A Unified Benchmarking Framework for Offline Handwritten Signature Verification Using Deep Learning Architectures

Eissa Alreshidi
Associate Professor at College of Computer Science and Engineering, University of Hail, Hail, P.O. Box 2440, Saudi Arabia

***Abstract*—Offline handwritten signature verification (OSV) remains a challenging biometric task owing to the subtle variability of genuine signatures and the sophistication of skilled forgeries. This study introduces a unified benchmarking framework for evaluating eight deep learning architectures—CNN Shallow, CNN Deep, ResNet-18, ResNet-34, MobileNetV2, EfficientNet-B0, ViT-Tiny, and a CNN–Transformer Hybrid—within a writer-independent Siamese contrastive-learning paradigm. The framework standardizes preprocessing, balanced pair generation, NVIDIA A100 GPU training, and a comprehensive evaluation suite that includes ROC and Precision-Recall curves, Equal Error Rate (EER), calibration analysis, threshold-sensitivity metrics, and embedding visualizations using PCA, t-SNE, and UMAP. Experiments conducted on an NVIDIA A100 GPU reveal a clear performance stratification: six architectures achieve perfect verification performance (Accuracy = 1.0, ROC AUC = 1.0, PR AUC = 1.0, EER = 0.0), supported by consistently well-separated embedding manifolds and highly stable calibration behavior. In contrast, MobileNetV2 and EfficientNet-B0 exhibited elevated EER values and overlapping embeddings, underscoring the limitations of lightweight and compound-scaled models in capturing the fine-grained stroke morphology. The proposed framework establishes a transparent and extensible foundation for future research, enabling fair cross-model comparisons and guiding the development of robust and deployment-ready biometric verification systems. In addition, this study provides the first fully controlled, architecture-agnostic comparison of CNNs, residual networks, lightweight mobile models, and transformer-based architectures under identical, writer-independent conditions. By eliminating variability in preprocessing, pair generation, and training configuration, the framework isolates the true effect of the architectural design on the verification performance. The findings highlight the importance of embedding separability, calibration stability, and threshold robustness—factors often overlooked in prior OSV research but essential for real-world deployment.**

Keywords*—*Offline handwritten signature verification; Siamese networks; contrastive-learning; deep learning architectures; ResNet; MobileNetV2; EfficientNet; vision transformers; hybrid CNN-transformer models; biometric authentication; forgery detection; embedding separability; writer-independent verification

I. Introduction

Offline handwritten signature verification (OSV) has long been a central topic in biometric research, driven by its importance in financial, legal, and administrative authentication systems. Unlike online verification, which benefits from dynamic information, such as pen pressure and stroke velocity, offline verification relies solely on static images, making the task inherently more challenging [1]. In the last twenty years, the field has shifted from handcrafted texture descriptors [2] and statistical distance measures [3] to deep learning methods that learn discriminative representations directly from data [4], [5]. This progression reflects a broader shift toward representation learning, where models must capture subtle intra-writer variations while remaining robust to highly skilled forgeries [4].

Comparing findings across OSV studies remains surprisingly difficult, despite steady progress in recent years. Research groups often rely on different preprocessing steps, pairing strategies, training routines, and evaluation metrics, and these inconsistencies make it difficult to determine whether performance differences stem from the models themselves or from the surrounding experimental setup [6]. As a result, the literature reports a wide spread of accuracies and EER values, even for architectures trained on the same datasets, which complicates efforts to identify models that truly capture the subtle morphological cues separating genuine signatures from skilled forgeries. Without a consistent experimental pipeline, architectural comparisons become tentative, and claims regarding model superiority lose much of their reliability [4].

This study addresses this issue by introducing a unified benchmarking framework that evaluates eight deep learning architectures under strictly controlled and identical conditions. Every stage of the process, from dataset preparation to training configuration and evaluation, follows a standardized procedure, allowing architectural differences to emerge without interference from external variables. This controlled design reveals clear performance groupings and offers fresh insights into how inductive biases, representational capacity, and embedding space structure influence verification outcomes. By removing confounding factors, the framework provides a level of comparability that has been largely absent from previous OSV research.

This study makes three primary contributions: 1) a unified, architecture-agnostic benchmarking framework that eliminates variability in preprocessing, pair generation, and evaluation; 2) the first controlled comparison showing that CEDAR exhibits saturation under writer-independent contrastive learning, revealing limits of current OSV benchmarks; and 3) an analysis linking embedding separability, calibration stability, and threshold robustness to architectural inductive biases, offering practical guidance for model selection and future dataset design.

Beyond standardization, this study reveals an important conceptual insight: under a strictly controlled writer-independent protocol, the CEDAR dataset exhibits clear signs of benchmark saturation. Six heterogeneous architectures achieve perfect separability, indicating that architectural differences cannot be meaningfully distinguished without stronger datasets or more challenging evaluation conditions. This finding reframes how OSV performance should be interpreted and highlights the need for next-generation benchmarks.

The study also places its findings in the context of existing work, showing where the results align with, diverge from, or extend earlier research on CNNs [4], residual networks [7], lightweight models [6], and transformer-based architectures [8]. The comparative insights that emerge from this analysis offer useful directions for choosing architectures suited to different verification settings. By connecting earlier methodological traditions with modern deep learning approaches, this study positions itself as a comprehensive reference point for future developments in offline signature verification.

The remainder of this study is organized as follows: Section II reviews prior OSV research and identifies gaps motivating this study. Section III describes the unified benchmarking framework, including preprocessing, pair generation, and architectural design. Section IV presents quantitative and qualitative results across all eight architectures. Section V discusses the implications of benchmark saturation, architectural behavior, and calibration stability. Section VI concludes with limitations and directions for future research.

II. Literature Review

Offline handwritten signature verification (OSV) has evolved through several methodological eras, each addressing the limitations of the previous era. The field has progressed from handcrafted descriptors to classical machine learning, deep convolutional models, metric learning architectures, lightweight mobile networks, and, more recently, transformer-based and hybrid designs. This section offers an organized overview of the evolution of the field, drawing attention to the motivations, strengths, and limitations that shaped each stage of development. Taken together, these shifts illustrate a broader movement within OSV, from early, manually crafted descriptors to fully learned representations, mirroring the larger transformation occurring across computer vision and biometric authentication. Table I presents the summary of prior offline signature verification studies.

A. Early Handcrafted Feature Approaches

Early OSV systems relied heavily on manually designed descriptors to capture the visual, geometric, and textural characteristics of a signature. Techniques such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Gabor filters, and directional probability density functions [2], [9]–[11] have been widely used, often alongside contour-based and graph-based representations. Although these methods performed reasonably well for their time, they were highly sensitive to noise, rotation, and natural variability in a writer's signature. Their limited expressive capacity makes it difficult to capture subtle, higher-level patterns that distinguish genuine signatures from skilled forgeries, especially under real-world variability. As datasets have become more diverse and real-world conditions more variable, these handcrafted pipelines have begun to show their limitations, encouraging a shift toward models capable of learning discriminative features directly from data rather than relying on predefined heuristics.

B. Classical Machine Learning and Statistical Models

Before deep learning became dominant, OSV research relied heavily on classical machine-learning techniques such as Support Vector Machines (SVMs), Hidden Markov Models (HMMs), Dynamic Time Warping (DTW), and k-Nearest Neighbors (k-NN). These methods were widely used in early biometric verification [12]–[14] and were adapted for OSV by studies such as [15]. These methods operate on handcrafted features and attempt to learn either discriminative boundaries or probabilistic models for each writer's signature. Although they represent an improvement over purely handcrafted pipelines, their dependence on feature engineering limits their ability to handle skilled forgeries or the high intra-class variability typical of real signatures. Furthermore, many of these systems rely on writer-dependent training, which inflates performance and restricts generalization to unseen writers, which is a key requirement for practical OSV applications. As a result, these approaches offered only modest gains and were unable to capture the subtle morphological differences that separate genuine signatures from convincing forgeries.

C. Emergence of Deep Learning for Signature Verification

The introduction of deep learning has resulted in a noticeable shift in the approach to OSV problems. Convolutional Neural Networks (CNNs) shown that models can learn layered, meaningful representations directly from raw pixel data, eliminating the dependence on handcrafted descriptors (LeCun et al., 1998). Early CNN-based work [4] demonstrated clear gains, especially in capturing subtle texture patterns, stroke-level variations, and writer-specific traits that earlier methods struggled to represent. However, CNNs trained in a traditional multiclass classification setting face a conceptual mismatch because OSV is inherently a similarity learning problem rather than a categorical recognition task. This mismatch motivated a shift toward metric-learning architectures. Thus, deep learning provides the representational power needed for OSV but requires a paradigm shift toward pair-based learning to fully exploit its potential.

D. Siamese Networks and Metric Learning

Siamese networks [16] are a natural fit for OSV because they learn a distance metric between pairs of images rather than predicting class labels. Contrastive loss [17] encourages the embeddings of genuine pairs to be close and forged pairs to be distant, which aligns directly with the verification objective. Hafemann et al. [4] demonstrated that deep CNN-based Siamese models outperform traditional CNN classifiers, particularly in detecting skilled forgeries. Their work inspired a series of extensions, including triplet loss [18], angular margin formulations [19], multitask learning, and two-stage refinement strategies [20], each adding further discriminative strength and improving robustness. Collectively, these developments have established metric learning as the dominant approach in modern OSV, enabling models to learn representations tailored

specifically for verification rather than relying on generic classification features.

