

From a practical point of view, this diagnostic capability is highly relevant. It enables developers and system operators to understand not only that a model fails but also why it fails and under which conditions. This insight can guide targeted improvements, such as refining data augmentation, adjusting acquisition protocols, or selecting more suitable compression schemes.

6.3 Practical Considerations

Beyond methodological contributions, the findings of this thesis have important implications for real-world biometric systems.

6.3.1 Implications for biometric system design

The demonstrated robustness and interpretability of attention-based models suggest several design guidelines for biometric systems. First, architectures that rely on global reasoning are better suited to handle variability in acquisition conditions, sensor characteristics, and data quality. Second, explainability should be considered a core requirement rather than an optional feature, especially in security-sensitive applications.

Attention analysis can inform system-level decisions, such as the choice of preprocessing techniques, region-of-interest extraction strategies, and compression standards. Observing how attention responds to different configurations, designers can select pipelines that preserve discriminative information and minimize vulnerability to degradation.

6.3.2 Operational and security aspects

In operational environments, biometric systems must be reliable, accountable and resilient to evolving threats. Attention-based explainability supports these requirements by providing transparent evidence of how decisions are made. This is particularly important in contexts where automated decisions may have legal or ethical implications.

In addition, the ability to interpret the behavior of the model improves the security assessment. Attention maps can reveal whether a system relies on legitimate biometric cues or artifacts that attackers could potentially exploit. In this sense, explainability contributes not only to transparency but also to proactive vulnerability analysis.

6.4 Limitations and Future Work

Although the results of this thesis demonstrate the potential of attention-based models, several limitations must be acknowledged.

6.4.1 Limitations of the proposed approach

Vision Transformers generally require higher computational resources than lightweight convolutional models, which may limit their deployment on resource-constrained devices. In addition, although attention maps provide valuable insight, their interpretation remains largely qualitative, introducing subjectivity into the analysis. The dependence on large-scale pretraining datasets also raises questions about bias transfer and domain mismatch.

6.4.2 Directions for future research

Future work should address these limitations by developing quantitative metrics for explainability, enabling objective evaluation of attention stability and semantic alignment. Extending attention-based frameworks to multi-modal biometric systems represents another promising direction, as attention mechanisms are naturally suited to dynamic fusion of heterogeneous signals. Finally, a deeper investigation of the interaction between explainability, robustness, and security—particularly under adversarial conditions—could further strengthen the practical relevance of these models.

6.5 Final Conclusions

This thesis has investigated the role of attention-based deep learning models in biometric recognition and security, with a specific focus on explainability, robustness, and generalization. Through two extensive experimental studies—one on wrist-vein biometric verification and one on iris Presentation Attack Detection—the work has demonstrated that Vision Transformers represent a powerful and versatile framework for addressing both recognition and anti-spoofing tasks under realistic operating conditions.

A central contribution of this thesis lies in showing that attention mechanisms provide benefits that extend beyond improved numerical performance. In both biometric modalities considered, attention-based models consistently learn representations that align with meaningful physiological or physical structures. In wrist-vein recognition, attention emphasizes vascular patterns and suppresses irrelevant background information, leading to improved verification stability between sessions and subjects. In PAD of the iris, attention highlights inconsistencies introduced by spoofing instruments, such as unnatural texture distributions, photometric irregularities, and structural discontinuities, enabling reliable detection of presentation attacks even in challenging scenarios.

An important outcome of this work is the identification of a strong relationship between explainability and robustness. The experimental analysis reveals that the models that produce stable and semantically coherent attention maps are also those that generalize more effectively across datasets, acquisition conditions, and image quality levels. Conversely, when attention becomes unstable or shifts toward non-biometric regions, performance degradation often follows. This observation supports the broader claim that explainability can be interpreted as an indicator of model reliability, rather than as a purely qualitative visualization tool. In this sense, attention maps offer a window into the internal reasoning of the model and provide action-

able insight into its strengths and weaknesses.

