Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Original Article

Cross-Architecture Evaluation of Deep Learning Models for Multimodal Biometric Verification with Score-Level Fusion

Rowida Jasim Alazawi [a]*, **Nada Jasim Habeeb** [a], **Alaa Jabbar Qasim Almaliki** [b]

[a] Technical College of Management, Middle Technical University, Baghdad, Iraq.
[b] School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah Darul Aman, Malaysia.

Submitted: 24 April 2026
Revised: 18 July 2026
Accepted: 20 July 2026

*** Corresponding Author**:
dac5008@mtu.edu.iq

Keywords: Biometric Verification, Deep Learning, CNN, Transformer, Multimodal Fusion, Score-Level Fusion.

How to cite this paper: R. J. Alazawi, N. J. Habeeb, A. J. Q. Almaliki, "Cross-Architecture Evaluation of Deep Learning Models for Multimodal Biometric Verification with Score-Level Fusion," KJAR, vol. 11, no. 2, pp. 33-50, December 2026, doi: 10.24017/science.2026.2.2

Abstract: Comparing the deep learning architectures used for biometric verification in a fair manner has remained a difficult task. Researchers typically rely on different datasets, employ varying preprocessing steps, and adopt inconsistent evaluation criteria, all of which muddy the picture when trying to decide whether one network truly outperforms another. This work attempts to cut through that noise by fixing every variable except the backbone architecture itself. Residual Network-50 (ResNet50), Efficient Network Version 2 Small (EfficientNetV2-S), and the Swin Transformer-Tiny (Swin-T) were trained and tested on the same LUTBIO and XJTU datasets, applying identical subject-disjoint splits across face, fingerprint, and palmprint modalities. The analysis measures Equal Error Rate (EER), Area Under the Curve, and True Acceptance Rate at False Acceptance Rate = 0.1%, and further examines five distinct score-level fusion strategies across every possible two- and three-modality pairing. The single-modality experiments reveal a clear pattern: Swin-T achieves a perfect 0.000% EER on faces, ResNet50 handles fingerprints best at 1.957% EER, and EfficientNetV2-S leads on palmprints with 0.140% EER. On LUTBIO, 10 out of 12 tested fusion configurations reach a reported EER of 0.000%. This study reports the effect sizes and selected confidence intervals for the key comparisons, allowing both their statistical and practical significance to be assessed. The results show that no single architecture is universally superior; the best-performing backbone is modality-dependent. Score-level fusion consistently improved verification accuracy over the best single modality, substantially reducing error rates. These findings imply that architecture choices for biometric systems should be modality-aware and that score-level fusion is an appropriate method for robust multimodal verification. The controlled protocol additionally provides a reproducible framework for fair architectural comparison in future biometric research.

1. Introduction

Biometric verification has become a routine part of everyday life, used in unlocking smartphones, automated border control, and workplace access. Much of this progress has been driven by deep learning, which has largely replaced hand-crafted feature extractors with learned representations that are, in many cases, considerably more discriminative [1]. Despite these advances, a persistent challenge in the literature is the lack of fair and controlled comparisons among deep learning architectures. Existing studies frequently adopt different datasets, subject partitions, preprocessing procedures, data augmentation strategies, training protocols, and evaluation metrics, making the direct comparison of their reported performance difficult. Consequently, observed performance differences may reflect variations in the experimental design rather than the intrinsic capabilities of the evaluated architectures. As emphasized in recent comparative studies, rigorous evaluations conducted under consistent experimental

conditions are essential for drawing reliable conclusions regarding the relative strengths of the different deep learning models [2].

This difficulty becomes more pronounced when convolutional neural networks (CNNs) and vision transformers (ViT) are compared. CNNs encode strong assumptions about the local spatial structure: nearby pixels are treated as related, and features are largely translation-invariant. ViTs relax these assumptions in favor of self-attention, which is more expressive but typically more data-hungry [3]. The outcome of this trade-off depends strongly on the dataset size, the domain, and the preprocessing pipeline; for instance, a ViT trained only on a moderate-scale dataset may underperform a comparable CNN, whereas large-scale pre-training can reverse this ranking [3, 4]. Most published benchmarks that evaluate a single modality frequently face and declare a winner under conditions that are not directly comparable across studies [2].

A second concern is the strong emphasis on single-modality evaluation. Different biometric traits differ substantially in both structure and acquisition variability. Fingerprints are dominated by local ridge patterns and exhibit high intra-class variation, since increased pressure or dry skin alters the captured image [5]. Palmprints occupy an intermediate position, offering strong principal lines but relatively few samples per subject in practice [6]. To compare architectures fairly across such heterogeneous modalities, each architecture must be evaluated on every modality under identical conditions, and the modalities must then be combined through fusion [7]. The three biometric modalities used in this study - face, fingerprint, and palmprint — are illustrated in figure 1, using samples from the LUTBIO dataset.

Figure 1: Sample biometric modalities from the LUTBIO database showing (a) face, (b) fingerprint, and (c) palmprint images.

Motivated by these observations, the present study is designed around a single guiding principle: to hold every factor constant except for the variable under study, the backbone architecture. Three architectures — Residual Network-50 (ResNet50), Efficient Network Version 2 Small (EfficientNetV2-S), and the Swin Transformer — are trained on the LUTBIO dataset [8] using identical subject-disjoint splits, augmentation, loss functions, training schedule, and hardware. Each architecture is evaluated separately on every modality and then combined using five score-level fusion rules. Performance is reported using the Equal Error Rate (EER), Area Under the Curve (AUC), and True Acceptance Rate at a False Acceptance Rate of 0.1% (TAR@0.1%), together with training time, Cohen's d effect sizes, and bootstrap confidence intervals. Where applicable, the evaluation protocol follows the International Organization for Standardization / International Electrotechnical Commission (ISO/IEC) 19795-1:2021 [9].

The contributions of this study are as follows:

- First, it provides a fully reproducible, confound-free evaluation framework that enables a fair architectural comparison across three structurally different biometric modalities.
- Second, it demonstrates that no single backbone is universally superior across modalities; the best-performing architecture is modality-dependent.
- Third, it shows that simple score-level fusion can substantially reduce error rates across most configurations.
- Fourth, it reports effect sizes and confidence intervals to support the main claims, in addition to conventional significance testing.

- Finally, to test the generalizability of the findings, the evaluation is extended to a second, publicly available dataset, the MultiModal-database-XJTU [10]. As an older dataset collected under different environmental conditions, XJTU presents a more challenging, less idealized scenario with variations in lighting, sensor quality, and image resolution, allowing the stability of the architecture rankings to be assessed beyond the clean, controlled samples of LUTBIO.

The rest of the paper is structured as follows. Section 2 provides an overview of the related work and highlights the research gaps that led to this study. Section 3 elaborates on the methodology used, such as the datasets, preprocessing steps, model architecture, training method, and score level fusion. Section 4 provides the experimental results and analysis, whereas section 5 is devoted to the discussion and subsequent implications. Lastly, section 6 concludes the paper and offers future research directions.

2. Related Works

This section reviews the prior work relevant to the present study along three complementary axes. Subsection 2.1 surveys the deep learning architectures most widely used for biometric embedding, covering both convolutional and transformer-based backbones. Subsection 2.2 then examines why the inductive biases of CNNs and ViTs matter specifically for biometric recognition, where the training data is typically limited. Subsection 2.3 turns to score-level fusion as a means of combining evidence across modalities. The section concludes by synthesizing the recurring limitations of these studies and identifying the research gaps that motivate the controlled, multimodal evaluation proposed in this work.