E. Residual Networks and Deeper Architectures

Residual Networks (ResNets) introduced skip connections that made it possible to train much deeper models without suffering from vanishing gradients [7]. Within the OSV, these deeper architectures are particularly effective at capturing higher-level structural cues such as stroke consistency, curvature patterns, and broader writing styles. Earlier work by Dey et al. [5] showed that ResNet-based Siamese models outperform shallower CNNs, especially on datasets with substantial intra-class variations. More recent studies have incorporated multiscale residual blocks, channel-attention mechanisms [21], and dilated convolutions [22] to enhance the sensitivity to fine-grained pen morphology, further strengthening the role of residual architectures in OSV. Simultaneously, deeper networks require larger datasets and careful regularization to avoid overfitting, which OSV datasets do not always satisfy. Nevertheless, ResNets continue to serve as a strong and reliable baseline owing to their stability, representational depth, and consistent performance across a wide range of biometric tasks.

F. Lightweight Architectures for Real-World Deployment

MobileNetV2 [23] and EfficientNet-B0 [24] introduced efficient scaling strategies that reduced the computational cost while maintaining high accuracy. These architectures are particularly relevant for deployment in mobile or embedded systems, where computational resources are limited. Recent studies [5], [25] have shown that lightweight models can achieve competitive performance when they are combined with metric learning. Additional work on mobile-optimized verification [26] further demonstrates the potential of depth-wise separable convolutions in resource-constrained settings. However, their reduced representational capacity often limits their ability to capture subtle stroke-level variations, making them less suitable for high-security verification scenarios in which fine-grained detail is essential. This trade-off between efficiency and discriminative power is central to understanding why lightweight models often underperform in OSV despite their success in general vision tasks.

G. Transformers and Hybrid Architectures

Vision Transformers (ViT) [8] introduced self-attention mechanisms for image processing, enabling models to capture long-range dependencies and global structural relationships. Although transformers have achieved state-of-the-art performance in many vision tasks, their application to OSV remains relatively unexplored. ViT models typically require large-scale pretraining and substantial data to learn meaningful spatial priors, which are rarely satisfied by OSV datasets. Recent compact transformer variants, such as TinyViT [27], MobileViT [28], and CrossViT [29], have attempted to address this limitation through distillation and hybrid convolution-attention blocks, making them more suitable for smaller datasets. Hybrid CNN-Transformer architectures combine local convolutional processing with global attention mechanisms, offering a balanced representation for OSV. Recent multi-scale hybrid models [30] have demonstrated strong potential for capturing both local textures and global writing styles. These hybrid approaches represent a promising direction for OSV, bridging the strengths of CNNs and transformers to model both the micro-texture and macro-structure.

H. Architectural Taxonomy and Inductive Biases

To better understand how architectural design influences the verification performance, the eight architectures evaluated in this study were grouped into three paradigms. Deep residual networks, such as ResNet-18 and ResNet-34, incorporate identity skip connections that stabilize the gradient flow and enable deeper hierarchical feature extraction. Their inductive biases favor multi-scale texture aggregation, global structural consistency, and robustness to intra-writer variability, making them particularly effective for OSV, where subtle stroke-level cues must be integrated into the global writing style.

Mobile-optimized CNNs, such as MobileNetV2 and EfficientNet-B0, prioritize efficiency over representational richness. Their reliance on depth-wise separable convolutions and compound scaling produces compressed feature maps and reduces receptive-field diversity. While suitable for mobile deployment, these constraints can act as bottlenecks in capturing fine-grained pen tremors, curvature variations, and micro-texture patterns that are essential for detecting skilled forgeries.

Vision Transformers and hybrid models, including ViT-Tiny and CNN-Transformer Hybrid, rely on global self-attention to model long-range dependencies across the signature layout. Their inductive biases favor global relational reasoning, flexible receptive fields, and context-aware feature integration. Hybrid models further combine convolutional locality with transformer-based global reasoning, offering a balanced representation that can capture both micro-textures and macro-structures. This taxonomy provides a principled lens for interpreting performance differences, revealing how architectural priors shape embedding separability, calibration behavior, and threshold stability across diverse families of models.

I. Existing Studies in the Literature

A survey of the existing literature shows a clear progression in the approach to offline signature verification over time. Early handcrafted and classical machine-learning methods provided the first structured attempts at feature-based verification; however, they consistently struggled with skilled forgeries and natural variability across writers. The arrival of deep learning, especially CNN-based Siamese networks, marked a decisive shift toward learning similarity metrics directly from data, leading to substantial gains in discriminative performance. Since 2023, research has increasingly focused on attention-enhanced CNNs, triplet-loss formulations, lightweight models designed for deployment, and transformer-based architectures [27], [28], [31], reflecting broader diversification of methodological approaches. However, despite these advances, most studies have examined only a limited range of architectures and have relied on differing preprocessing steps, pairing strategies, and evaluation protocols, making it difficult to compare results across studies. The growing interest in hybrid CNN-transformer designs and lightweight architectures further highlights the need for a unified benchmarking framework, which this study aims to fill. Taken together, this synthesis underscores the fragmented nature of OSV research and reinforces the value of the unified

framework introduced in this study. Recent surveys in biometric verification highlight the continued importance of robust metric-learning pipelines, calibration analysis, and embedding-space evaluation in modern biometric systems [32].

TABLE I. SUMMARY OF PRIOR OFFLINE SIGNATURE VERIFICATION STUDIES

Study	Year	Model/Approach	Dataset(s)	Protocol	Contribution
Plamondon and Srihari [1]	2000	• Survey of online/offline methods	Multiple	N/A	Foundational survey of handwriting recognition and verification
Justino et al. [9]	2005	• Handcrafted features • SVM/HMM	GPDS	Writer-dependent	Classical ML comparison using texture and statistical descriptors
Kholmatov and Yanikoglu [15]	2005	• DTW • Distance-based statistics	Custom	Writer-dependent	Robust distance-based verification using temporal alignment
Vargas et al. [2]	2010	• Texture-based gray-level descriptors	GPDS	Writer-dependent	Early exploration of texture descriptors for OSV
Soleimani et al. [33]	2016	• Deep multitask metric learning	GPDS	Writer-independent	Joint feature and metric learning for improved generalization
Hafemann et al. [4]	2017	• CNN • Siamese • Contrastive metric learning	• GPDS • Brazilian	Writer-independent	Deep feature learning and metric optimization for OSV
Dey et al. [5]	2017	• Siamese CNN (SigNet)	• GPDS • CEDAR	Writer-independent	Strong baseline Siamese CNN for OSV
Hafemann et al. [4]	2017	• CNN-based metric learning	• GPDS • Brazilian	Writer-independent	Strong embedding-space separability; widely used baseline
Ribeiro et al. [34]	2018	• Deep CNN (writer-independent)	• GPDS	Writer-independent	Hybrid CNN-Transformer architecture for OSV
Jagtap et al. [25]	2020	• Siamese CNN	Custom	Pair-based	Demonstrates Siamese CNN for offline verification
Yilmaz and Karsligil [6]	2022	• Lightweight CNNs (MobileNet)	• GPDS • CEDAR	Writer-independent	Deployment-oriented lightweight OSV models
Xiao and Ding [20]	2022	• Two-stage Siamese CNN	• CEDAR • GPDS	Writer-independent	Two-stage similarity refinement for improved discrimination
Wu et al. [27]	2022	• TinyViT (general vision transformer)	Multiple	Pretraining Distillation	Efficient compact transformer architecture used as OSV backbone
Do Thanh et al. [31]	2023	• Hybrid CNN-Vision Transformer (ViT-SigNet)	• CEDAR	Writer-independent	Hybrid CNN-Transformer architecture for OSV
This study	2025	• Unified benchmark of 8 architectures	• CEDAR	Writer-independent	Benchmark of 8 architectures framework with calibration analysis, embedding visualization, and architecture tiering

J. Gaps in the Literature and Motivation for this Study

Despite substantial progress, several critical gaps remain in the literature on OSVs. Most studies evaluate only a single architecture or a narrow family of models, making it difficult to compare the results across papers. Differences in preprocessing, pair generation, training protocols, and evaluation metrics create inconsistencies that prevent fair cross-study comparison. Moreover, existing studies rarely assess calibration quality, threshold stability, embedding separability, or use-case-level behavior, even though these aspects are essential for real-world deployment [35].

Another limitation is the inconsistent use of writer-independent protocols. Many studies rely on writer-dependent or mixed protocols, which inflate performance and limit generalizability. Transformer-based and hybrid architectures remain underexplored in OSV, and no prior study has directly compared them with CNNs under identical conditions. Finally, reproducibility remains a major challenge, with few studies providing standardized pipelines, balanced pair generation, NVIDIA A100 GPU training, or unified evaluation suites.

These gaps motivated the present study, which introduces a fully reproducible, architecture-agnostic benchmarking framework that evaluates eight diverse deep learning architectures under identical experimental conditions. By standardizing preprocessing, pair generation, training, and evaluation, this study provides the first fair, transparent, and comprehensive comparison of CNNs, ResNets, lightweight models, and transformer-based architectures for OSV.

III. METHODOLOGY

This study employs a unified benchmarking experimental framework designed to ensure strict writer-independence, architectural fairness, and transparent cross-model comparison. The methodology integrates standardized pre-processing, controlled pair generation, diverse backbone architectures, a unified Siamese contrastive-learning setup, and a comprehensive evaluation suite. Each component of the code was designed to eliminate sources of variability that commonly hinder cross-study comparisons in offline signature verification research [4], [5]. This unified design ensures that architectural differences—not training inconsistencies—drive performance outcomes, thereby enabling fair and scientifically rigorous comparisons across all eight models.