Investigating image compression further reinforces this perspective. By analyzing the behavior of attention-based models under different compression standards, this thesis demonstrates that Vision Transformers are inherently more resilient to data degradation than conventional convolutional architectures. Even when local micro-textures are partially destroyed by compression, global relational cues remain accessible to the self-attention mechanism, allowing the model to preserve discriminative power. The synergy observed between Transformers and modern learned compression techniques highlights the importance of considering the entire acquisition and preprocessing pipeline when designing biometric systems.

Beyond methodological contributions, this work also has significant implications for practical implementation. Biometric systems increasingly operate in unconstrained environments, where variability in sensors, bandwidth, and attack strategies is unavoidable. The combination of robustness and explainability offered by attention-based models addresses key requirements in such contexts, including transparency, accountability, and trustworthiness. By enabling operators to understand why a decision was made, attention-based explainability supports not only technical validation, but also operational accountability in security-sensitive applications.

At the same time, this thesis acknowledges the limitations of the proposed approaches. Attention-based models typically require greater computational resources, and the interpretation of attention maps remains largely qualitative. These limitations do not diminish the value of the contributions presented here; rather, they define a clear research agenda for future work. Advancing toward quantitative measures of explainability, exploring multi-biometric biometric fusion, and investigating the interaction between attention mechanisms and adversarial threats represent promising directions that naturally build on the foundations established in this thesis.

In conclusion, this work demonstrates that attention mechanisms constitute a fundamental and unifying approach to biometric recognition and security. By bridging performance, robustness, and explainability, Vision Transformers enable the development of biometric systems that are not only accurate but also transparent and reliable. The insights presented in this thesis contribute to a deeper understanding of how modern deep learning models can be responsibly applied in security-critical domains and provide a solid basis for future research in explainable and robust biometric systems.

References

- [1] Renu Sharma and Arun Ross, “D-NetPAD: An explainable and interpretable iris presentation attack detector,” in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020.
- [2] Zhaoyuan Fang, Adam Czajka, and Kevin W. Bowyer, “Robust iris presentation attack detection fusing 2d and 3d information,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 510–520, 2021.
- [3] Venkat Sai Swarup, Debanjan Sadhya, Vinal Patel, and Kanjar De, “Presentation attack detection in iris recognition through convolution block attention module,” in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, 2022.
- [4] Gaurav Jaswal, Raghavendra Ramachandra, and Christoph Busch, “Dfcanet: Deep feature correlation attention network for iris presentation attack detection,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 1, pp. 12–25, 2024.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv*, 2021.
- [6] Anil K Jain, Arun Ross, and Salil Prabhakar, “An introduction to biometric recognition,” *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4–20, 2004.
- [7] Arun Ross, Karthik Nandakumar, and Anil K. Jain, *Handbook of Multi-biometrics*, Springer, 2006.
- [8] Kevin W. Bowyer, Karen Hollingsworth, and Patrick J. Flynn, “A survey of iris biometrics research: 2008 update,” *Pattern Recognition*, vol. 41, no. 7, pp. 1149–1163, 2008.

- [9] N. Miura et al., “Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification,” *Machine Vision and Applications*, vol. 15, pp. 194–203, 2007.
- [10] Author Wang, “Title of the fingerprint paper,” *Journal Name*, vol. 1, pp. 1–10, 2006.
- [11] X. Liu et al., “A survey of deep learning for biometrics,” *Journal Name*, vol. XX, no. YY, pp. ZZZ–ZZZ, 2017.
- [12] Maria De Marsico et al., “Deep learning for biometrics: Advances and challenges,” *Journal Name*, vol. XX, no. YY, pp. ZZZ–ZZZ, 2020.
- [13] Anil K. Jain et al., “Biometric recognition: Recent advances and challenges,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. XX, no. YY, pp. ZZZ–ZZZ, 2021.
- [14] Finale Doshi-Velez and Been Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv*, 2017.
- [15] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 93, 2018.
- [16] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Andres Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al., “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [17] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [18] Matthew D Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” in *13th European Conference on Computer Vision*, 2014.
- [19] Andrea F. Abate, Michele Nappi, Danilo Riccio, and Giovanni Sabatino, “Performance evaluation of biometric systems,” *Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 3, pp. 537–554, 2007.