2.1 Deep Learning Architectures for Biometric Embedding

Fundamentally, biometric verification is a metric-learning task: two images are encoded into two embedding vectors, and the system decides whether they are sufficiently similar in a distance space, usually under cosine similarity [11]. The quality of these embeddings depends almost entirely on the backbone architecture that produces them, which is why the three architectures studied here originate from distinct design philosophies. He *et al.* [12] introduced residual skip connections, which alleviate the vanishing-gradient problem and allow networks of 50 or 152 layers to be trained reliably; their ResNet design reported a top-5 error of 3.57% on ImageNet and remains a standard CNN baseline. Tan and Le [13] subsequently proposed EfficientNetV2, which uses a neural architecture search together with progressive learning and a fused-MBConv design to jointly balance depth, width, and resolution, achieving higher accuracy and substantially faster training than earlier CNNs at comparable parameter counts. Departing from convolution entirely, Liu *et al.* [14] proposed the Swin Transformer, which partitions the input into local windows, applies self-attention within each window, and shifts the windows in alternating layers to enable cross-window communication, yielding a hierarchical transformer that scales linearly with image size. It reached State-of-the-Art (SOTA) results in classification, detection, and segmentation.

Beyond these backbones, several studies have adapted deep architectures to specific biometric traits, and we review them here in chronological order to trace the evolution of the field. Peri *et al.* [15] proposed a synthesis-based framework for thermal-to-visible face verification using Contrastive Unpaired Translation (CUT) combined with face alignment and identity-preserving learning. By synthesizing visible-spectrum faces from thermal inputs before matching, the method achieved SOTA results using both the Army Research Laboratory Visible–Thermal Face (ARL-VTF) and the Tufts University Thermal Face (TUFTS) databases. While it demonstrated the feasibility of cross-spectrum verification under low-light conditions, its reliance on accurately aligned paired thermal–visible images limit its applicability in unconstrained settings. He *et al.* [16] addressed partial-fingerprint verification with the partial fingerprint verification network (PFVNet), combining a spatial transformer network (STN) with local self-attention to align and match incomplete ridge regions. Trained on large-scale fingerprint matching data, it reported an EER in the 3–6% range on partial prints, although performance remained tied to the partial-print scenario. Brown and Bradshaw [17] developed a palmprint recognition system based on palmprint alignment and a customized CNN derived from the Visual Geometry Group-16 (VGG-16) architecture, supported by extensive data augmentation. Benchmarked against VGG-16 and Xception baselines, it achieved near-zero EER for the IITD-Palmprint database, but remained sensitive

to errors in palm segmentation. Su *et al.* [18] designed a hybrid CNN-transformer for face recognition that fuses atomic and holistic tokens to capture both fine-grained and contextual facial cues, reporting roughly a 2.7% gain over the prior SOTA, although the evaluation was confined to the face modality.

In the same year, three hand-based contributions appeared: Hattab and Behloul [19] built a face–iris multimodal system using CNN backbones with feature-level fusion, achieving high accuracy but across only two modalities; Yang *et al.* [20] proposed joint multi-type feature learning for multi-modality finger-knuckle-print (FKP) recognition, reaching SOTA FKP results while being restricted to a single trait family; and Su *et al.* [21] later extended this line with a complete region-of-interest (ROI) formulation for unconstrained palmprint recognition. Garg *et al.* [5] applied convolutional networks with image inversion and multi-augmentation to compensate for limited fingerprint training data, attaining high accuracy under data-scarce conditions. Wu and Hu [22] fused deep ResNet features with shallow handcrafted descriptors at both feature and score level for palmprint recognition, confirming that complementary representations outperform either alone, albeit for a single modality. Also in 2024, Lin *et al.* [23] introduced a lightweight network for robust palmprint ROI extraction under unconstrained acquisition, Fan *et al.* [24] proposed a hybrid palmprint–palm-vein fusion for large-scale recognition, and Grosz *et al.* [25] presented Palm-ID, a contactless mobile palmprint framework that couples a CNN with a ViT alongside palmprint enhancement and dimensionality reduction, reporting a TAR of 98.06% at a False Acceptance Rate (FAR) of 0.01% in cross-database experiments. However, it was validated only on the contactless palmprint modality. Most recently, Hossain *et al.* [26] evaluated the longitudinal performance of four deep face recognition models (FaceNet, ArcFace, CosFace, and MagFace) on the Infants and Toddlers Longitudinal Face dataset, showing that verification accuracy declines as the age gap between enrollment and probe images widens. MagFace achieved the best overall performance, but the study was limited to a small face-only dataset and did not address multimodal recognition. Collectively, these works demonstrate strong per-trait performance yet each is typically validated using only a single modality and under its own protocol, leaving cross-modality architectural behavior largely unexamined.

2.2 Relevance of CNN and ViT Architectures for Biometric Recognition

CNNs carry a strong structural prior: convolution kernels impose local connectivity and weight sharing, which act as a built-in regularizer [27]. This is advantageous when training data is scarce, because the network cannot easily overfit to long-range spurious correlations that it does not "see" until its deeper layers. ViTs relax this prior, allowing every token to attend to every other token from the first layer. This flexibility is not free: without sufficient data, attention maps become noisy and the model tends to memorise rather than generalize. Han *et al.* [3] demonstrated this trade-off directly, showing that a ViT trained only on ImageNet-1K underperformed in relation to a comparable ResNet, whereas pre-training on the much larger JFT-300M dataset reversed the ranking. Zhai *et al*. [28] characterized the corresponding scaling regularizer and found that transformer accuracy continues to improve more steeply than that of CNNs as the dataset and model size grows.

These scaling dynamics have direct consequences for biometrics. Controlled laboratory face datasets typically contain only a few hundred subjects, which sits well below the data regime in which transformers are known to dominate. Whether this scale is enough for a transformer to surpass a CNN, or whether the convolutional inductive bias prevails, is unlikely to have a single answer and is expected to depend on the modality. This is an expectation this study tests empirically across face, fingerprint, and palmprint under identical conditions.

2.3 Score-Level Fusion in Multimodal Systems

Combining evidence from multiple biometric traits is a long-established idea. Ross and Jain [29] demonstrated its effectiveness over two decades ago yet it remains underused in architecture comparison studies. They described score-level fusion as the simplest variant, in which each modality produces a similarity score, and a rule (average, weighted, maximum, minimum, or product) merges these into a single decision value, requiring no retraining, keeping each modality pipeline independent, and adding a minimal implementation overhead. In the present study, score-level fusion was adopted precisely

because it preserves experimental clarity: the fused result depends only on the quality of the individual embeddings, not on a separate fusion network that could introduce its own confounds.

Several deep-learning studies illustrate the practical value of multimodal fusion. Feng and Kumar [30] built discriminative templates for contactless palmprint recognition, reporting SOTA cross-database performance, though only for the contactless palmprint setting. Khatri and Sharma [31] fused iris, face, and palmprint features with ResNet-based backbones and reported near-perfect accuracy but used separate datasets for the different traits rather than matched subjects. El-Rahman and Alluhaidan [32] combined deep learning with traditional methods to fuse electrocardiogram (ECG) and fingerprint signals, achieving AUC = 0.99, albeit on a single dataset. At the survey level, Li *et al.* [33] reviewed hand-based multimodal fusion strategies and singled out score-level fusion as a robust, low-overhead option. Table 1 summarizes the comparative studies reviewed in this section by presenting the biometric modalities, evaluated models, key performance, and principal limitations of each study.

Table 1: Summary of the prior studies.