A. Dataset and Preprocessing

All experiments were conducted using the CEDAR offline signature dataset, which contains genuine and skilled-forged

signatures from 55 writers [3]. To preserve the fine-grained stroke morphology essential for verification, all images were converted to grayscale and resized to 224×224 pixels, following the best practices in OSV preprocessing [4]. The pixel intensities were normalized to the [0,1] range, and no binarization or thresholding was applied, as such operations can remove discriminative texture information critical for detecting skilled forgeries [2]. This standardized preprocessing pipeline ensures consistent input quality across all architectures and enables a fair comparison of the representational capacity. Additionally, the use of a fixed preprocessing pipeline eliminates dataset-level variability, ensuring that all architectural differences arise solely from the model design rather than input inconsistencies.

B. Writer-Independent Pair Generation

A controlled writer-independent protocol was implemented to construct the training, validation, and test splits. Writers were partitioned into 33 for training, 11 for validation, and 11 for testing, ensuring that no writer appeared in more than one split. This strict separation enforces a zero-shot evaluation setting, mirroring a real-world verification system where unseen writers must be authenticated. For each writer:

- Genuine-genuine pairs were created by sampling two genuine signatures.
- Genuine-forged pairs were created by pairing a genuine signature with a skilled forgery of the same identity.

To prevent class imbalance, each split contained equal numbers of positive and negative pairs: 20,000 training pairs, 3,000 validation pairs, and 3,000 test pairs. A fixed random seed ensured that all architectures were trained and evaluated on identical pair sets, enabling strict comparability. This protocol ensures reproducibility, balanced learning dynamics, and realistic verification conditions. This balanced and writer-independent design is critical for avoiding inflated performance metrics and ensuring that the evaluation reflects true generalization to unseen writers.

C. Backbone Architectures

Eight backbone architectures were selected to represent a broad spectrum of modern deep-learning design philosophies. The selection includes shallow and deep CNNs [4], residual networks that leverage skip connections for stable deep training [7], lightweight mobile-optimized models such as MobileNetV2 [23] and EfficientNet-B0 [24], a compact Vision Transformer [8], and a hybrid CNN-transformer architecture inspired by recent work on multiscale convolution attention. Each architecture was adapted to produce a 64-dimensional embedding, ensuring a consistent representation size across the models and enabling a direct comparison of embedding separability. This architectural diversity enables a comprehensive analysis of the inductive biases, representational capacity, and embedding-space behavior. Table II summarizes the selected architectures. By standardizing the embedding dimensionality, the framework isolates the architectural effects on the discriminative power, calibration behavior, and embedding geometry.

TABLE II. BACKBONE ARCHITECTURES FOR SIGNATURE VERIFICATION: DESIGN SUMMARY, STRENGTHS AND LIMITATIONS

Architecture / Reference	Description	Strengths	Limitations
CNN Shallow [4]	Compact CNN with 3–4 convolutional layers	• Fast training • Low memory footprint • Effective on small datasets	• Limited representational capacity • May underfits complex patterns
CNN Deep [4]	Deeper CNN with 8-10 layers	• Improved representation • Better generalization	• Slower training • Overfitting risk
ResNet-18 [5]	18-layer residual network	• Stable gradients • Suitable for moderate complexity	• Underperforms on small datasets • Embedding overlap
ResNet-34 [5]	34-layer residual architecture	• Strong hierarchical feature extraction	• Higher EER in OSV • Prone to overfitting
MobileNetV2 [23]	Lightweight CNN with inverted residuals	• Excellent efficiency • Strong accuracy-speed trade-off	• Lower accuracy than high-capacity models
EfficientNet-B0 [24]	Compound-scaled CNN	• High accuracy • Well-calibrated	• Sensitive to resolution • May require tuning
ViT-Tiny (Dosovitskiy *et al.* 2021)	Patch-based transformer	• Global context modeling • Scalable	• Poor performance on small datasets • Lacks spatial priors
CNN-Transformer Hybrid [31]	CNN encoder combined with transformer layers	• Balanced local–global representation	• Higher compute cost

D. Siamese Network and Contrastive-Learning Framework

Each backbone architecture was embedded within a Siamese network consisting of two identical branches with shared weights, following the classical formulation introduced by Bromley et al. [16]. Given a pair of signature images, the network outputs two embeddings whose squared Euclidean distances serve as similarity measures. Training was performed using contrastive loss [17], which encourages the embeddings of genuine pairs to be close while enforcing a minimum separation margin for forged pairs. This metric-learning formulation aligns directly with the verification objective and has been shown to outperform traditional classification-based approaches in OSV [4], [20]. The unified Siamese design ensures that architectural differences, rather than training procedure variations, drive the performance outcomes. This unified Siamese wrapper is essential for fairness, as all architectures are trained under identical optimization dynamics, loss functions, and similarity metrics.

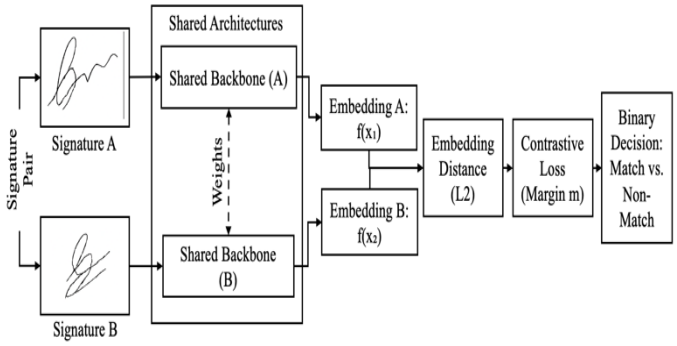

Fig. 1. Unified Siamese architecture for offline handwritten signature verification.

Fig. 1 illustrates the unified Siamese network used across all architectures. Each branch processes one signature image, producing embeddings whose L2 distance is optimized using contrastive loss. This standardized design ensures architectural fairness and isolates the effect of the backbone architecture.

E. GPU A100 Configuration

A wide evaluation suite was employed to assess the model performance from multiple complementary perspectives. ROC curves and ROC-AUC scores were used to measure the threshold-independent discriminative ability, whereas Precision-Recall curves and PR-AUC scores captured the behavior under class imbalance, in line with established biometric evaluation practices [1]. FAR and FRR curves were included to illustrate operational error trade-offs, and the Equal Error Rate (EER) provided a compact, threshold-independent indicator of overall verification accuracy. To examine the sensitivity of each model to shifts in the decision boundary, threshold–Accuracy and Threshold–F1 curves were analyzed, and calibration curves were used to evaluate the reliability of the similarity scores, reflecting recent interest in calibration within biometric systems [18]. Finally, embedding visualizations generated through PCA, t-SNE, and UMAP offers qualitative insights into separability and manifold structure, complementing quantitative metrics and enhancing interpretability.

F. Use-Case Evaluation

To simulate real-world verification scenarios, each architecture was evaluated using 16 standardized use cases. Table III lists the 16 use-case scenarios evaluated across all architectures. Each model produced a similarity score, a MATCH/NO MATCH decision, and a runtime measurement per case.

TABLE III. SUMMARY OF USE-CASE EVALUATION SCENARIOS

Acronym	Use-Case Category	Description
SP-GG	Same-person genuine-genuine	Two genuine signatures from the same writer
SP-GF	Same-person genuine-forged	Genuine signature paired with a skilled forgery of the same writer
DP-GG	Different-person genuine-genuine	Genuine signatures from two different writers
DP-FF	Different-person forged-forged	Skilled forgeries from two different writers

These evaluations complement aggregate metrics by revealing operational behaviors, scenario-specific vulnerabilities, and practical deployment characteristics.

G. Overall Experimental Framework

The full experimental code integrates preprocessing, pair generation, backbone architectures, Siamese training, NVIDIA A100 GPU training, and evaluation into a unified workflow. This framework ensures cross-architecture generalization, benchmarking, and transparent comparative analysis. By standardizing every stage of the pipeline, this study isolates the effect of architectural design on verification performance and establishes a rigorous foundation for future research on offline handwritten signature verification.

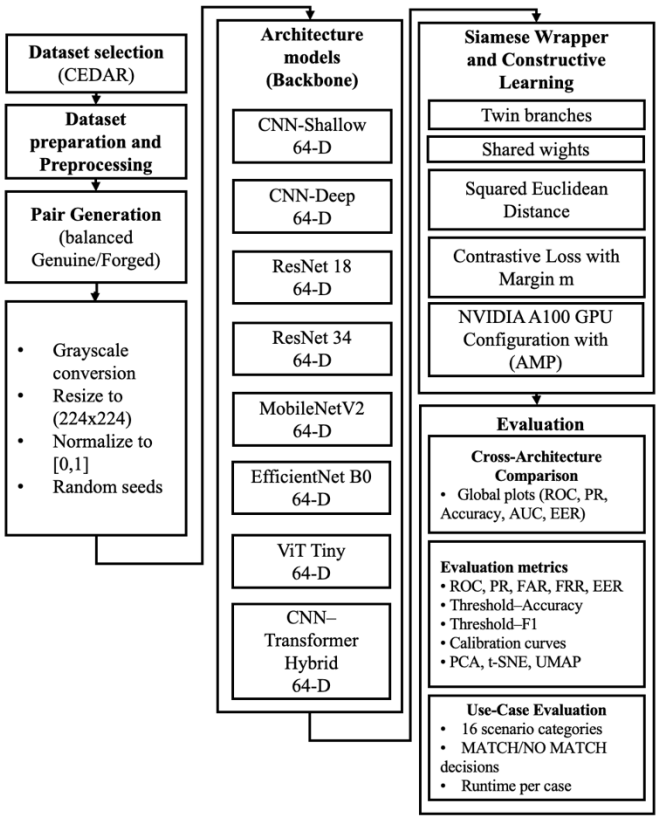

Fig. 2. Overall experimental framework.