- [20] Robert Geirhos et al., “Shortcut learning in deep neural networks,” *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.
- [21] Andreas Holzinger et al., “Towards causability in explainable ai,” *Frontiers in Artificial Intelligence*, vol. 5, pp. 1–10, 2022.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [23] Hila Chefer, Shir Gur, and Lior Wolf, “Transformer interpretability beyond attention visualization,” 2021.
- [24] Author Jain, “Attention in neural networks,” *Journal of Attention*, vol. 1, pp. 1–10, 2019.
- [25] Sarah Wiegrefe and Yuval Pinter, “Attention is not not explanation,” *arXiv preprint arXiv:1908.04626*, 2019.
- [26] Adam Czajka and Kevin W. Bowyer, “Biometric explainability: A survey,” *IEEE Transactions on Technology and Society*, vol. 2, no. 2, pp. 76–88, 2021.
- [27] Pedro C Neto, Tiago Gonçalves, João Ribeiro Pinto, Wilson Silva, Ana F Sequeira, Arun Ross, and Jaime S Cardoso, “Explainable biometrics in the age of deep learning,” *arXiv*, 2022.
- [28] Anil K. Jain, “Trust in biometric recognition: Challenges and opportunities,” *Journal of Biometrics*, vol. XX, no. YY, pp. ZZZ–ZZZ, 2022.
- [29] Adam Czajka and Kevin W Bowyer, “Performance of iris presentation attack detection techniques on post-mortem samples,” in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–9.
- [30] Raghavendra Ramachandra and Christoph Busch, “Presentation attack detection methods for face recognition systems: A comprehensive survey,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.
- [31] Kiran B. Raja, Ramachandra Raghavendra, Vamsi Kiran Vemuri, and Christoph Busch, “Iris presentation attack detection: Generalization and efficiency,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2238–2252, 2016.

- [32] N. Poh et al., “Variability in biometric systems,” *Journal Name*, vol. XX, pp. XX–XX, 2019.
- [33] John Daugman, “How iris recognition works,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 21–30, 2004.
- [34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.
- [35] Daniele Albano et al., “Explainable vision transformers for medical image analysis,” *Journal of Medical Imaging*, vol. 10, no. 1, pp. 1–15, 2023.
- [36] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, “A survey of methods for explaining black box models,” *ACM Comput. Surv.*, vol. 51, no. 5, Aug. 2018.
- [37] Christian Rathgeb and Andreas Uhl, “A survey on biometric cryptosystems and cancelable biometrics,” *EURASIP Journal on Information Security*, vol. 2011, no. 3, pp. 1–25, 2011.
- [38] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [39] Karthik Nandakumar and Anil K. Jain, “Template protection: Theory and practice in biometric systems,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2113–2126, 2015.
- [40] Adam Czajka and Kevin W. Bowyer, “Presentation attack detection for iris recognition: An assessment of the state-of-the-art,” *ACM Comput. Surv.*, vol. 51, no. 4, 2018.
- [41] P. Drozdowski, C. Rathgeb, A. Dantcheva, C. Busch, and M. Gomez-Barrero, “Demographic bias in biometrics: A survey on an emerging challenge,” *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

