Scientific Reports

https://doi.org/10.1038/s41598-026-43252-x

Article in Press

A secure and explainable multimodal biometric system using trust adaptive fusion for face and fingerprint

Received: 26 December 2025

Accepted: 3 March 2026

Published online: 19 March 2026

Cite this article as: Chitrapu P., Morampudi M.K. & Kalluri H.K. A secure and explainable multimodal biometric system using trust adaptive fusion for face and fingerprint. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-43252-x

Pavani Chitrapu, Mahesh Kumar Morampudi & Hemantha Kumar Kalluri

A Secure and Explainable Multimodal Biometric System Using Trust Adaptive Fusion for Face and Fingerprint

Pavani Chitrapu[1], **Mahesh Kumar Morampudi**[1], **and Hemantha Kumar Kalluri** [1,*]

[1,*]Department of Computer Science and Engineering, School of Engineering and Applied Sciences, SRM University - AP, Neerukonda-Kuragallu Village, Mangalagiri, Andhra Pradesh, 522240, India.
[*]hemanthakumar.k@srmap.edu.in

ABSTRACT

Multimodal biometrics are able to improve the accuracy and security of authentication by integrating more than one biometric characteristic, minimizing errors, and maximizing the resistance to attacks. The primary drawback of multimodal biometric verification is the complexity of the systems that are introduced by multiple sensors, more computing, and fusion issues. Multimodal feature extraction methods are inadequate in traditional feature extraction methods as they generate modality-specific, handcrafted representations which are not robust, compatible and discriminative enough to support effective feature-level fusion. Deep learning feature extractors produce robust, discriminative, and fusion-friendly representations which are very important in multimodal biometric authentication systems to enhance accuracy and reliability. Trust and confidence are crucial in multimodal biometric authentication systems utilizing deep learning, as the models operate as black boxes, handle irreversible biometric data, and make high-impact security decisions. This motivates the development of a secure, explainable, multimodal biometric authentication framework. The proposed system is a privacy-preserving and explainable multimodal biometric solution that combines deep learning, trust-adaptive fusion, and encrypted domain matching. It utilizes MobileNet for extracting discriminative features. A Trust Adaptive Fusion (TAF) Strategy adjusts the contribution of each modality based on its quality or confidence, enhancing the robustness against the noisy inputs. The fused features are secured using the Cheon-Kim-Kim-Song (CKKS) homomorphic encryption, without revealing the raw biometric data. Transparency is enhanced with the help of the Grad-CAM, which provides interpretability of the model's decision. The proposed system is evaluated on the CASIA-FaceV5 and CASIA-FingerprintV5 datasets, demonstrates the low error rate of 0.0038 on fused feature representation.

KEYWORDS: Multimodal Biometrics, Privacy Preservation, Explainable AI, Homomorphic Encryption, Trust-Adaptive Fusion.

1 Introduction

Nowadays, biometric systems are very common in order to authenticate people securely. They use physical/behavioural characteristics such as fingerprints or facial features to determine who a person is. Unimodal systems are easy to find, but they are not always reliable when the input is noisy or incomplete[1]. Multimodal biometric systems (combining two or more traits, including face and fingerprint) are the solution to this problem, as they not only improve recognition accuracy but also improve resilience to spoofing and presentation attacks[2]. The features that were employed in earlier systems were handcrafted i.e. edges, textures or local descriptors like Local Binary Pattern (LBP) and Scale-Invariant Feature Transform (SIFT)[34]. These features are easy and quick, but they face noise, lighting or pose problems[56]. The current Deep Learning (DL), in particular Convolutional Neural Networks (CNNs), is able to learn meaningful patterns when presented with raw data.[7]. Attention mechanisms have enhanced the performance of DL models largely in recent times. These mechanisms enable the system to concentrate on the most significant areas within the input image like facial peculiarity or even unique fingerprint patterns[8].