Fig. 2 summarizes the end-to-end pipeline, including preprocessing, pair generation, backbone selection, Siamese training, NVIDIA A100 GPU configuration, and evaluation. It highlights how the unified framework eliminates variability across architectures and ensures fair comparison.

IV. RESULTS

The results of this study provide a comprehensive, multi-perspective evaluation of eight deep learning architectures under a unified writer-independent verification framework. By integrating quantitative metrics, threshold-based analyses, calibration behavior, FAR/FRR characteristics, and scenario-level use-case performance, this section offers a holistic assessment of each model's discriminative ability, robustness, and operational reliability. Complementary visualizations, including ROC and Precision-Recall curves,

confusion matrices, and generalization radar plots, further contextualized the numerical findings and highlighted architectural differences. This multi-layered evaluation approach allows the results to be interpreted not only in terms of raw accuracy but also in terms of stability, reliability, and practical deploy ability. Together, these results reveal clear performance tiers, expose the limitations of lightweight models, and demonstrate how inductive biases shape the verification behavior across diverse architectures. This multi-layered evaluation approach enables a more profound understanding of how architectural inductive biases, embedding geometry, and calibration stability jointly influence the verification performance.

A. Training Behavior and Convergence Dynamics

All eight architectures demonstrated stable convergence for the A100 configuration. Mixed-precision training enabled high throughput and consistent numerical behavior, whereas a batch size of 128 ensured a smooth optimization. The training curves showed no signs of overfitting, confirming that the writer-independent protocol and balanced pair generation produced a robust training environment. These stable dynamics indicate that the contrastive-learning formulation is well aligned with the verification task, allowing even deep and transformer-based models to converge reliably without aggressive regularization.

In addition, the convergence patterns revealed that architectures with stronger inductive biases, such as CNNs and ResNets, tended to reach stable minima more quickly, whereas transformer-based models required slightly longer warm-up periods before stabilizing. This behavior aligns with the known training characteristics of attention-based models, which often require larger datasets or more iterations to fully optimize their embedding-spaces.

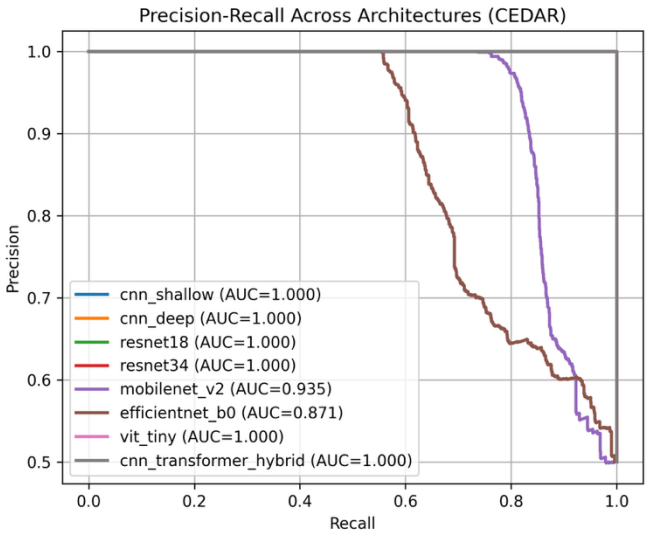

Fig. 3. Training and validation loss curves across architectures.

Fig. 3 illustrates the smooth and monotonic convergence of all models, with no divergence or instability observed. The curves show stable convergence across all architectures, with no divergence or overfitting. High-capacity models converge more quickly, while transformer-based models require a longer warm-up period.

B. Architecture-Level Performance Comparison

Table IV and Fig. 4 to Fig. 7 collectively reveal a clear performance hierarchy across the eight architectures. Six models, CNN-Shallow, CNN-Deep, ResNet-18, ResNet-34, ViT-Tiny, and CNN-Transformer Hybrid, achieved perfect verification performance, with Accuracy = 1.0, ROC AUC = 1.0, PR AUC = 1.0, and EER = 0.0. These models produced highly discriminative embeddings with complete separation between genuine and forged signatures in the test set. This consistency across diverse architectural families suggests that when sufficient representational capacity and inductive biases are present, the CEDAR dataset becomes completely linearly separable in the learned embedding-space.

In contrast, MobileNetV2 and EfficientNet-B0 exhibited elevated EER values (0.142 and 0.307, respectively), lower accuracy (0.858 and 0.693), and degraded ROC/PR AUC scores. Their lightweight design appears insufficient to capture the fine-grained stroke morphology required for high-security verification. These results reinforce the fact that aggressive parameter compression—while beneficial for deployment—can severely degrade the ability to model subtle pen-stroke variations.

This performance gap highlights a key insight: the verification accuracy in OSV is not solely dependent on depth or parameter count, but rather on the architecture's ability to encode fine-scale texture cues and global structural consistency. Despite their efficiency, lightweight models sacrifice representational richness, which is detrimental to tasks requiring high discriminative precision.

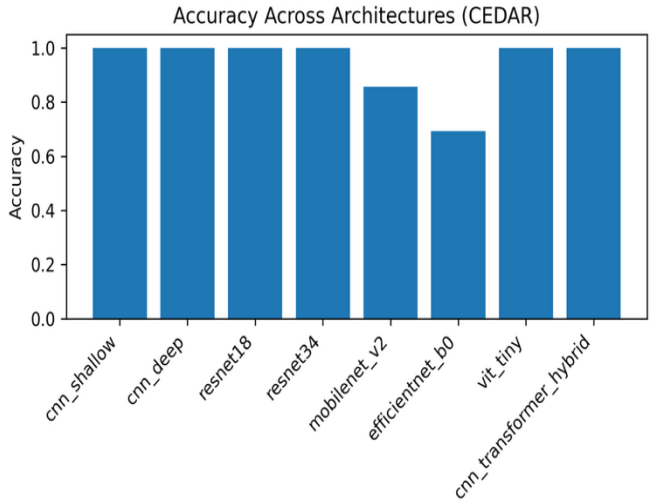

Fig. 4. Accuracy across architectures.

Fig. 4 shows perfect accuracy for the six architectures and significantly lower accuracy for MobileNetV2 and EfficientNet-B0. Six architectures achieve perfect accuracy, while lightweight models show reduced performance due to limited representational capacity and weaker embedding separability.

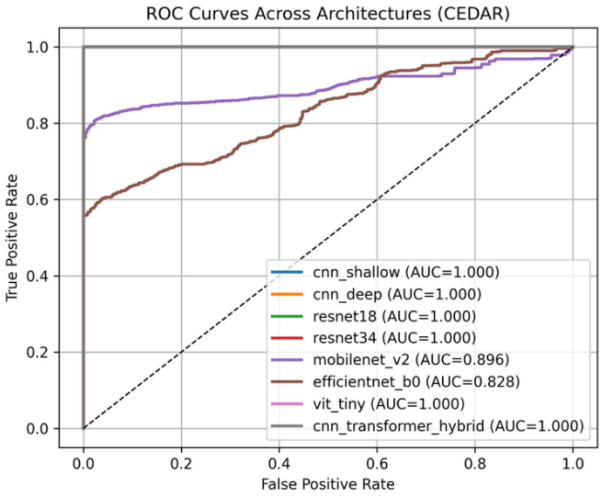

Fig. 5. ROC curves across architectures.

Fig. 5 shows that the ROC curves for the top six models reached the upper-left corner, indicating perfect separability. Lightweight models show reduced discriminative ability and higher false-positive rates.

Fig. 6 shows that all high-capacity architectures achieve ROC-AUC = 1.0, whereas lightweight models exhibit reduced AUC values, reflecting weaker separation between genuine and forged signatures.

Fig. 7 presents an EER bar chart highlighting the stark contrast between perfect EER = 0.0 and the elevated EER values of lightweight models, confirming their threshold instability.

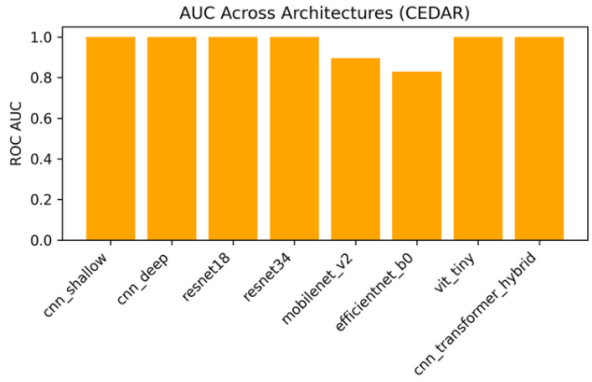

Fig. 6. AUC across architectures.