- [42] Yilong Yin, Lili Liu, and Xiwei Sun, “Sdumla-hmt: A multimodal biometric database,” in *Biometric Recognition: 6th Chinese Conference, CCBR 2011, Beijing, China, December 3-4, 2011. Proceedings 6*. Springer, 2011, pp. 260–268.
- [43] Christian Rathgeb and Christoph Busch, “Compression of biometric data: Iris, fingerprint and face,” *IET Biometrics*, vol. 3, no. 4, pp. 222–233, 2014.
- [44] H. Hofbauer and A. Uhl, “Compression of biometric data: Iris, fingerprint and face,” in *2016 International Conference on Biometrics (ICB)*, 2016, pp. 1–8.
- [45] Patrick J. Flynn et al., “The impact of jpeg compression on biometric systems,” in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2008, pp. 1–6.
- [46] Adam Czajka, Zhaoyuan Fang, and Kevin Bowyer, “Iris presentation attack detection based on photometric stereo features,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.
- [47] Kien Nguyen, Hugo Proença, and Fernando Alonso-Fernandez, “Deep learning for iris recognition: A survey,” *ACM Comput. Surv.*, vol. 56, no. 9, 2024.
- [48] Cunjian Chen and Arun Ross, “An explainable attention-guided iris presentation attack detector,” in *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2021.
- [49] Rocco Albano, Lorenzo Giusti, Emanuele Maiorana, and Patrizio Campisi, “Explainable vision transformers for vein biometric recognition,” *IEEE Access*, vol. 12, pp. 60436–60446, 2024.
- [50] Anjith George and Sébastien Marcel, “On the effectiveness of vision transformers for zero-shot face anti-spoofing,” in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.
- [51] R.W. Ives, B.L. Bonney, and D.M. Etter, “Effect of image compression on iris recognition,” in *2005 IEEE Instrumentation and Measurement Technology Conference Proceedings*, 2005, vol. 3, pp. 2054–2058.
- [52] Robert Ives, Daniel Bishop, Yingzi Du, and Belcher Craig, “Iris recognition: The consequences of image compression,” *EURASIP Journal on Advances in Signal Processing*, vol. 2010, 02 2010.

- [53] John Daugman and Cathryn Downing, “Effect of severe image compression on iris recognition performance,” *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 52–61, 2008.
- [54] Soumyadip Rakshit and Donald M. Monro, “An evaluation of image sampling and compression for human iris recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 605–612, 2007.
- [55] I.J. Shaikh and S.M. Mukane, “Iris image compression using JPEG & its effect on recognition performance,” *International Journal of Scientific & Engineering Research*, vol. 6, no. 2, 2015.
- [56] Heinz Hofbauer, Christian Rathgeb, Johannes Wagner, Andreas Uhl, and Christoph Busch, “Investigation of better portable graphics compression for iris biometric recognition,” in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2015.
- [57] Gio Arturo S. Martinez, Reese Elijah B. Moralde, Noel B. Linsangan, and Rayellee Myrtle Laire Ang, “A comparative analysis between the performance of the extracted features of JPEG and PNG on a Raspberry Pi iris recognition system,” in *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*, 2023, pp. 811–816.
- [58] R. Kabacinski and K. Kowalski, “Vein pattern database and benchmark results,” *Electronic Letters*, vol. 47, no. 20, pp. 1127–1128, 2011.
- [59] Önsen Toygar, Felix O. Babalola, and Yıltan Bitirim, “Fyo: A novel multimodal vein database with palmar, dorsal and wrist biometrics,” *IEEE Access*, vol. 8, pp. 82461–82470, 2020.
- [60] Ibrahim Majid Mohammed and Nor Ashidi Mat Isa, “Contrast limited adaptive local histogram equalization method for poor contrast image enhancement,” *IEEE Access*, vol. 13, pp. 62600–62632, 2025.
- [61] James S. Doyle and Kevin W. Bowyer, “Robust detection of textured contact lenses in iris recognition using BSIF,” *IEEE Access*, vol. 3, pp. 1672–1683, 2015.
- [62] David Yambay et al., “Livdet iris 2017 — iris liveness detection competition 2017,” in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017.