Biometric traits cannot be replaced or reset in case they are stolen, unlike passwords. This is why securing biometric information is an essential need. To address this problem, researchers have suggested several schemes of biometric template protection, such as cancellable biometrics and homomorphic encryption (HE)[9]. HE is unique since it enables matching on encrypted data directly, thus keeping the features secure at all times. HE, more specifically, allows calculation on encrypted data without the need to decrypt the encrypted information and provides high privacy. HE schemes may be categorised into Partially Homomorphic Encryption (PHE)[10], Somewhat Homomorphic Encryption (SHE)[11], and Fully Homomorphic Encryption (FHE)[10] with the latter being the most powerful and appropriate to carry out secure biometric computations. Of these, the CKKS version of FHE is utilized in the proposed system because it is the most efficient to deal with biometric feature matching.

The majority of DL biometric systems are black boxes and thus users cannot easily understand why a particular choice was made[12]. The other issue is security. Moreover, our Trust- Adaptive Fusion (TAF), rather than the basic concatenation approach, assigns greater importance to trustworthy inputs and lesser to the invalid ones, generating more powerful and reliable outputs.

1.1 Contributions

The key contributions of the proposed system are as follows:

- Integrated MobileNet with channel attention to extract features, which is efficient to recognize discriminative and modality-specific features in strong representations of multiple modalities.
- Introduced a trust-adaptive fusion approach that dynamically weights face and fingerprint modalities according to their reliability to enhance recognition with noisy or poor-quality inputs.
- Secured through the deployment of the CKKS homomorphic encryption scheme on the merged features in such a way that matching the encrypted features is privacy preserving, and it does not require access to the unencrypted data.
- Grad-CAM visualization has been incorporated to explain the visualization, as the salient areas that determine the decisions made by the model, thus giving it a more transparent look.

The rest of the paper is structured as follows: Section 2 reviews related work. Section 3 explains our proposed method. Section 4 presents results and comparisons. Section 5 concludes with key findings and future directions.

1.2 Notations Used

The notations used in the proposed system are summarized in the table. 1.

Table 1. Notations used in the proposed methodology

Notation	Description
MN+CA	MobileNet with Channel Attention Model
F	Face modality
FP	Fingerprint modality
F_Features	Face features
FP_Features	Fingerprint features
X	Reference template
Y	Probe template
P_k	Public key
S_k	Secret key
T	Threshold
D_s	Distance between two templates
Enc(X)	Encrypted reference template
Enc(Y)	Encrypted probe template

2 Related Work

The literature on secure and efficient multimodal biometric fusion has been studied extensively. Barni et al.[13] developed a privacy-preserving authentication framework using the SPDFZ tool, demonstrating low EER and rapid computation, as well as security against malicious adversaries. Walia et al.[14] enhanced the reliability of multimodal recognition by using evolutionary optimization to resolve conflicts among classifiers. Dwivedi et al.[15] suggested weighted score-level fusion of cancelable biometrics, with very low EER and high genuine match rates. Rathgeb et al.[16] enhanced the security of templates by adding face and iris features into a Bloom filter representation, making the template more resistant to reconstruction attacks. Aleem et al.[17] proved that combining traits such as face and fingerprint achieves more reliable results than unimodal systems and suggested a multimodal biometric recognition system that would incorporate both face features and fingerprint features. The methodology involves Gabor-based extraction of fingerprint features and ELBP-LNMF extraction of facial features and dimensionality reduction and score level fusion through sum rule. The experiments on the ORL face dataset and FVC fingerprint dataset obtained a recognition accuracy of 99.59% and a EER of 0.035%. Vallabhadas et al.[18] introduced a cancelable biometric framework using a Cancelable Convolutional Neural Network (C-CNN) for iris and fingerprint modalities. By embedding random transformations, the approach generated cancelable templates that provided template security while