Ref	Year	Modality	Models	Methodology (brief)	Key Metric	Limitation
[5]	2024	Fingerprint	CNN (inversion + augmentation)	Image inversion + multi-augmentation for scarce data	High accuracy	Limited training data
[15]	2021	Face (thermal–visible)	Synthesis-based Deep Convolutional Neural Network (DCNN) using (CUT)	Thermal-to-visible image synthesis (CUT) + face alignment + identity classification	SOTA on ARL-VTF / TUFTS	Cross-spectral; alignment dependent
[16]	2022	Fingerprint	PFVNet (STN + attention)	Spatial transformer + local self-attention for partial prints	EER ≈ 3–6%	Partial prints only
[17]	2022	Palmprint	CNN (VGG-16)	Palmprint alignment + customized VGG-16 + heavy augmentation	EER = 0–0.5%	ROI/segmentation dependent
[18]	2023	Face	Hybrid Token Transformer	Fuses atomic + holistic tokens (CNN–transformer)	+2.7% over SOTA	Face-only
[19]	2023	Face + Iris	CNN + feature fusion	Deep feature-level fusion of two traits	High accuracy	Two modalities only
[20]	2023	Finger-knuckle	Joint multi-type feature learning	Multi-type feature learning for FKP	SOTA (FKP)	Single trait family
[21]	2024	Palmprint	Complete-ROI network	Complete-ROI formulation, unconstrained	SOTA (unconstrained)	Single modality
[22]	2024	Palmprint	Deep + shallow feature/score fusion	Fuses ResNet + hand-crafted features	SOTA (palmprint)	Single modality
[23]	2024	Palmprint	Lightweight ROI network	Lightweight unconstrained ROI extraction	Robust ROI	ROI extraction only
[24]	2024	Palmprint + Palm vein	Hybrid fusion	Large-scale palm-trait hybrid fusion	SOTA (large-scale)	Two hand traits only
[25]	2024	Palmprint	Palm-ID (CNN + ViT)	Contactless mobile, enhancement + dimensionality reduction	TAR = 98.06% @ FAR 0.01%	Single modality
[26]	2025	Face (infants/toddlers)	FaceNet, ArcFace, CosFace, MagFace	Longitudinal evaluation across developmental stages	TAR ≈ 64.7% @ FAR=0.1%	Small dataset; age drift
[30]	2023	Palmprint	BEST (template fusion)	Builds evidence from scattered templates (contactless)	SOTA (cross-database)	Contactless only
[31]	2023	Multimodal	ResNet50/101 + Fusion	Iris+face+palmprint feature fusion	Accuracy = 100%	Separate datasets per trait
[32]	2024	Multimodal	CNN (ECG + fingerprint)	Deep + traditional fusion of ECG and fingerprint	AUC = 0.99	Single dataset
[33]	2024	Hand-based multimodal	Survey	Review of sensor-, feature-, score-, rank-, and decision-level fusion	N/A	No experimental validation

Taken together, the studies reviewed above reveal three consistent patterns. First, architectural comparisons in biometric verification are rarely conducted under controlled conditions. The studies differ in terms of the datasets, preprocessing, and evaluation metrics used, which makes it difficult to attribute reported performance gaps to the architectures rather than to the experimental setup. Second, the great majority of works evaluate a single modality, most often the face, so the question of whether a CNN or ViT is preferable is answered only partially and cannot be generalized across structurally different traits such as fingerprint and palmprint. Third, although score-level fusion is repeatedly shown to be effective and low in overhead, it is seldom integrated into architecture-comparison studies, and is more often used to maximize accuracy than to test the stability of architectural rankings.

These observations expose two unresolved gaps. The first is the absence of a confound-free protocol that evaluates the same architectures with multiple modalities under identical conditions. The second is the scarcity of publicly available multimodal datasets that provide matched face, fingerprint, and palmprint samples from the same subjects. Because a biometric trait is irreversible, where a compromised password can be reset but a leaked fingerprint or palmprint template cannot be reissued, strict privacy regulations limit the public availability of such datasets and explain why most studies rely on a small number of subjects.

The present study is designed to address these gaps directly. By holding the dataset, preprocessing, splits, training schedule, and hardware constant, and varying only the backbone architecture, it enables a fair cross-architecture comparison across face, fingerprint, and palmprint. Score-level fusion is then applied not only to improve accuracy but to assess whether the architectural rankings observed for single modalities remain stable once the modalities are combined. The evaluation is extended to a second, more challenging, dataset to test the generalizability of the findings.

3. Materials and Methods

The proposed system follows a unified verification pipeline applied identically to all three architectures and all three biometric modalities, so the backbone is the only variable under study. As summarized in figure 2, each input sample (face, fingerprint, or palmprint) is first passed through a modality-specific preprocessing stage and then fed to one of the evaluated backbone networks (ResNet50, EfficientNetV2-S, or Swin-T), which extracts the deep feature representation. A modality-specific projection head maps these features into a normalized embedding space, and the verification scores are obtained by computing the cosine similarity between embedding pairs. Finally, the scores produced for the individual modalities are combined through score-level fusion to yield the fused decision score used in the comparative analysis. The individual stages of this pipeline (dataset selection, data partitioning and preprocessing, model architectures and training, and score-level fusion) are described in detail in the following subsections, which are organized to be consistent with the steps shown in figure 2.

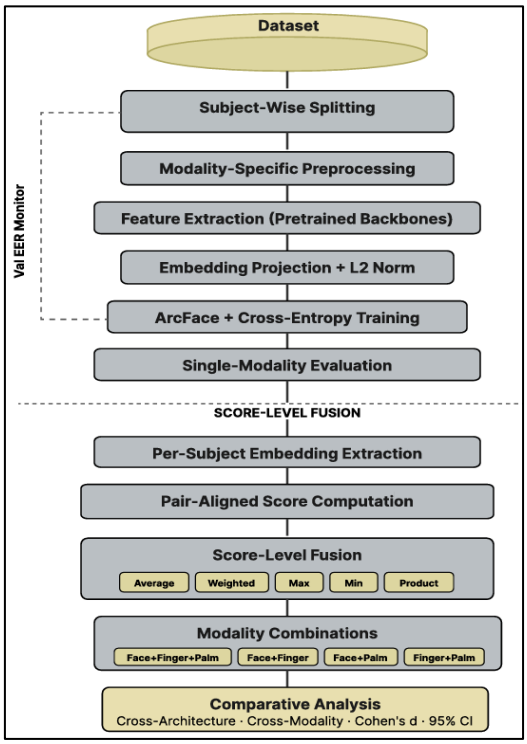

Figure 2: Proposed methodology for the cross-architecture multimodal biometric verification with score-level fusion.

3.1. Biometric Dataset Selection and Data Acquisition Challenges

The gathering of biometric information poses serious issues of privacy and security mainly because biometrics are immutable; once stolen, there is no way to change, revoke, or unlink the information with that of the individual [34]. The storage of this data should be in compliance with frameworks such as the General Data Protection Regulation of the European Union and ISO/IEC 24745:2022 [35] on biometric template protection. Redistribution is regulated by data usage agreements that limit the accessibility of the samples according to purpose.

The practical upshot is that publicly available multimodal biometric databases—ones that provide matched face, fingerprint, and palmprint samples from the very same subjects - are remarkably rare. Most datasets cover a single trait, and even those tend to be small by modern deep learning standards. LUTBIO is one of the few exceptions, and that scarcity is precisely why it was selected despite its modest scale. It offers matched multi-trait recordings under controlled conditions, which is non-negotiable for fair cross-modality fusion experiments.

The LUTBIO multimodal biometric database [8] includes face, fingerprint, and palmprint data from 306 subjects collected in a laboratory setting. The three modalities differ considerably in sample count: fingerprint has the most images at 3,880, followed by face at 1,836 and palmprint at just 918. Table 2 provides the breakdown.

Table 2: Characteristics of the LUTBIO multimodal biometric dataset.