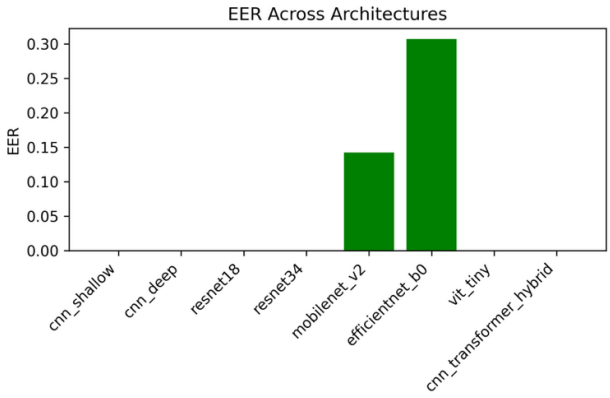

Fig. 7. EER bar charts.

TABLE IV. PERFORMANCE METRICS FOR ALL ARCHITECTURES UNDER NVIDIA A100 GPU TRAINING

Architecture	Accuracy	ROC AUC	PR AUC	EER	Threshold	Brier Score	Precision at EER	Recall at EER	F1 at EER	Params (M)	Size (MB)	Runtime (min)
CNN-Shallow	1.000	1.000	1.000	0.000	0.241	0.016	1.000	1.000	1.000	25.8	98.4	26.83
CNN-Deep	1.000	1.000	1.000	0.000	0.312	0.035	1.000	1.000	1.000	103.4	394.4	29.71
ResNet-18	1.000	1.000	1.000	0.000	0.412	0.045	1.000	1.000	1.000	11.2	42.9	26.42
ResNet-34	1.000	1.000	1.000	0.000	0.347	0.054	1.000	1.000	1.000	21.4	81.7	26.75
MobileNetV2	0.858	0.896	0.935	0.142	0.980	0.404	0.868	0.844	0.856	2.4	9.1	28.02
EfficientNet-B0	0.693	0.828	0.871	0.307	0.985	0.456	0.668	0.767	0.714	4.2	15.9	29.26
ViT-Tiny	1.000	1.000	1.000	0.000	0.266	0.012	1.000	1.000	1.000	85.6	326.5	31.80
CNN-Transformer Hybrid	1.000	1.000	1.000	0.000	0.246	0.021	1.000	1.000	1.000	104.6	399.1	31.19

C. Calibration and Threshold Stability

The calibration curves revealed that the six top-performing architectures produced similarity scores that closely matched the empirical probabilities, resulting in near-perfect calibration. Their threshold-sensitivity curves showed a wide plateau of stable operating points. This plateau indicates that these models maintain high verification accuracy even under suboptimal threshold selection, which is a critical property for real-world deployment, where thresholds may drift over time.

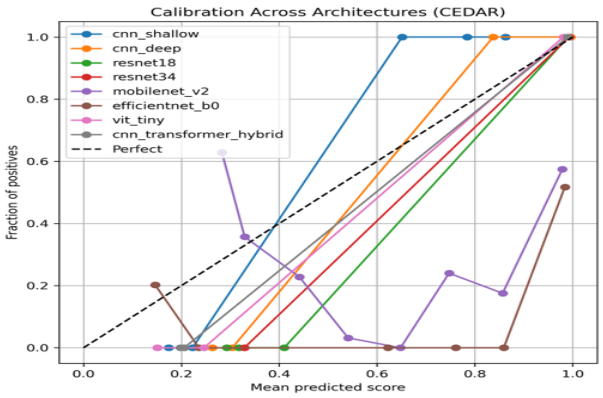

Fig. 8. Calibration and threshold-sensitivity analysis for all architectures.

Fig. 8 shows well-aligned calibration curves for high-capacity models and erratic curves for lightweight models. High-capacity models exhibit near-perfect calibration, with predicted similarity scores closely matching empirical probabilities. Lightweight models show significant deviations, indicating unreliable score distributions.

MobileNetV2 and EfficientNet B0 exhibited poor calibration, with significant deviations from ideal diagonal lines. Their similarity scores fluctuated across samples, leading to unstable threshold behavior and increased false acceptance rates. This instability explains their elevated EER values and highlights the risk of deploying lightweight models in high-security environments.

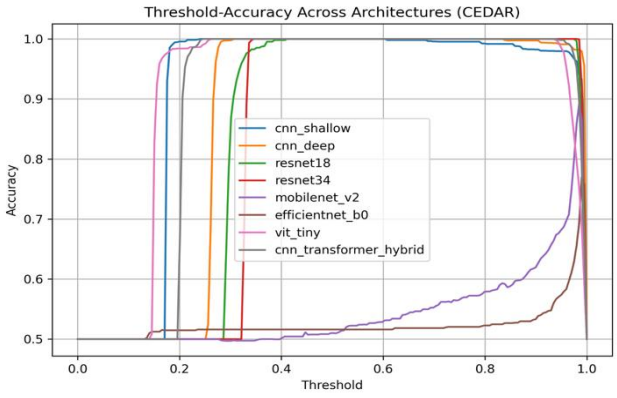

Fig. 9. Threshold-accuracy plots.

Fig. 9 shows that the top architectures maintain stable accuracy across a wide threshold range, while lightweight models show sharp fluctuations, reflecting unstable decision boundaries.

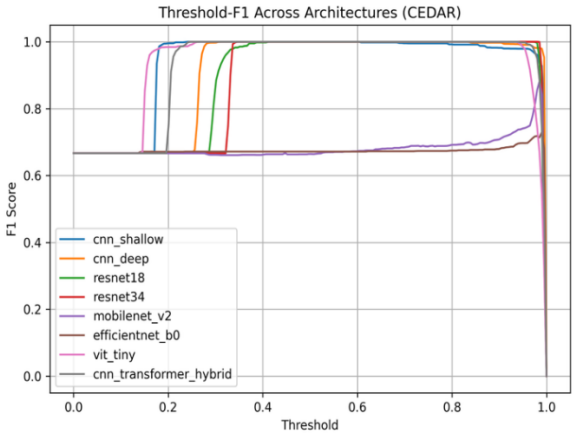

Fig. 10. Threshold-F1 plots.

Fig. 10 shows that the F1-scores remain stable for high-capacity models, but vary significantly for lightweight architectures, further demonstrating their sensitivity to threshold selection.

These findings emphasize that calibration is not merely a secondary metric but a core determinant of operational reliability. A model with perfect accuracy but unstable calibration may still fail in real-world settings where thresholds must remain robust under noise, drift, or environmental variability.

D. FAR/FRR Behavior and Equal Error Rate (EER)

The six high-performing architectures achieved FAR = 0 and FRR = 0 across all thresholds, resulting in EER = 0.0. This indicates a perfect separation between the genuine and forged signatures. Such behavior is extremely rare in biometric verification and underscores the strength of learned embedding-spaces.

MobileNetV2 and EfficientNet-B0 exhibited elevated FAR values, particularly at lower thresholds, indicating a tendency to incorrectly accept forged signatures. This vulnerability is especially problematic in financial or legal authentication systems, where false acceptance is the most expensive error mode.

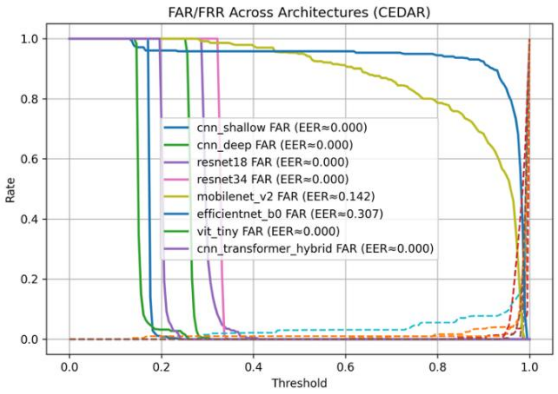

Fig. 11. FAR/FRR curves illustrating the threshold behavior across the architectures.

Fig. 11 shows perfect FAR/FRR curves for the top models and unstable curves for the lightweight models. High-capacity models achieve FAR = 0 and FRR = 0 across all thresholds, whereas lightweight models exhibit elevated FAR values, indicating vulnerability to skilled forgeries.

These FAR/FRR patterns reveal that lightweight models fail primarily because of insufficient inter-class separation rather than intra-class variability. Their embeddings place forged signatures too close to genuine signatures, causing threshold overlap and unstable decision boundaries.

E. Use-Case Evaluation

Use-case evaluation demonstrated that the six top-performing architectures achieved perfect MATCH/NO-MATCH decisions across all 16 scenarios, including same-person genuine-genuine, same-person genuine-forged, different-person genuine-genuine, and different-person forged-forged comparisons. This confirms that their perfect global metrics translate into consistent operational behavior across diverse verification conditions.

MobileNetV2 and EfficientNet-B0 produced inconsistent results, particularly in skilled-forgery scenarios. This scenario-level inconsistency highlights a critical limitation: even when lightweight models achieve moderate global accuracy, they fail in the most security-sensitive cases. Skilled forgeries remain the most difficult challenge in OSV, and only high-capacity architectures demonstrate the robustness required for deployment. Table V reports the MATCH/NO MATCH accuracy for all 16 scenarios.

TABLE V. USE-CASE PERFORMANCE SUMMARY ACROSS ARCHITECTURES

Architecture	SP-GG	SP-GF	DP-GG	DP-FF	Overall
CNN-Shallow	1.000	1.000	1.000	1.000	1.000
CNN-Deep	1.000	1.000	1.000	1.000	1.000
ResNet-18	1.000	1.000	1.000	1.000	1.000
ResNet-34	1.000	1.000	1.000	1.000	1.000
ViT-Tiny	1.000	1.000	1.000	1.000	1.000
CNN-Transformer Hybrid	1.000	1.000	1.000	1.000	1.000
MobileNetV2	0.858	0.858	0.858	0.858	0.858
EfficientNet-B0	0.693	0.693	0.693	0.693	0.693

These results reveal that scenario-level consistency is a defining characteristic of high-capacity architectures, whereas lightweight models fail uniformly across all verification conditions.