preserving recognition performance. Experiments on standard iris and fingerprint datasets achieved an EER of 0.038%. Li et al.[19] introduced a cancellable multi-biometric system which combined the features of fingerprint and finger vein. Their framework offered template security, revocability and strong resistance to attacks, and high recognition accuracy, being EER 0.09%. Sasikala et al.[20] concentrated their research on the fingerprint and retina modalities. Attention EfficientNet B7 was used to extract features, and the hash codes were generated and compressed by Diagonal Hash Compression (DHC). The Sparrow Search Optimization (SSO) technique was also applied to further optimize and minimize loss and false acceptance. Their system attained a balance of accuracy and security with the EER of 0.12%. Dang et al.[21] proposed the Adaptive Vector Transformation (AVET), a cancellable biometric scheme. Their method was tested on CASIA-FaceV5 (face), FVC2002 (fingerprint), CASIA-Iris (iris), and PolyU (palmprint) datasets, achieved 96–98% accuracy with EERs of 1.7–2.4%. With multimodal fusion, performance improved to 99% accuracy and 2.53% EER. The scheme ensures privacy and cancellability, but validated only on benchmark datasets. Purohit et al.[22] introduced a multimodal biometric fusion model based on deep neural networks (DNNs), which incorporated pixel-level, feature-level, and score-level fusion. Vijay et al.[23] have created a multimodal system using a deep belief network (DBN) and an algorithm known as a hybrid Chicken Earthworm Optimization (CEO). The model had specificity, sensitivity, and accuracy of 98.79%, 95.85% and 95.36% respectively, using finger images, ear images, and iris images. El Mehdi et al.[24] demonstrated that cascade decision-level fusion of fingerprint, finger-vein, and face modalities significantly improves recognition accuracy, but the use of multiple hand-crafted features and used complex fusion rules increases computational complexity. Mwaura et al.[25] developed a multimodal biometric system that combines face and fingerprint at the match-score level. They used SIFT for feature extraction, Hamming distance with Best Bin First (BBF) for matching, and a weighted sum rule for fusion. Using data from FVC 2004 (fingerprints) and Face94 (faces), the system reached 92.5% accuracy, which was better than using only face (90%) or only fingerprint (82.5%). Mehdi et al.[24] proposed a multimodal biometric system combining fingerprint, finger-vein, and face modalities using CNN-based feature extraction. Fingerprint and face features were classified with Softmax, while finger-vein features were used with a Random Forest classifier. Pre-processing included K-means segmentation for fingerprints and exposure fusion for finger-vein images. The individual scores were fused using a weighted sum rule. Kazi et al.[26] developed a multimodal biometric system that uses face, fingerprint, and signature for authentication. They tested a wide range of methods, including 13 feature extraction techniques (such as PCA combined with DCT, DWT, and SVD) and 32 different matching approaches. To improve performance, they applied two fusion strategies—score-level fusion using sum and max rules with normalization, and decision-level fusion using the AND rule. Tables. 2 and 3 gives an overview of multimodal biometric works comparatively focusing on different fusion approaches and performance, security trade-offs in accordance to different datasets.

Multimodal biometric authentication systems currently available have been shown to be more accurate than unimodal systems, though there are a few limitations that have yet to be addressed. Numerous previous studies center their research on the fixed fusion techniques, where all modalities are considered equally reliable, and thus, they are susceptible to high-noise or poor-quality input. Moreover, privacy-preserving mechanisms are commonly considered separately on fusion, and few adaptations of encrypted-domain matching into adaptive multimodal models are realized. Also, the majority of deep fusion methods are based on the focus on performance and lack transparency of model choices, lowering user confidence and interpretability of the system. Hence, no single multimodal system is available to concurrently take into account modality reliability, biometric template security, and explainability.