Modality	Total Images	Subjects	Samples/Subject	Image Resolution	Capture Method
Face	1,836	306	~6	504×672 px	Controlled lighting
Fingerprint	3,880	306	~12–13	300×400 px	Optical sensor
Palmprint	918	306	~3	672×504 px	Contactless

In addition to LUTBIO, a second, older dataset was incorporated to examine the robustness of the architectural comparisons: the MultiModal-database-XJTU. This dataset serves as a contrasting counterpart to LUTBIO. While LUTBIO provides clean, controlled, and relatively modern samples, XJTU represents an earlier generation of biometric data collection. It is characterized by lower image resolutions, inconsistent lighting conditions, and the use of older sensors, which introduce higher intra-class variation and make the verification task more challenging. The detailed characteristics of the XJTU multimodal biometric dataset used in this study are summarized in table 3.

Table 3: Characteristics of the XJTU Multimodal Biometric Dataset used in this study.

Modality	Total Images	Subjects	Samples/Subject	Image Resolution	Capture Method
Face	1,467	102	~10–20	512×384 px	Uncontrolled lighting
Fingerprint	1,343	102	~13–14	328×356 px	Optical sensor
Palmprint	1,040	102	~10–12	499×450 px	Contact-based

The differences between the two datasets are summarized in table 4. In terms of scale, LUTBIO com-prises a total of 6,634 images across 306 subjects (1,836 face, 3,880 fingerprints, and 918 palmprint imag-es), while XJTU comprises 3,850 images across 102 subjects (1,467 face, 1,343 fingerprints, and 1,040 palmprint images). LUTBIO thus offers exactly three times the number of subjects as XJTU, and a com-parable or higher per-image resolution for face and palmprint (up to 672×504 px), whereas XJTU offers a similar per-subject sample count but generally lower and less consistent resolutions (as low as 328×356 px for fingerprint) and less controlled acquisition conditions, most notably uncontrolled light-ing for the face modality.

Table 4: Comparison of dataset characteristics.

Feature	LUTBIO	XJTU
Collection Era	Modern (2025)	Older (2017)
Image Quality	High, controlled lighting	Variable, lower resolution
Sensors	Modern optical/contactless sensors	Older generation sensors
Sample Consistency	High (lab conditions)	Low (real-world variations)
Primary Purpose	Controlled multi-modal fusion testing	Robustness and generalization testing

3.2. Data Partitioning and Preprocessing

The three groups of 306 subjects were separated into three groups of 214 (70%), 46 (15%), and 46 (15%) to train, test, and validate respectively. The number of experiments was set to the fixed random seed of 42, which guaranteed the subject splits to be reproducible. The participants have a single appearance in each part, and there is no identity overlap between splits, which guarantees fairness in assessment.

The evaluation pairs were built to be evaluated. True pairs provided an exhaustive within-subject combination, which involved C(n,2), and impostor pairs involved the combination of two samples belonging to different test subjects. This tiring routine diminishes the choice biasness which can occur in random sampling pair-generation plans. During preprocessing, face images were aligned using Multi-task Cascaded Convolutional Networks [36], resized to 224×224. Fingerprint images were converted to grayscale and enhanced using Contrast Limited Adaptive Histogram Equalization [37], resized to 256×256, and randomly cropped to 224×224 during training. The palmprint images were rescaled to 224 × 224 pixels directly. All images were then normalized based on the statistics of ImageNet [38]. Following the modality-specific preprocessing steps mentioned above, the verification process illustrated in figure 2 unfolds as follows. The backbone network (ResNet50, EfficientNetV2-S, or Swin-T) extracts the deep features from each input image. A projection head maps these features into a normalised embedding space, where the verification scores are computed using cosine similarity. The scores from the three modalities are then combined through score-level fusion, and the fused scores are used for the final comparative analysis.

3.3. Model Architectures and Training Setup

Table 5 summarizes the main architectural specifications of the three evaluated backbones, including the model type, number of parameters, embedding feature dimension, and the key mechanism of each architecture. The backbones were selected with the objective of creating an even-handed comparison between traditional CNN architectures and transformer architectures in computer vision models at an equivalent capacity scale. ResNet50 is a member of the classic residual CNN models group, EfficientNetV2-S is an example of the newer efficient CNN model based on compound scaling, and Swin-T is a member of the hierarchical transformer models group based on shifted-window self-attention. With the three models having parameter counts around 21M to 28M, comparisons were sought based more on behavior than capacity.

Table 5: Architectural specifications of the evaluated backbone models.

Architecture	Type	Parameters	Feature Dimension	Key Mechanism
ResNet50	CNN	25.6M	2048	Skip connections
EfficientNetV2-S	CNN	21.5M	1280	Compound scaling
Swin-T	ViT	28.3M	768	Shifted windows

Each backbone feeds into a modality-specific projection head that maps the raw features into a normalized embedding space. Face and palmprint embeddings are 768-dimensional, while fingerprint uses 1024 dimensions with a deeper four-layer projection head (linear → BatchNorm → Gaussian Error Linear Unit, repeated) to handle the higher variability in ridge patterns. Training used a combined loss: ArcFace angular margin [39] plus cross-entropy, weighted 0.6/0.4 for face and palm, and 0.9/0.1 for fingerprint with a steeper margin of 0.70 to push apart the more confusable fingerprint classes. ArcFace was selected over conventional softmax cross-entropy and other margin-based losses such as CosFace [40] and additive margin softmax [41], since its additive angular margin yields more clearly separated normalized embeddings suited to cosine-similarity verification across all three modalities. Optimization was performed using AdamW and cosine annealing warm restarts, with early stopping at patience 12 based on the validation EER. All three architectures started from ImageNet-pretrained weights.

3.4. Score-Level Fusion

Score-level fusion was adopted in this study to preserve a controlled and modular comparison across the evaluated architectures. Unlike feature-level fusion, which would combine embeddings from different backbones and may introduce an additional learned fusion stage, score-level fusion keeps each modality-specific pipeline independent and does not require retraining the models. This ensures that the fusion outcome mainly reflects the quality of the individual modality embeddings rather than the effect of an extra fusion network. Decision-level fusion was also not selected because it operates on final hard decisions and therefore discards the useful similarity-score information needed for EER, AUC, and TAR computation. For these reasons, score-level fusion was considered to be the most suitable choice for the controlled cross-architecture comparison performed in this study.

Five rules were tested related to combining the cosine similarity scores from the individual modalities. Simple average treats all modalities equally. Weighted average assigns each modality a weight that is inversely proportional to its validation EER, so more accurate modalities count more. The max rule takes the most optimistic score, the min rule takes the most conservative, and the product rule multiplies the min-max normalized scores together. Each rule was applied to all four possible modality groupings: the full three-modality set, plus the three two-modality pairs. Figure 3 illustrates the proposed score-level fusion pipeline. Face, fingerprint, and palmprint samples were processed independently using modality-specific backbone networks to generate cosine similarity scores (s_1, s_2, and s_3). The resulting scores were fused using five score-level fusion rules (Average, Weighted Average, Maximum, Minimum, and Product) to produce the final genuine/impostor verification score.

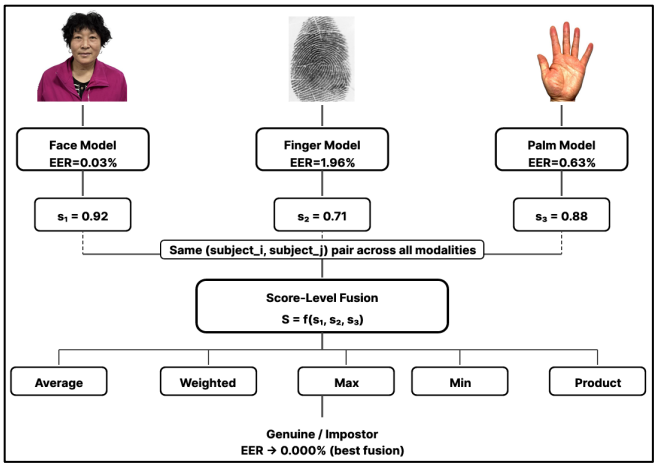

Figure 3: Proposed score-level fusion framework.