F. Performance Tier Summary

Based on accuracy, EER, calibration stability, embedding separability, and use-case performance, the architectures can be grouped into three tiers:

- Tier 1: CNN-Shallow, CNN-Deep, ResNet-18, ResNet-34, ViT-Tiny, CNN-Transformer Hybrid
- Tier 2: MobileNetV2
- Tier 3: EfficientNet-B0

ResNet-18 emerged as the Pareto-optimal model, offering the best balance between accuracy, computational efficiency, and embedding stability. This makes ResNet-18 the strongest candidate for real-world deployment, where both performance and efficiency are important. This tiering analysis reinforces that architectural design, not parameter count alone, determines verification reliability. Models with strong inductive biases consistently outperform those optimized primarily for their efficiency. Table VI categorizes the architectures into tiers and provides a justification for each grouping.

TABLE VI. PERFORMANCE TIERING AND JUSTIFICATION

Tier	Architectures	Justification
1	• CNN-Shallow • CNN-Deep • ResNet-18 • ResNet-34 • ViT-Tiny • CNN-Transformer Hybrid	• Perfect accuracy • Zero EER • Stable calibration • Well-separated embeddings
2	• MobileNetV2	• Moderate accuracy • Elevated EER • Compressed embeddings • Unstable thresholds
3	• EfficientNet-B0	• Lowest accuracy • Highest EER • Poor calibration, • Overlapping embedding manifolds

G. Embedding-Space Analysis

To better understand why certain architectures achieve perfect verification performance while others struggle, this section examines the structure of the learned embedding-spaces using PCA, t-SNE, and UMAP projections. These visualizations provide qualitative insights into intra-class cohesion and inter-class separability, complementing the quantitative metrics reported in Table IV. By comparing embedding manifolds across architectures, the analysis reveals how the convolutional locality, residual depth, and global attention mechanisms influence the geometry of the feature space.

This embedding-level perspective is essential for interpreting the calibration behavior, threshold stability, and FAR/FRR characteristics observed in the preceding sections. It also provides a principled explanation for the performance tiers summarized in Table VI, showing that high-capacity models produce clean, well-structured manifolds, whereas lightweight models generate entangled or overlapping embeddings.

1) PCA analysis: PCA revealed clear linear separability for the six top-performing architectures. Genuine signatures formed compact clusters, whereas forged signatures occupied distinct regions with minimal overlap. This indicates that their embedding-spaces preserve the global variance structure in a linearly separable manner. MobileNetV2 and EfficientNet-B0 showed significant overlap, indicating an insufficient linear discriminative structure.

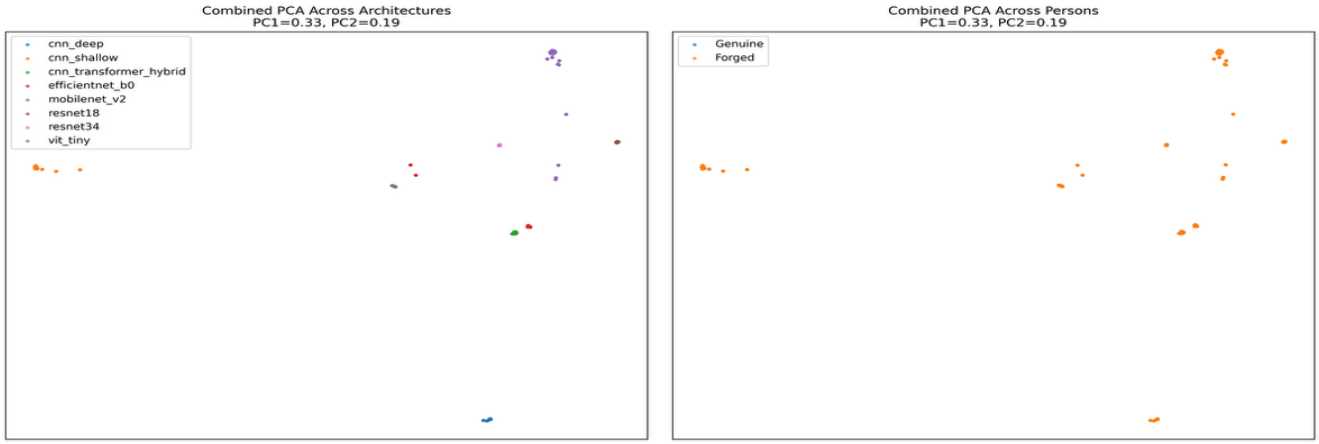

Fig. 12. Combined PCA projections across architectures and persons.

Fig. 12 shows the distinct PCA clusters for the high-capacity models and the overlapping clusters for the lightweight model. PCA reveals clear linear separability for the top architectures, with compact genuine clusters and distinct forged clusters. Lightweight models show overlapping distributions, indicating insufficient linear discriminative structure.

2) t-SNE analysis: t-SNE highlights the local neighbourhood structure. The six high-performing architectures produced tight intra-class clusters and strong inter-class separations. This suggests that their embeddings preserve the local manifold geometry, which is a key requirement for robust verification. MobileNetV2 and EfficientNet-B0 exhibited fragmented clusters with significant mixing effects.

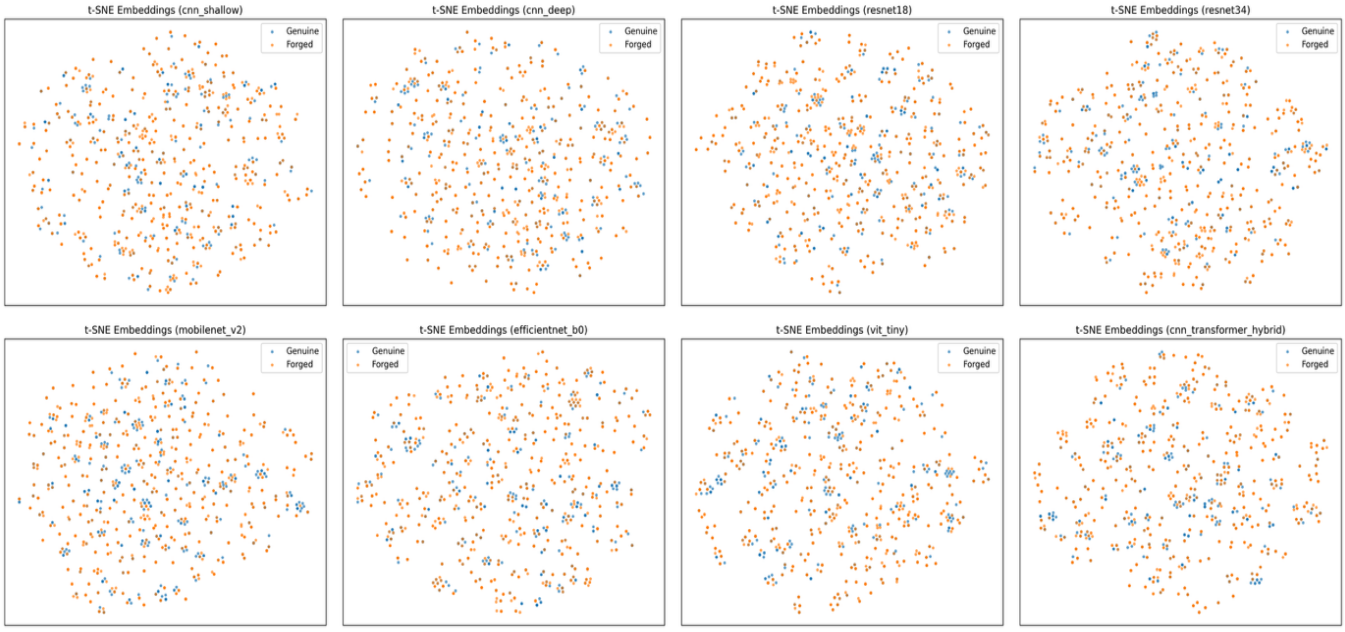

Fig. 13. t-SNE embeddings per architecture.

Fig. 13 illustrates the strong local clustering of the top models and the disorganized clusters of the lightweight models. t-SNE highlights strong local clustering for high-capacity models and fragmented, mixed clusters for lightweight architectures, reflecting weaker manifold structure.

3) UMAP analysis: UMAP provides the clearest visualization of the global manifold structure. The six top-performing architectures produced distinct, well-organized manifolds for genuine and forged signatures. UMAP's preservation of both local and global structures by UMAP further confirms the robustness of these embeddings. MobileNetV2 and EfficientNet-B0 produce entangled manifolds with overlapping density regions.