3 Methodology

The proposed system uses a client-server architecture with a Trusted Authenticator (TA) to provide secure multimodal biometric authentication relying on the face and fingerprint traits, as illustrated in Figure.1. The proposed system contains enrollment and verification stages. In the enrollment stage, the client device captures face and fingerprint images of the user, and generates the discriminative features using the MN+CA model, which produces modality-specific confidence scores as well. The identified features are combined on a feature level through the fusion process based on a confidence and trust mechanism, where the fusion mechanism is informed by the confidence score and can decide which features are to be combined in order to create a stronger and more reliable fusion. The fused feature vector is then transmitted to the TA, which encrypts it with FHE and stores it in a cloud server in a secure way as the reference template. During the verification step, new face and fingerprint samples are run through the same MN+CA model to obtain features and confidence scores. The features are combined with the confidence and trust-based feature-level fusion strategy, and sent to the TA, whereby the fused-up one is encrypted to create a probe template. The encrypted probe template is matched against the stored encrypted reference template, and the outcome is compared with a threshold by the TA to either accept or reject the user. The proposed system provides secure biometric templates management and uses confidence and trust-aware fusion to provide secure and privacy-preserving multimodal authentication.

The modules of the proposed system are: 1).Preprocessing 2).Feature Extraction 3).Trust-Adaptive Fusion 4).Privacy-Preserving Matching with Explainability. These modules combined offer a complete solution that is reliable, secure, and understandable. Each of the modules are explained in the following subsections.

Table 2. Taxonomy of multimodal biometric fusion approaches by modality, fusion level, and architectural design.

Paper / Year	Modalities	Fusion Level	Backbone / Extractor
Aleem et al.[17], (2020)	Face, Fingerprint	Score-level	Gabor-based extraction of fingerprint features and ELBP-LNMF extraction of facial features
Vallabhadas et al.[18] (2024)	Fingerprint, Iris	Feature-level	Cancellable CNN and 1D log Gabor Filters
Batouche et al.[27] (2025)	Face, Fingerprint	Feature Level fusion	1D CNN and Pretrained Resnet-50
Zhou et al.[28] (2025)	Iris, Periocular traits	Binary Mask Cross folding fusion	Fine Grained Hybrid Attention Dual path Network (FGHADP-Net) with hybrid attention
Zhao et al.[29] (2023)	Face, Iris	Decision Level Fusion	MobileNetV2 with hashing layer
Naeem et al.[30] (2024)	Face, Knuckle, palm and iris	DCT Based Fusion	DCT and Adaptive Filter
Elsheikh et al.[31] (2024)	Fingerprint, palmprint, Iris, face	Image Level fusion using DCT	DCT and Minimum Average Corelation Energy(MACE) filter
Wang et al.[32] (2022)	Face , Fingerveins	2-Channel CNN feature fusion	Alexnet and VGG19
El rahman et al.[31] (2024)	ECG, Fingerprint	Decision and Score level Fusion	VGG-Net, Minutiae and Signal processing
Lee et al.[33] (2021)	Face, Fingerprint	Feature level Fusion	Facenet and Kernel-PCA on Minutiae Cylinder code
Kim et al.[34] (2022)	Face, Periocular traits	Cancellable Softmaxout fusion network	Resnet-50
Vallabhadas et al.[35] (2023)	Iris, fingerprint	Feature level fusion	Minutiae based distances and pretrained VGG-16
Morampudi et al.[36]	Fingerprint, Iris	Feature level fusion	Minutiae based and Iris Codes
Li et al.[19]	Finger, FingerVein	feature level fusion	1.Randomly permuted with a permutation seed and 2. Concatenated by the AND operartion
Sasikala et al.[20] (2025)	Fingerprint, Retina	Feature Level	Attention EfficientNet B7
Dang et al.[21] (2022)	Face, Fingerprint, Iris and Palmprint	feature level fusion	FaceNet, VGG-based CNN, Minutiae Cylinder Code, Hand crafted statistical features
Purohit et al.[22] (2021)	ear, palm, fingerprint	optimal feature-level fusion	Gabor, HMSB, shape and texture feature extractor
Vijay et al.[23](2021)	Finger, Ear and Iris	Score level fusion	Handcrafted (Ear-Texture features + BiComp mask, Fingervein-Radon Transform–based features +BiComp and Iris-Daugman's Rubber Sheet Model + texture features)
Kazi et al.[26] (2024)	Face, Fingerprint, and signature	Score-level Fusion and decision level fusion	PCA combined with DCT,DWT and SVD
Jha et al.[37] (2025)	Face , Voice	mutual information feature level fusion	Optimized Ensemble learning