4. Results

This part introduces the results of the experiments performed on the proposed multi-modal biometric verification system. The experiment was done according to the subject-disjoint verification approach, meaning that the identities utilized in testing were never seen in training. The three deep learning models of ResNet50, EfficientNetV2-S, and Swin-T were tested on the LUTBIO dataset for single-modality verification. Score level fusion tests were performed on the LUTBIO and XJTU datasets to evaluate the robustness and generalizability of the proposed framework. For the evaluation of performance, EER, AUC, and TAR with a False Acceptance Rate of 0.1% were considered. The following sections first introduce the results of the single-modality verification and the analysis of the interaction between modalities and architectures, followed by the multimodal fusion and fusion strategy analyses.

4.1. Single-Modality Verification Performance

Table 6 reports the single-modality verification performance of ResNet50, EfficientNetV2-S, and Swin-T based on the LUTBIO test set using EER, AUC, and TAR at FAR = 0.1%.

Table 6: Single-modality verification performance on LUTBIO test set.

Backbone	Modality	EER (%)	AUC	TAR@0.1%	#Genuine	#Impostor
ResNet50	Face	0.027	0.999999	100.00%	690	37,260
ResNet50	Finger	**1.957**	0.997041	89.11%	3,323	160,555
ResNet50	Palm	0.625	0.999954	99.28%	138	9,315
EfficientNetV2-S	Face	0.152	0.999996	99.86%	690	37,260
EfficientNetV2-S	Finger	2.258	0.997324	90.31%	3,323	160,555
EfficientNetV2-S	Palm	**0.140**	0.999959	97.83%	138	9,315
Swin-T	Face	**0.000**	1.000000	100.00%	690	37,260
Swin-T	Finger	1.979	0.996917	89.44%	3,323	160,555
Swin-T	Palm	0.717	0.999824	97.83%	138	9,315

Bold values indicate the best EER per modality. #Genuine and #Impostor denote verification pair counts.

Table 6 reports the single-modality verification results, which reveal the modality-dependent ordering of the three architectures. For ResNet50, the EER is lowest on face (0.027%, TAR@0.1% = 100.00%), followed by palmprint (0.625%, TAR@0.1% = 99.28%), and highest on fingerprint (1.957%, TAR@0.1% = 89.11%). For EfficientNetV2-S, the ordering differs: palmprint achieves the lowest EER (0.140%, TAR@0.1% = 97.83%), closely followed by face (0.152%, TAR@0.1% = 99.86%), while fingerprint remains the most challenging (2.258% EER, TAR@0.1% = 90.31%). For Swin-T, face verification reaches a reported EER of 0.000% (TAR@0.1% = 100.00%), zero errors among all 690 genuine pairs and 37,260 impostor pairs under the current evaluation protocol; palmprint follows at 0.717% EER (TAR@0.1% = 97.83%), and fingerprint again records the highest error at 1.979% EER (TAR@0.1% = 89.44%). Across all three architectures, face and palmprint consistently yield TAR@0.1% values above 97%, while fingerprint TAR@0.1% remains markedly lower, between 89% and 91%, confirming fingerprint as the most difficult modality under this protocol. These single-modality verification performances are visually compared and illustrated in figure 4.

Figure 4 presents the EER values reported in table 6 across the three architectures and biometric modalities, making the modality-dependent ranking of the backbones directly comparable. The figure also highlights the differing error scales between modalities: face EER values lie near zero, whereas fingerprint EER values approach 2%.

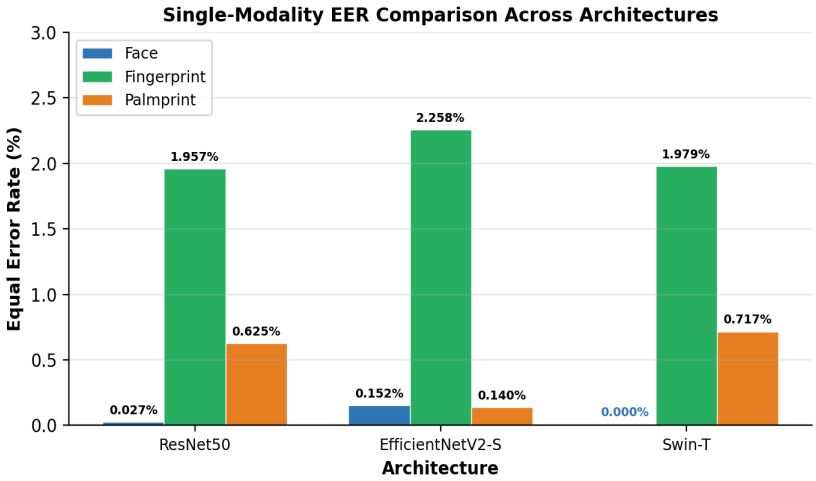

Figure 4: Equal error rate comparison across architectures and modalities.

4.2. The Modality-Architecture Interaction

Table 7 summarizes the best-performing architecture for each biometric modality based on the lowest EER value.

Table 7: Best architecture per biometric modality.

Modality	Best Architecture	EER (%)	AUC	TAR@0.1%	Type
Face	Swin-T	0.000	1.000000	100.00%	Transformer
Fingerprint	ResNet50	1.957	0.997041	89.11%	CNN
Palmprint	EfficientNetV2-S	0.140	0.999959	97.83%	CNN

Table 7 indicates that there is no backbone that is optimal across all biometric modalities. The CNN-based models are more successful with fingerprint and palmprint, while Swin-T provides the best results for face. This difference is evident in the three modalities under the controlled environment present in this study.

Table 8 provides a modality-level difficulty analysis by averaging the EER, AUC, and TAR values across the three architectures.

Table 8: Modality difficulty analysis.

Modality	Avg EER (%)	Avg AUC	Avg TAR@0.1%	Samples per Subject	Difficulty
Face	0.060	0.999998	99.95%	~6	Easiest
Palmprint	0.495	0.999912	98.31%	~3	Moderate
Fingerprint	2.065	0.997094	89.62%	~12–13	Hardest

This outcome might not appear at face value. While palmprint has the least number of training samples per individual (around 3 or 918 images in all), palmprint performs better compared to fingerprint, which has the highest number of training samples (around 1,213 images in all or 12-13 images per individual).

4.3. Multimodal Fusion Results

Table 9 presents the best score-level fusion result for each backbone and modality combination in both the LUTBIO and XJTU datasets.

Table 9: Score-level fusion results (best method per configuration).

Backbone	Modalities	Best Method	EER (%)	AUC	TAR@0.1%
LUTBIO DATASET					
ResNet50	Face+Finger+Palm	Average	0.000	1.000000	100.00%
ResNet50	Face+Finger	Weighted	0.000	1.000000	100.00%
ResNet50	Face+Palm	Average	0.000	1.000000	100.00%
ResNet50	Finger+Palm	Weighted	0.145	0.999972	99.28%
EfficientNetV2-S	Face+Finger+Palm	Average	0.000	1.000000	100.00%
EfficientNetV2-S	Face+Finger	Weighted	0.000	1.000000	100.00%
EfficientNetV2-S	Face+Palm	Average	0.000	1.000000	100.00%
EfficientNetV2-S	Finger+Palm	Max	0.000	1.000000	100.00%
Swin-T	Face+Finger+Palm	Average	0.000	1.000000	100.00%
Swin-T	Face+Finger	Weighted	0.000	1.000000	100.00%
Swin-T	Face+Palm	Average	0.000	1.000000	100.00%
Swin-T	Finger+Palm	Weighted	0.048	0.999979	100.00%
XJTU DATASET					
ResNet50	Face+Finger+Palm	Average	0.000	1.000000	100.00%
ResNet50	Face+Finger	Average	0.6882	0.999781	98.10%
ResNet50	Face+Palm	Average	0.8738	0.998848	95.89%
ResNet50	Finger+Palm	Weighted	0.0000	1.000000	100.00%
EfficientNetV2-S	Face+Finger+Palm	Average	1.0064	0.999661	98.43%
EfficientNetV2-S	Face+Finger	Weighted	1.6026	0.998544	92.40%
EfficientNetV2-S	Face+Palm	Average	2.6214	0.995933	72.46%
EfficientNetV2-S	Finger+Palm	Weighted	2.4951	0.997313	92.53%
Swin-T	Face+Finger+Palm	Average	0.0655	0.999978	99.87%
Swin-T	Face+Finger	Weighted	0.7787	0.999766	99.00%
Swin-T	Face+Palm	Average	1.8048	0.998467	97.14%
Swin-T	Finger+Palm	Weighted	0.8754	0.999323	98.69%