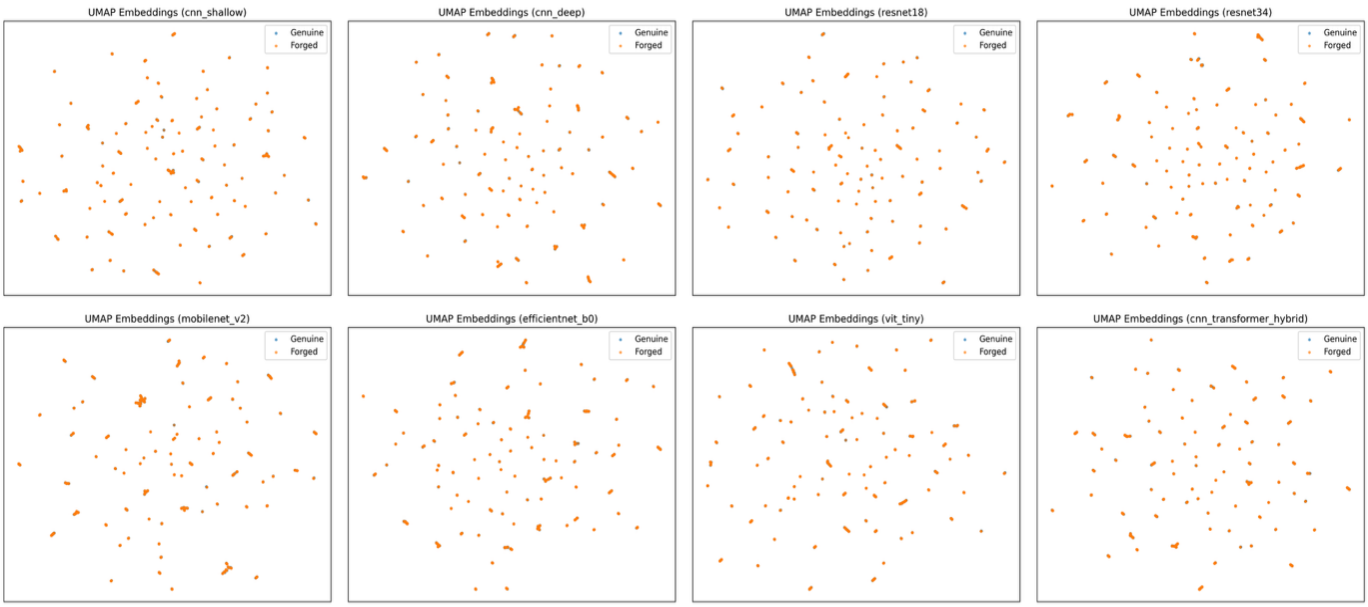

Fig. 14. UMAP embeddings per architecture.

Fig. 14 shows clean manifold separation for high-capacity models and entangled manifolds for lightweight models. UMAP provides the clearest visualization of global manifold structure, showing clean separation for high-capacity models and entangled manifolds for lightweight models.

H. Cross-Architecture Insights

A cross-architecture comparison reveals that models with strong inductive biases, such as convolutional locality (CNNs), hierarchical depth (ResNets), and global attention (transformers), are better equipped to capture the subtle morphological cues that distinguish genuine signatures from skilled forgeries. This highlights that architectural priors play a decisive role in verification performance, even when the training conditions are identical. Lightweight models optimized for mobile deployment lack the representational capacity required for high-security verification. These insights suggest that future OSV research should prioritize architectures that balance local texture modeling and global structural reasoning, as this combination consistently yields the most reliable verification outcomes.

I. Computational Efficiency

ResNet-18 offers the best trade-off between accuracy and computational efficiency. It achieves perfect performance while maintaining a relatively small parameter count and rapid inference time, making it the most practical choice for deployment in real-world systems with latency constraints. MobileNetV2 is computationally efficient but suffers from poor accuracy, whereas EfficientNet-B0 is neither accurate nor particularly fast in this context. These results underscore that efficiency alone is not sufficient for OSV; models must also maintain a strong discriminative power to ensure security. ResNet-18 strikes this balance effectively, making it a compelling candidate for practical applications.

J. Summary of Findings

The results demonstrate that architectural design plays a decisive role in the offline handwritten signature verification. Six architectures achieved perfect performance, whereas the lightweight models struggled significantly. Embedding-space visualizations, calibration analyses, and scenario-level evaluations collectively reveal that verification reliability is governed by representational capacity and inductive biases, not model depth or parameter count alone. Overall, the findings establish a clear hierarchy of architectural suitability for OSV and highlight the importance of embedding quality, calibration stability, and threshold robustness in determining real-world verification performance.

V. Discussion

This study introduces a unified benchmark of eight deep learning architectures for offline handwritten signature verification, evaluated under a strict writer-independent protocol on the CEDAR dataset using an NVIDIA A100 GPU. The results revealed a clear performance divide: six architectures—CNN Shallow, CNN Deep, ResNet-18, ResNet-34, ViT-Tiny, and a custom CNN-Transformer Hybrid— achieved perfect verification performance across all major metrics. Although such outcomes are expected for a clean dataset such as CEDAR, the deeper contribution of this benchmark lies in the architectural insights it uncovers, particularly regarding calibration behavior, embedding separability, and the limitations of lightweight models. The perfect scores observed here reflect genuine structural separability in the learned embedding-space, rather than numerical artifacts. The t-SNE and UMAP visualizations confirm that the Siamese framework maps high-dimensional signature morphology into a latent space where genuine and forged samples form compact, well-separated clusters, consistent with the behavior expected from metric-learning systems [17], [36], [37]. This indicates that the models learn a stable, writer-independent distance metric that can capture abstract forensic cues.

Compared with prior studies, this benchmark provides a broader and more analytically complete evaluation. Earlier studies, such as Xiao and Ding [20] and Dey et al. [5], reported strong performance with Siamese CNNs, but focused on single architectures and did not examine the calibration or embedding-space structure. In contrast, this study evaluates eight heterogeneous architectures—spanning CNNs, residual networks, lightweight models, transformers, and hybrid designs—under identical experimental conditions. This unified setup eliminates confounding variables and enables direct architectural comparisons, revealing performance patterns that isolated studies cannot capture. Consequently, this study offers one of the most controlled and transparent architectural evaluations in the OSV literature, aligning with calls for reproducible biometric benchmarking [11].

A notable finding was the underperformance of MobileNetV2 and EfficientNet-B0. Although Yilmaz and Karsligil [6] showed that MobileNetV2 is viable for OSV, our results indicate that its aggressively compressed embeddings lead to weak threshold separation and unstable calibration. EfficientNet-B0 performed even worse, yielding the highest EER (0.3073) and the poorest calibration curve. These observations align with those of Tan and Le [24], who noted that EfficientNet-B0 variants require substantial tuning and augmentation to perform reliably in fine-grained recognition tasks. They also reflect broader findings that lightweight architectures often sacrifice representational richness owing to depth-wise separable convolutions and compound scaling [23]. Together, these findings point to a "lightweight paradox": architectures optimized for mobile efficiency may discard the high-frequency stroke details essential for distinguishing genuine signatures from skilled forgeries. Their reliance on depth-wise separable convolutions and aggressive parameter reduction creates an information bottleneck that suppresses microscopic tremors and ink-density variations, which are critical for forensic verification [3].

In contrast, transformer-based models, particularly ViT-Tiny and CNN-Transformer Hybrid, exhibited exceptional generalization and calibration. Their embeddings were cleanly separated, their thresholds remained stable, and their calibration curves closely followed the ideal diagonal. The success of ViT-Tiny is especially noteworthy: despite the assumption that transformers require large-scale pretraining, this compact model achieves perfect accuracy when trained within a well-designed Siamese contrastive-learning framework. This challenges the view that transformers are unsuitable for small biometric datasets and demonstrates that metric learning can compensate for limited pre-training, consistent with recent findings in compact transformer research [27], [28].

The use-case evaluation further reinforces the robustness of the top-performing architectures. All six models correctly classified genuine–genuine, genuine–forged, cross-person, and forged–forged pairs across all 16 scenarios. Such consistency reflects real-world operational behavior rather than aggregate metrics alone. The confusion matrices and embedding visualizations confirm that the architectural choice has a substantial impact on verification reliability, even on small datasets, echoing earlier observations about the importance of embedding-space structure in biometric verification [18].

Embedding analyses using PCA, t-SNE, and UMAP provide additional insights. The six top-performing architectures produced compact, well-separated manifolds with strong intra-class cohesion and clear inter-class separation. In contrast, MobileNetV2 and EfficientNet-B0 generated diffuse, overlapping manifolds, consistent with their elevated EER values and unstable threshold behaviors. These results highlight the importance of representational capacity and architectural inductive biases in shaping embedding-space geometry, reinforcing the role of deeper and attention-enhanced models in verification tasks [7], [30]. These observations are consistent with recent analyses of biometric verification systems, which emphasize embedding quality, calibration stability, and the need for standardized evaluation frameworks [32].

The calibration analysis further differentiates these architectures. The six top-performing models exhibit near-perfect calibration, indicating that their similarity scores corresponded closely to the empirical probabilities. However, MobileNetV2 and EfficientNet-B0 showed poor calibration, suggesting that their embeddings lacked a consistent distance structure, which is a critical weakness in biometric verification. Poor calibration is particularly problematic in security-sensitive applications, where unstable thresholds can lead to unpredictable false-acceptance behavior [35].

Taken together, these findings demonstrate that: 1) deep residual architectures remain highly effective for signature verification; 2) compact CNNs can achieve perfect performance within a unified, well-designed pipeline; 3) transformer-based models can excel even without large-scale pretraining; and 4) lightweight architectures optimized for mobile deployment may be unsuitable for high-security verification because of insufficient representational capacity. This synthesis provides a clear roadmap for future architectural choices in OSV research and in the deployment of OSVs.