Table 3. Comparison of evaluation protocols, datasets, and reported performance metrics in multimodal biometric systems.

Paper / Year	Datasets Used	Protocol	Metrics Reported	Key Takeaway/Limitation
Aleem et al.[17], (2020)	FVC 2000 DB1 & DB2, ORL(AT&T) and YALE	identification and verification	Recognition Accuracy, EER, FAR, FRR, ROC	Score level fusion improves the recogntion accuracy, but dependence on handcrafted features limits scalability.
Vallabhadas et al.[18] (2024)	Casia iris v3 and FVC 2002 db2 and CMDB	Authentication (Verification)	EER	Multiplies fused biometric features by a user-provided seed before CNN processing to ensure template cancelability.
Batouche et al.[27] (2025)	Face Recognition dataset from kaggle, Sokoto Conventry	Identification (1:N) with incremental dataset updates	FAR, FRR, FPR and FRR, ROC-AUC, Accuracy, precision, recall, f1 score	Scalable user addition without full retraining and minimal forgetting.
Zhou et al.[28]	IITDV1, MMUv1	Verification (1:1) cloud-based	EER, Unlinkability	Limited robustness evaluation; small-scale datasets
Zhao et al.[29] (2023)	VGG-Face2, CFP, CASIA-Iris-Thousand, CASIA-Iris-Lamp	Authentication (Verification)	EER and GAR	Accurate but limited to face–iris modalities.
Naeem et al.[30] (2024)	AT&T face, IIT Delhi knuckle/palm print/iris databases	Verification (1:1)	EER and AROC	Robust to noise; seed-dependent template generation
Elsheikh et al.[31] (2023)	AT&T face, IIT Delhi finger knuckle, IIT Delhi touchless palmprint, and IIT Delhi iris databases	Verification (1:1)	EER, FAR, Average Accuracy, Average Authentication Time, AROC	Very high accuracy; low Gaussian noise sensitivity.
Wang et al.[32] (2022)	CASIA-WebFace, Finger Vein USM (FV-USM), and SDUMLA-FV	Identification (1:N) with cross-validation	Accuracy, AUC	Self-attention-based bimodal weighting
El rahman et al.[38] (2024)	MIT-BIH ECG and FVC2004 databases.	Identification	ROC,AUC	Sequential ECG–fingerprint multimodal fusion using CNN achieves higher accuracy than unimodal systems. Tested only on controlled datasets
Lee et al.[33] (2021)	FVC and LFW	Verification (1:1) employing the FVC matching protocol	EER, Processing time	Removes the need for physical tokens by binding transformation keys directly to fused biometric vectors using XOR.