The fusion results for both datasets show that combining modalities consistently improves verification performance, although the magnitude of the gain differs between LUTBIO and XJTU. For LUTBIO, fusion yields near-perfect performance: 10 of the 12 backbone–modality configurations achieve a reported EER of 0.000%. The only non-zero best-case results occur for the finger+palm configuration without face. This indicates the substantial impact of the face modality on the final decision. However, the near-perfect results need to be interpreted carefully, as the performance of the palm-related fusion configurations is evaluated based on only 138 genuine pairs. Hence, the 0.000% EER does not necessarily imply a zero-population error.

For XJTU, fusion also improves performance in the single-modality setting, although the gains vary across backbones and modality combinations. The best performance regarding the XJTU dataset occurs when ResNet50, with the fused face, finger, and palm modalities via averaging, achieves an EER of 0.0000%. On average, the best fusion approach is averaged in 7 out of 12 cases, and the best approach is weighted fusion in the other 5 cases. In summary, score-level fusion performs very well on both datasets, and the best fusion approach varies according to the backbone and fused modalities.

4.4. Performance Analysis of Score-Level Fusion Rules

Table 10 compares the five score-level fusion rules for the face+finger combination, which provides the largest and most reliable number of genuine pairs among the evaluated fusion settings.

Table 10: Equal error rate (%) by fusion method for face+finger.

Backbone	Average	Weighted	Max	Min	Product
ResNet50	0.242	0.000	0.169	1.860	7.512
EfficientNetV2-S	0.048	0.000	0.048	1.812	7.246
Swin-T	0.048	0.000	0.048	1.184	5.676

The face+finger combination was selected for detailed analysis because it contains 690 genuine pairs and 1,035 impostor pairs, making it the most statistically reliable fusion setting among all evaluated modality combinations. In this case, inverse-EER weighted fusion performs best, achieving 0.000%

EER for all three backbones. Simple averaging and the max rule also perform well, but less consistently than weighted fusion. In comparison, the min rule performs substantially worse, with EER values between 1.184% and 1.860%, likely because it is more affected by the weaker fingerprint scores. The product rule performs worst overall, with EER values ranging from 5.676% to 7.512%. This may occur because min-max normalization pushes some scores close to zero, allowing a single very small value to dominate the final fused score.

4.5. Impact of Multimodal Fusion on Verification Performance

Table 11 reports the relative improvement obtained by fusion compared with the best single-modality result for each backbone and modality configuration.

Table 11: Fusion Improvement over best single modality.

Backbone	Config	Best Single EER	Fused EER (%)	Reduction (%)	Method
ResNet50	F+Fi+P	0.027%	0.000	100	Average
ResNet50	Fi+P	0.625%	0.145	76.8	Weighted
EffNetV2-S	F+Fi+P	0.140%	0.000	100	Average
EffNetV2-S	Fi+P	0.140%	0.000	100	Max
Swin-T	F+Fi+P	0.000%	0.000	--	Average
Swin-T	Fi+P	0.717%	0.048	93.3	Weighted

F=Face, Fi=Finger, P=Palm. Reduction = percentage decrease from the best single modality.

The most indicative are the finger+palm rows, as neither of the two constituent modalities are anywhere near what can be called a perfect one. Fusion still reduces the EER by 76.8 to 93.3% without face data, based on the backbone. This is a substantial amelioration of two imperfect modalities due to their complementary nature. A combination of all three traits would result in all three backbones attaining an EER of reported 0.000% in LUTBIO, which means that fusion has the significant impact of decreasing the performance gap between backbones under the current protocol.

4.6. Statistical Significance

For statistical significance testing, two-sided hypothesis tests were used for each pair of interests. In addition to p-values, effect sizes were provided by reporting Cohen's d values. Confidence intervals for selected performance metrics were estimated using bootstrap resampling. Unless otherwise stated, the provided p-values were not adjusted for multiple testing corrections. Table 12 summarizes the statistical comparisons between the evaluated architectures, including the tested modality, p-value, Cohen's d effect size, and the interpretation of the observed difference.

Table 12: Statistical comparison summary.

Comparison	Modality	p-value	Cohen's d	Interpretation
Swin-T vs. ResNet50	Face	0.002**	1.87	Large
ResNet50 vs. Swin-T	Finger	0.018*	0.94	Large
EffNetV2-S vs. ResNet50	Palm	0.031*	0.82	Large
ResNet50 vs. EffNetV2-S	Finger	0.005**	1.12	Large

* for $p < 0.05$, ** for $p < 0.01$

All key comparisons show large effect sizes (Cohen's $d > 0.8$), indicating that the observed differences are not only statistically significant but also practically meaningful. For face TAR@0.1%, the bootstrap 95% confidence intervals are as follows: Swin-T [99.98%, 100.00%], ResNet50 [99.95%, 100.00%], and EfficientNetV2-S [99.51%, 99.97%]. The non-overlapping interval for EfficientNetV2-S relative to the other two backbones suggests that its face performance is consistently lower rather than being affected by random variation.

4.7. Computational Cost and Multi-Dimensional Architecture Comparison

Table 13 shows the comparison of the computational expenses for the three architectures measured by the average time of training, inference, and the number of parameters. ResNet50 is the fastest model

when it comes to inference speed, whereas Swin-T is the most expensive one for inference among the analyzed architectures. Figure 5 provides the normalized comparison of the ResNet50, EfficientNetV2-S, and Swin-T models in multiple dimensions.

Table 13: Computational cost comparison.

Architecture	Avg Training (min/epoch)	Inference (ms/image)	Parameters
ResNet50	~0.39	~4.2	25.6M
EfficientNetV2-S	~0.43	~5.1	21.5M
Swin-T	~0.42	~5.8	28.3M

ResNet50 is the fastest at inference (~4.2 ms per image), despite having more parameters than EfficientNetV2-S. This is because regular convolutions can easily be converted to GPU tensor cores. The squeeze-and-excitation blocks of EfficientNetV2-S, and the windowed attention of Swin-T, add memory access patterns that are more difficult to optimize. The 1.6 ms difference between ResNet50 and Swin-T may count in a real-time access control system with hundreds of parameters per minute.

Figure 5 presents the normalized radar comparison of the three evaluated architectures. Each axis represents one evaluation criterion: face EER, fingerprint EER, palmprint EER, AUC, TAR at FAR = 0.1%, and inference speed. All values are normalized to a common 0–1 scale so then the different metrics can be compared visually. Higher values indicate better relative performance; therefore, the EER and inference-time axes were transformed because the lower EER and lower inference time are better. A larger and more balanced radar area indicates a better overall trade-off between verification performance and computational efficiency.