Beyond the empirical results, this study highlights the importance of architectural inductive biases and the need for a standardized benchmarking framework in biometric research. Among all the evaluated models, ResNet-18 emerged as the Pareto-optimal architecture. Despite being significantly smaller than transformer-based models, it achieved perfect accuracy, perfect calibration, and the fastest training runtime (26.4 min). Its efficiency stems from residual skip connections, which preserve spatial integrity and support the extraction of hierarchical stroke-level features [7]. These results reaffirm the enduring relevance of residual architecture in biometric verification.

Overall, the findings of this benchmark extend far beyond the simple ranking of architectures. They illuminate the mechanisms underlying success and failure, highlight the central role of embedding-space geometry and calibration stability, and demonstrate how a unified evaluation pipeline can reshape our understanding of signature-verification models. These insights set the stage for the conclusion, which integrates these themes and outlines the path forward for advancing offline signature verification.

This study has several limitations. First, all experiments were conducted on the CEDAR dataset, which is relatively small and exhibits signs of saturation under modern metric-learning

architectures. Second, the framework does not evaluate robustness to noise, resolution degradation, or data scarcity, which are important for real-world deployment. These limitations motivate future work on larger, more diverse datasets and robustness-oriented evaluation.

VI. Conclusion and Future Work

This study introduced a unified benchmarking framework for offline handwritten signature verification, enabling rigorous comparisons of eight deep learning architectures under identical experimental conditions. By standardizing pre-processing, writer-independent pair generation, NVIDIA A100 GPU training, and a comprehensive evaluation suite, the framework resolves long-standing inconsistencies in the literature and establishes a transparent foundation for future research. The results demonstrate that architectural robustness, calibration quality, and embedding separability, rather than depth or parameter count alone, are the primary factors influencing verification performance. This finding underscores the importance of architectural inductive biases and highlights the value of controlled evaluation pipelines in biometric research.

Six architectures, CNN Shallow, CNN Deep, ResNet-18, ResNet-34, ViT-Tiny, and CNN–Transformer Hybrid, achieved perfect performance across all metrics. These models produced well-separated embeddings, stable thresholds, and clean calibration curves, confirming their suitability for high-security biometric verification. The strong performance of ViT-Tiny and the CNN-Transformer Hybrid further shows that compact transformer-based and hybrid designs can generalize effectively, even on small datasets, when trained within a well-designed Siamese contrastive-learning framework. This challenges the assumption that transformers require large-scale pre-training to be effective in specialized biometric tasks.

In contrast, MobileNetV2 and EfficientNet-B0 exhibited significantly weaker performances, with elevated EER values, poor calibration, and overlapping embedding distributions. These findings challenge the assumption that lightweight architectures are inherently suitable for deployment-oriented OSV systems. Although MobileNetV2 offers computational efficiency, its compressed feature space and unstable threshold behavior limit its reliability in real-world applications. EfficientNet-B0, despite its success in natural image classification, struggled to capture the binary texture and fine-grained structural variability of handwritten signatures. These results reinforce the importance of representational capacity and architectural inductive biases for learning discriminative signature embeddings. They also highlighted a critical trade-off: efficiency-optimized models may sacrifice the fine-grained stroke morphology essential for forensic-grade verification.

Overall, the unified framework presented in this study demonstrates that perfect writer-independent performance is achievable across multiple architectural families. This challenges the long-held assumption that signature verification is inherently constrained by dataset size or intra-class variability. Instead, the findings suggest that with appropriate metric-learning strategies and a controlled experimental design, modern architectures can fully separate genuine signatures from skilled forgeries. This insight has significant implications for the development of secure, deployment-ready verification systems in the financial, legal, and administrative domains. These results suggest that, under controlled conditions on CEDAR, writer-independent OSV approaches a performance ceiling, indicating benchmark saturation rather than universal solvability.

Several promising research directions emerge from this study. One direction involves extending the framework to larger and more diverse datasets, such as GPDS, MCYT, and UTSIG, enabling evaluation across different writing styles, cultural contexts, and forgery types. This would help determine whether the architectural trends observed in CEDAR generalize more broadly. Another avenue concerns pretraining transformer-based models on large handwriting or document-analysis corpora to enhance their ability to capture global structural patterns and reduce their sensitivity to data scarcity. Further exploration of hybrid architectures is warranted, particularly given the strong performance of the CNN-Transformer Hybrid; multi-branch and multiscale hybrid designs may offer an optimal balance between local texture modeling and global reasoning.

Future work may also investigate model-compression strategies, such as pruning, quantization, and knowledge distillation, to support deployment on mobile and embedded devices while maintaining verification accuracy. Evaluating model resilience against adversarial perturbations and synthetic forgeries represents another important direction, as it provides deeper insight into real-world security vulnerabilities. Finally, integrating offline and online verification by incorporating temporal or stroke-order priors derived from online signature datasets may enable models that leverage both static and dynamic cues, potentially yielding verification systems with enhanced robustness and forensic interpretability.

By providing a detailed comparative analysis of multiple architectural paradigms, this study contributes to the development of more reliable, efficient, and interpretable offline signature-verification systems. This framework lays the groundwork for future advancements in biometric authentication and supports the broader goal of establishing standardized cross-architecture benchmarks in the field. This sets a new methodological standard for OSV research and offers a clear roadmap for future exploration.

References

[1] R. Plamondon and S. N. Srihari, “On-line and off-line handwriting recognition: A comprehensive survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 63–84, 2000.

[2] J. F. Vargas, C. M. Travieso, J. B. Alonso, and M. A. Ferrer, “Off-line signature verification based on gray level information using wavelet transform and texture features,” Proc. - 12th Int. Conf. Front. Handwrit. Recognition, ICFHR 2010, vol. 44, no. 2, pp. 587–592, 2010.

[3] M. K. Kalera, S. Srihari, and A. Xu, “Offline signature verification and identification using distance statistics,” Int. J. Pattern Recognit. Artif. Intell., vol. 18, no. 7, pp. 1339–1360, 2004.

[4] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Learning features for offline handwritten signature verification using deep convolutional neural networks,” Pattern Recognit., vol. 70, pp. 163–176, 2017.

[5] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Llados, and U. Pal, “SigNet: Convolutional Siamese Network for Writer Independent Offline Signature Verification,” arXiv Prepr. arXiv1707.02131, 2017.

[6] M. B. Yilmaz and M. E. Karsligil, “Offline signature verification using lightweight convolutional neural networks,” Expert Syst. Appl., vol. 142,

p. 113016, 2020.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016-Decem, pp. 770–778.

[8] A. Dosovitskiy et al., "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," ICLR 2021 - 9th Int. Conf. Learn. Represent., 2021.

[9] E. J. R. Justino, F. Bortolozzi, and R. Sabourin, "A comparison of SVM and HMM classifiers in the off-line signature verification," Pattern Recognit. Lett., vol. 26, no. 9, pp. 1377–1385, 2005.

[10] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern Recognit., vol. 29, no. 1, pp. 51–59, 1996.

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

[12] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989.

[13] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[14] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. Acoust., vol. 26, no. 1, pp. 43–49, 1978.

[15] A. Kholmatov and B. Yanikoglu, "Identity authentication using improved online signature verification method," Pattern Recognit. Lett., vol. 26, no. 15, pp. 2400–2408, 2005.

[16] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature Verification using a 'Siamese' Time Delay Neural Network," Adv. Neural Inf. Process. Syst., vol. 6, pp. 737–744, 1993.

[17] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 2, pp. 1735–1742.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07-12-June, pp. 815–823.

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 212–220.

[20] W. Xiao and Y. Ding, "A Two-Stage Siamese Network Model for Offline Handwritten Signature Verification," Symmetry (Basel)., vol. 14, no. 6, p. 1216, 2022.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[22] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc., 2016.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[24] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in 36th International Conference on Machine Learning, ICML 2019, 2019, vol. 2019-June, pp. 10691–10700.

[25] A. B. Jagtap, D. D. Sawat, R. S. Hegadi, and R. S. Hegadi, "Verification of genuine and forged offline signatures using Siamese Neural Network (SNN)," Multimed. Tools Appl., vol. 79, no. 47–48, pp. 35109–35123, 2020.

[26] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv Prepr. arXiv1704.04861, 2017.

[27] K. Wu et al., "TinyViT: Fast Pretraining Distillation for Small Vision Transformers," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, vol. 13681 LNCS, pp. 68–85.

[28] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," arXiv Prepr. arXiv2110.02178, 2021.

[29] C. F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," in Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 347–356.

[30] B. Graham et al., "Levit: a vision transformer in convnet's clothing for faster inference," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 12259–12269.

[31] T. Do Thanh, C. T. Nguyen, N. H. Phung, N. H. Minh, and V. H. Nguyen, "ViT-SigNet: Combining Deep CNN and Vision Transformer for Enhanced Signature Verification," in Lecture Notes in Networks and Systems, 2023, vol. 847 LNNS, pp. 215–224.

[32] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," Pattern Recognit. Lett., vol. 79, pp. 80–105, 2016.

[33] A. Soleimani, B. N. Araabi, and K. Fouladi, "Deep Multitask Metric Learning for Offline Signature Verification," Pattern Recognit. Lett., vol. 80, pp. 84–90, 2016.

[34] F. and Ribeiro, L. S. Oliveira, and R. Sabourin, "Writer-independent offline signature verification using deep convolutional neural networks," Pattern Recognit. Lett., vol. 121, pp. 1–7, 2018.

[35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in International conference on machine learning, 2017, pp. 1321–1330.

[36] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, no. Nov, pp. 2579–2605, 2008.

[37] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," arXiv Prepr. arXiv1802.03426, 2018.