Kim et al.[34] (2022)	AR + Ethnic + FaceScrub + IMDB WIki + Pubfig + YTF	Authentication (Verification) with open-set evaluation	EER	Implements a learnable SoftmaxOut hashing module and diagonal compression to produce compact, discrete biometric templates.
Vallabhadas et al.[35]	Casia v1+FVC2004 and Casia v3+FVC2006 and Children multimodal biometric dataset	Authentication (Verification)	EER	Generates a secure 3D spiral curve by integrating 2D fingerprint features with iris-extracted vectors
Morampudi et al.[36]	FVC 2002 DB2 + IITD and CMDB	Authentication (Verification)	EER	Focused on privacy preservation but does not address adaptive fusion
Li et al.[19]	FVC2002, FVC2004, Saint Deem Finger Vein dataset	Verification	EER	Accuracy affected by the alignment and fusion strategy.
Sasikala et al.[20] (2025)	Digital retina images for vessel extraction (DRIVE) and FVC2004	Authentication (Verification)	Accuracy and EER	Tested only on benchmark datasets , improves accuracy and security
Dang et al.[21] (2022)	LFW, CFPW, CASIA-FaceV5, IITD-E, gait dataset	Verification and Identification	FAR, FRR, EER, ROC, AUC, RI, CMC curve	Enhances Biometric template security, while largely preserving recognition performance, with a small accuracy trade off compared to linear random projection.
Purohit et al.[22] (2021)	IITD, CASIA-palmprint, CASIA-FingerprintV5	Authentication (Identification)	Accuracy, specificity, sensitivity	Improves accuracy and security but increase the dimensions and computational cost
Vijay et al.[23](2021)	SDUMLA-HMT, AMI Ear dataset	Authentication (Verification)	Accuracy,TPR, TNR, ROC/TPR vs FPR	Improved Authentication accuracy , but limited scalability
Kazi et al.[26] (2023)	YALE, FVC2002 and KVKR	Authentication (Verification)	FAR, FRR, TAR, EER, Accuracy, ROC and DET curves, Precision, Recall, F1score, Sensitivity, Specificity	Multimodal biometric performance is improved due score level fusion, but the handcrafted features make the system complex and less scalable
Jha et al.[37] (2025)	Voxceleb1 and VidTIMIT	Authentication (Verification)	Accuracy, EER	Fusion is based on mutual information and optimized learning to face and voice, but without real-world and adversarial evaluation.