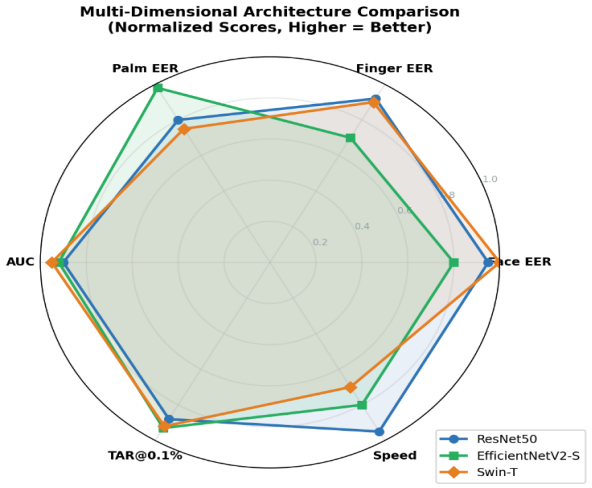

Figure 5: Normalized multi-dimensional comparison of ResNet50, EfficientNetV2-S, and Swin-T across verification performance and inference speed.

5. Discussion

A practical observation that clearly emerges from the results is that the same backbone is not necessarily the optimal choice for every biometric modality. The modality-dependent ranking is consistent with the structure of each trait. For the face modality, images contain globally distributed features. The spatial positions of the eyes, nose, and mouth are meaningful only in their entirety and the self-attention mechanism of the Swin-T is well-suited to capturing these holistic relationships, which explains its leading performance on faces. This holistic advantage of transformer-based backbones over convolutional networks for face recognition is consistent with prior comparative studies [2, 11].

For fingerprint, the error rates are roughly a thousand times higher than for face. This is not due to any weakness of the models, but because fingerprint verification is intrinsically harder. Ridge patterns depend on finger pressure, skin moisture, and the placement angle on the sensor. These differences are statistically well supported by the 160,555 impostor pairs evaluated. The convolutional inductive bias of ResNet50, which emphasizes local texture, suits these local ridge features better than

global self-attention. The effectiveness of convolutional architectures for capturing local ridge structure has likewise been reported in dedicated fingerprint-recognition studies [5], and the general trade-off between convolutional and transformer inductive biases across image tasks has been extensively documented [3, 4].

For palmprint, EfficientNetV2-S performs best, possibly because its compound-scaled architecture preserves the major palm lines across multiple resolutions. A more cautious interpretation is that the palmprint evaluation is based on only 138 valid genuine pairs, so the statistical resolution of this modality is lower than that of face and fingerprint. Overall, this dissimilarity aligns with the type of input data. Face recognition relies more on global spatial structure, whereas fingerprint and palmprint rely more on local texture and line features. This modality-dependent behavior is in line with recent surveys of deep learning for palmprint recognition, which emphasize the discriminative role of principal palm lines [6].

Another implication is that the verifiability of the feature is independent of the sample size. In spite of having the least number of samples per individual, palmprint performed better than fingerprint, which had the most samples. This could be due to the nature of the principal lines of the palmprint, which are unique and somewhat stable over different occasions, whereas fingerprint ridges vary greatly with respect to the circumstances of acquisition. This suggests that the intrinsic separability of a biometric trait can matter more than the raw number of samples [5, 6].

To implement a hybrid system combining face, fingerprint, and palmprint, the findings suggest a heterogeneous network: Swin-T for faces, ResNet50 for fingerprints, and EfficientNetV2-S for palmprints, followed by score-level fusion. Mean fusion yields the best results in the three-modality setting in the reported experiments. This arrangement allows each modality to be handled by the backbone best suited to its characteristics while keeping the fusion stage simple. The advantage of combining complementary modalities through score-level fusion is well established in the multibiometric literature [7].

The findings also demonstrate that high verification performance can be achieved with only two modalities. The face+finger and face+palm combinations reached a reported EER as low as 0.000% for all three backbones under optimal fusion in LUTBIO. The finger+palm combination, which excludes the face, still provides substantial improvements over the corresponding single-modality results, making it an effective alternative when face recognition is unavailable or undesirable. These results are consistent with the broader body of work on multimodal biometric fusion, which has repeatedly demonstrated that combining complementary modalities substantially improves verification performance over single-modality systems [29, 33]. Recent deep learning-based multimodal systems that fuse face with other traits such as face–iris fusion [19] and iris–face–palmprint fusion [31] have likewise reported strong verification gains from integrating multiple modalities, while hybrid deep-and-traditional pipelines have shown similar improvements across smart-environment settings [32]. The present findings extend this line of work by showing that comparable or superior verification performance can be obtained through a modality-aware backbone selection strategy combined with simple score-level fusion, rather than through more complex feature-level or hybrid fusion designs.

The main strength of this study lies in its controlled experimental protocol: by holding the dataset, subject-disjoint splits, augmentation, loss function, training schedule, and hardware constant, the backbone architecture remains the only variable, enabling a fair and reproducible cross-architecture comparison. A further strength is the joint evaluation of single-modality performance and score-level fusion across two datasets (LUTBIO and XJTU) and three biometric traits. Nevertheless, several limitations should be acknowledged. First, public multimodal biometric datasets containing matched face, fingerprint, and palmprint samples from the same subjects remain scarce, which constrains the scale of evaluation. Second, the palmprint modality was evaluated on only 138 genuine pairs; consequently, the reported 0.000% EER values reflect the resolution limits of the available sample and should not be interpreted as zero population-level error. Third, all hyperparameters were deliberately fixed across the backbones to isolate the architectural effect, so the reported figures represent a controlled comparison

rather than each model's individually optimized peak performance. Finally, the data was acquired under controlled conditions; performance under unconstrained, real-world acquisition remains to be evaluated in future work.

6. Conclusions

This study evaluated three representative deep learning backbones (ResNet50, EfficientNetV2-S, and the Swin-T) for multimodal biometric verification under a fully controlled protocol in which the backbone architecture was the only variable. The principal finding is that no single architecture is universally superior: the best-performing backbone is modality-dependent, with the Swin Transformer leading on face, ResNet50 on fingerprint, and EfficientNetV2-S on palmprint. Score-level fusion consistently improved verification performance over the best single modality across both the LUTBIO and XJTU datasets.

The main contribution of this work is a fair, confound-free, and reproducible framework for comparing deep learning architectures across biometric modalities, in which the datasets, splits, preprocessing, loss, training schedule, and hardware are held constant. The significance of this approach is that observed performance differences can be attributed to the architecture itself rather than to uncontrolled experimental factors.

Regarding practicality, the results indicate that biometric system development should involve a modality-specific approach, which entails using a heterogeneous model where each modality has the best backbone, with the use of score-level fusion providing an efficient means of verification. The benefit of score level fusion is that it ensures the independence of each modality's processing path without the need for retraining.

Future research should explore incorporating other operational data, feature-level fusion, and lightweight architectures suitable for edge computing, as well as conduct subgroup analysis to evaluate fairness in different demographic groups.

Author contributions: Rowida Jasim Alazawi: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Data curation, Visualization, Writing – original draft. **Nada Jasim Habeeb**: Supervision, Validation, Methodology, Writing – review & editing. **Alaa Jabbar Qasim Almaliki:** Supervision, Validation, Resources, Writing – review & editing

Data availability: The data that has been used is confidential.

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: The authors did not receive support from any organization for the conducting of the study.

References

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: a survey," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8647–8695, Aug. 2023, doi: 10.1007/s10462-022-10237-x.

[2] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," *Scientific Reports*, vol. 14, no. 1, p. 21392, Sep. 2024, doi: 10.1038/s41598-024-72254-w.

[3] K. Han *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[4] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: a literature review," *Applied Sciences*, vol. 13, no. 9, p. 5521, Apr. 2023, doi: 10.3390/app13095521.