- [63] Aidan Boyd, Zhaoyuan Fang, Adam Czajka, and Kevin W. Bowyer, “Iris presentation attack detection: Where are we now?,” *Pattern Recognition Letters*, vol. 138, pp. 483–489, 2020.
- [64] Rocco Albano, Filippo Battaglia, Emanuele Maiorana, Giuseppe Campobello, and Patrizio Campisi, “Effects of compression on attention-based iris presentation attack detection,” in *2025 33rd European Signal Processing Conference (EUSIPCO)*, 2025, pp. 1–5.
- [65] Alexey Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv*, 2021.
- [66] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002.
- [67] International Organization for Standardization, “Information technology – biometric presentation attack detection – part 1: Framework,” <https://www.iso.org/standard/53227.html>, 2016, ISO/IEC 30107-1:2016.
- [68] Gregory K Wallace, “The JPEG still picture compression standard,” *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.
- [69] JPEG AI, “Reference software for Rec. ITU-T T.840.1 — ISO/IEC 6048-1,” 2024–2025, <https://gitlab.com/wg1/jpeg-ai/jpeg-ai-reference-software>, Last accessed on 2025-06-23.
- [70] Andreas Uhl, Christoph Busch, Sébastien Marcel, and Raymond Veldhuis, *Handbook of Vascular Biometrics*, Springer, 2020.
- [71] Naoto Miura, Akio Nagasaka, and Takafumi Miyatake, “Extraction of finger-vein patterns using maximum curvature points in image profiles,” *IEICE Transactions on Information and Systems*, vol. E90-D, no. 8, pp. 1185–1194, 2007.
- [72] Hoang Thien Van, Chien Minh Duong, Giang Van Vu, and Thai Hoang Le, “Palm vein recognition using enhanced symmetry local binary pattern and sift features,” in *19th International Symposium on Communications and Information Technologies (ISCIT)*, 2019.
- [73] Kalaivani Sundararajan and Damon L Woodard, “Deep learning for biometrics: A survey,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–34, 2018.

- [74] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [75] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei, “Layercam: Exploring hierarchical class activation maps for localization,” *Trans. Img. Proc.*, vol. 30, pp. 5875–5888, Jan. 2021.
- [76] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, “Why should i trust you? explaining the predictions of any classifier,” *arXiv*, 2016.
- [77] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [78] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *IEEE conference on computer vision and pattern recognition*, 2016.
- [79] Manar Ramzy Dronky, Wael Khalifa, and Mohamed Roushdy, “Impact of segmentation on iris liveness detection,” in *2019 14th International Conference on Computer Engineering and Systems (ICCES)*, 2019.
- [80] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, “Demographic bias in presentation attack detection of iris recognition systems,” in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021.
- [81] Juan Tapia and Claudia Arellano, “Gender classification from iris texture images using a new set of binary statistical image features,” in *2019 International Conference on Biometrics (ICB)*, 2019.
- [82] Yung-Yao Chen, Sin-Ye Jhong, Chih-Hsien Hsia, and Kai-Lung Hua, “Explainable AI: a multispectral palm-vein identification system with new augmentation features,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 3, pp. 1–21, 2021.
- [83] Anil Sharma, Ajay Singh, Mayank Vatsa, and Richa Singh, “Dnetpad: Cnn-based presentation attack detection for iris recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1494–1506, 2020.

-
- [84] Author Fang and Others, “Title of the paper,” *Journal Name*, vol. XX, no. YY, pp. ZZZ–ZZZ, 2021.
 - [85] John Daugman, “The impact of compression on iris recognition performance,” *Journal Name*, vol. xx, pp. xx–xx, 2008, Please update with correct details.
 - [86] Christian Rathgeb, Andreas Uhl, and Peter Wild, “Effects of severe image compression on iris segmentation performance,” in *IEEE International Joint Conference on Biometrics*, 2014.