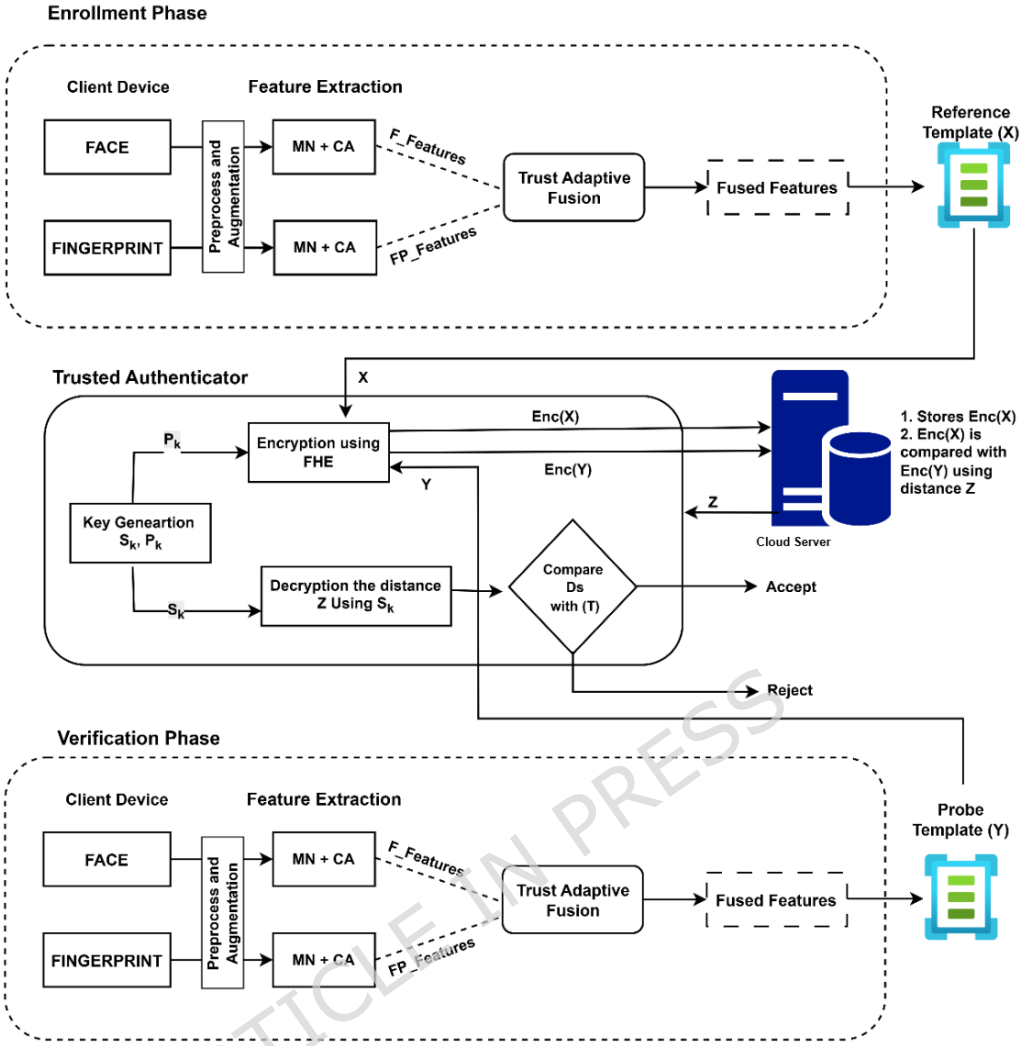


Figure 1. Proposed privacy-preserving multimodal biometric authentication

3.1 Preprocessing

Preprocessing of face modality

Raw facial images are usually susceptible to lighting, pose, and background noise, which may decrease the recognition accuracy. To address these problems, such important operations are carried out: face detection and normalization. The MTCNN (Multi-task Cascaded Convolutional Network)[39] is used to detect. MTCNN is divided into three steps: it initially suggests candidate face locales, secondly determines these faces by removing overlapping or low-confidence instances using Non-Maximum Suppression (NMS) and Intersection over Union (IoU) and lastly provides accurate face locales as well as facial features including the eyes, nose, and mouth. These landmarks assist in correcting pose variations by maintaining constant alignment even in the case of face tilt and rotation. Once the images are detected, the detection is normalized by resizing the image which standardizes the input and minimizes the impact of lighting variations. The preprocessing pipeline removes the variations by extracting normalized regions of the face, eliminates the background interference, and identifies the discriminative features, which allows the recognition model to produce robust performance regardless of pose, expressions and lighting conditions. The sample preprocessed images of the face are illustrated in the Figure.2.

Preprocessing of fingerprint modality

In order to have clean fingerprint images, which are capable of producing good fingerprint feature extractions, a structured preprocessing pipeline was used. It follows a series of four steps. The image is then smoothed with a Gaussian filter to remove small noise and retain significant ridge information. This is followed by thresholding by Otsu[40], which automatically isolates