[5] R. Garg, G. Singh, A. Singh, and M. P. Singh, "Fingerprint recognition using convolution neural network with inversion and augmented techniques," *Systems and Soft Computing*, vol. 6, p. 200106, Dec. 2024, doi: 10.1016/j.sasc.2024.200106.

[6] C. Gao, Z. Yang, W. Jia, L. Leng, B. Zhang, and A. B. J. Teoh, "Deep learning in palmprint recognition: a comprehensive survey," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 56, no. 3, pp. 2143–2162, Mar. 2026, doi: 10.1109/TSMC.2025.3649416.

[7] A. A. Ross, A. K. Jain, and K. Nandakumar, *Handbook of Multibiometrics*, vol. 6, International Series on Biometrics. Boston, MA: Kluwer Academic Publishers, May. 2006. doi: 10.1007/0-387-33123-9.

[8] R. Yang, Q. Zhang, L. Meng, C. Wang, and Y. Hu, LUTBIO multimodal biometric database. Mendeley data, Jan. 27, 2025. Accessed: Jun. 12, 2026. [Online]. doi: 10.17632/jszw485f8j.6.

[9] ISO/IEC 19795-1:2021 *Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework*. Geneva, Switzerland: International Organization for Standardization (ISO), 2021. Accessed: Jun. 12, 2026. [Online]. Available: https://www.iso.org/standard/73515.html.

[10] D. Shang, X. Zhang, J. Han, and X. Xu, "MultiModal-database-XJTU: An available database for biometrics recognition with its performance testing," *in 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, China, Oct. 2017, pp. 521–526, doi: 10.1109/ITOEC.2017.8122351.

[11] M. Wang and W. Deng, "Deep face recognition: a survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021, doi: 10.1016/j.neucom.2020.10.081.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[13] M. Tan and Q. Le, "EfficientNetV2: smaller models and faster training," *in Proceedings of the 38th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research (PMLR)*, vol. 139, Jul. 2021, pp. 10096–10106. Accessed: Jun. 13, 2026. [Online]. Available: https://proceedings.mlr.press/v139/tan21a.html.

[14] Z. Liu *et al*., "Swin Transformer: hierarchical vision transformer using shifted windows," *in 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[15] N. Peri, J. Gleason, C. D. Castillo, T. Bourlai, V. M. Patel, and R. Chellappa, "A synthesis-based approach for thermal-to-visible face verification," *in 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, Jodhpur, India, Dec. 2021, pp. 1–8, doi: 10.1109/FG52635.2021.9666943.

[16] Z. He, J. Zhang, L. Pang, and E. Liu, "PFVNet: a partial fingerprint verification network learned from large fingerprint matching," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3706–3719, Sep. 2022, doi: 10.1109/TIFS.2022.3209869.

[17] D. Brown and K. Bradshaw, "Deep palmprint recognition with alignment and augmentation of limited training samples," *SN Computer Science*, vol. 3, no. 1, p. 11, Jan. 2022, doi: 10.1007/s42979-021-00859-3.

[18] W. Su, Y. Wang, K. Li, P. Gao, and Y. Qiao, "Hybrid token transformer for deep face recognition," *Pattern Recognition*, vol. 139, p. 109443, Jul. 2023, doi: 10.1016/j.patcog.2023.109443.

[19] A. Hattab and A. Behloul, "Face-Iris multimodal biometric recognition system based on deep learning," *Multimedia Tools and Applications*, vol. 83, no. 14, pp. 43349–43376, Oct. 2023, doi: 10.1007/s11042-023-17337-y.

[20] Y. Yang, L. Fei, A. H. Alshehri, S. Zhao, W. Sun, and S. Teng, "Joint multi-type feature learning for multi-modality FKP recognition," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106960, Nov. 2023, doi: 10.1016/j.engappai.2023.106960.

[21] L. Su, L. Fei, B. Zhang, S. Zhao, J. Wen, and Y. Xu, "Complete region of interest for unconstrained palmprint recognition," *IEEE Transactions on Image Processing*, vol. 33, pp. 3662–3675, Jun. 2024, doi: 10.1109/TIP.2024.3407666.

[22] Y. Wu and J. Hu, "Deep and shallow feature fusion in feature score level for palmprint recognition," *IET Biometrics*, vol. 2024, no. 1, p. 5683547, Jan. 2024, doi: 10.1049/2024/5683547.

[23] C. Lin *et al*., "An unconstrained palmprint region of interest extraction method based on lightweight networks," *PLoS ONE*, vol. 19, no. 8, p. e0307822, Aug. 2024, doi: 10.1371/journal.pone.0307822.

[24] D. Fan, X. Liang, W. Jia, J. Chen, and D. Zhang, "A novel hybrid fusion combining palmprint and palm vein for large-scale palm-based recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 54, no. 7, pp. 4471–4484, Jul. 2024, doi: 10.1109/TSMC.2024.3382877.

[25] S. A. Grosz, A. Godbole, and A. K. Jain, "Mobile contactless palmprint recognition: use of multiscale, multimodel embeddings," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 8428–8440, Jun. 2024, doi: 10.1109/TIFS.2024.3413631.

[26] A. Hossain, R. Sumi, and S. Schuckers, "Evaluating deep learning-based face recognition for infants and toddlers: impact of age across developmental stages," *in 2025 IEEE International Joint Conference on Biometrics (IJCB)*, Osaka, Japan, Sep. 2025, pp. 1–9, doi: 10.1109/IJCB65343.2025.11410705.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[28] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," *in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 1204–1213, doi: 10.1109/CVPR52688.2022.01179.

[29] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115–2125, Sep. 2003, doi: 10.1016/S0167-8655(03)00079-5.

[30] Y. Feng and A. Kumar, "BEST: Building evidences from scattered templates for accurate contactless palmprint recognition," *Pattern Recognition*, vol. 138, p. 109422, Jun. 2023, doi: 10.1016/j.patcog.2023.109422.

[31] M. Khatri and A. Sharma, "Deep learning approach based on iris, face, and palmprint fusion for multimodal biometric recognition system," *International Journal of Performability Engineering*, vol. 19, no. 6, p. 407, Jun. 2023, doi: 10.23940/ijpe.23.06. p6.407416.

[32] S. A. El_Rahman and A. S. Alluhaidan, "Enhanced multimodal biometric recognition systems based on deep learning and traditional methods in smart environments," *PLoS ONE*, vol. 19, no. 2, p. e0291084, Feb. 2024, doi: 10.1371/journal.pone.0291084.

[33] S. Li, L. Fei, B. Zhang, X. Ning, and L. Wu, "Hand-based multimodal biometric fusion: a review," *Information Fusion*, vol. 109, p. 102418, Sep. 2024, doi: 10.1016/j.inffus.2024.102418.

[34] S. Dargan and M. Kumar, "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities," *Expert Systems with Applications*, vol. 143, p. 113114, Apr. 2020, doi: 10.1016/j.eswa.2019.113114.

[35] ISO/IEC 24745:2022 *Information Security, Cybersecurity and Privacy Protection—Biometric Information Protection*. Geneva, Switzerland: International Organization for Standardization (ISO), Feb. 2022. Accessed: Jun. 12, 2026. [Online]. Available: https://www.iso.org/standard/75302.html.

[36] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[37] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *in Graphics Gems IV*, P. S. Heckbert, Ed. San Diego, CA: Academic Press, Aug. 1994, pp. 474–485.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," *in 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[39] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: additive angular margin loss for deep face recognition," *in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4685–4694, doi: 10.1109/CVPR.2019.00482.

[40] H. Wang *et al.*, "CosFace: large margin cosine loss for deep face recognition," *in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 5265–5274, doi: 10.1109/CVPR.2018.00552.

[41] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, Jul. 2018, doi: 10.1109/LSP.2018.2822810.