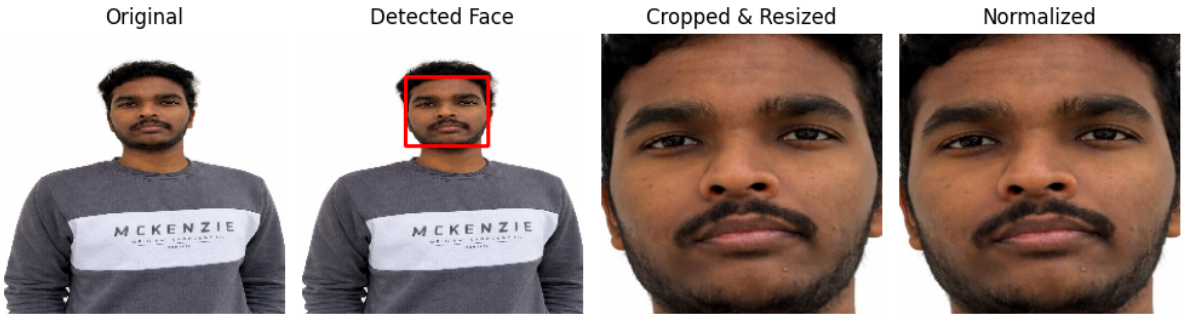


Figure 2. Face preprocessing pipeline

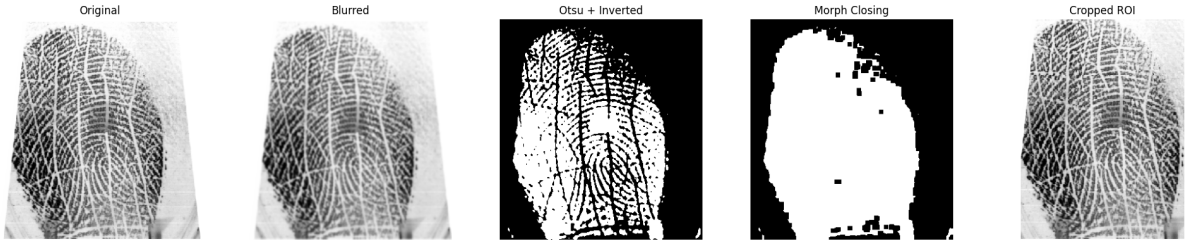


Figure 3. Fingerprint preprocessing pipeline

the ridges and the background and then inverts the image where the ridges are made brighter and more pronounced. Binarization is followed by a morphological closing operation to seal small holes, join fractured ridges and improve the structure of ridges. This enhances the visibility of the fingerprints. Lastly, the isolating of the Region of Interest (ROI) is done by the use of contour detection. The most prominent contour is picked, and the fingerprint is cut to represent only the pertinent area of the ridge and ignore the background noise. The sample preprocessed images of the fingerprint are illustrated in the Figure.3. The fingerprints undergo this pipeline of denoising, thresholding, ridges boosting, and ROI extraction, which standardizes and boosts the fingerprints, enabling them to be utilized more effectively in the process of extracting features.

3.2 Feature Extraction

In the feature extraction, MobileNetV2 was utilized as a backbone because of its lightweight architecture and demonstrated effectiveness in image recognition activities. To further enhance its ability, a Channel Attention (CA) mechanism was added, which allows the network to emphasize the most relevant feature maps and suppress the least relevant feature maps[41,42]. The CA block uses the output of the last convolutional layer of MobileNetV2 as depicted in the Figure. 4. It generalizes all the channel's information by global average and global max pooling, which is then fed through a common two-layer. The initial layer is used to downsample the channel dimension by a factor of r=8 using ReLU activation, and the second layer is used to upsample it to its original scale. The results of the combination and a sigmoid function are then used to get the channel attention weight, between 0 and 1. These weights are weighted with the original feature maps, which amplify the channels that are informative and disable the weak ones.

Table. 4 of the paper gives the detailed layer-wise architecture of the proposed feature extractor, whereas Figure 4 shows a visualization of its structure. The model targets significant biometric traits, like fingerprint ridges and facial shapes. The learned refined feature maps are afterwards compressed by global average pooling and through a dense layer (256 units, ReLU) to generate compact embeddings. This embedding serves as the final feature vector used for multimodal fusion. An additional Dense layer with Softmax activation is attached only to derive confidence scores reflecting classification certainty; this layer is not used for the matching process and is therefore not included in the feature extractor configuration. MobileNetV2 and CA can be used together to provide the system with strong and high-quality feature representations without causing any computational overhead. This was then incorporated into the trust-adaptive fusion mechanism, which led to a more robust and dynamic multimodal biometric system.

3.3 Trust-Adaptive Feature Level Fusion

Once each modality has been extracted, the system computes the confidence scores based on the softmax output of the trained model. A confidence score is defined as the maximum softmax probability associated with the predicted class. The scores

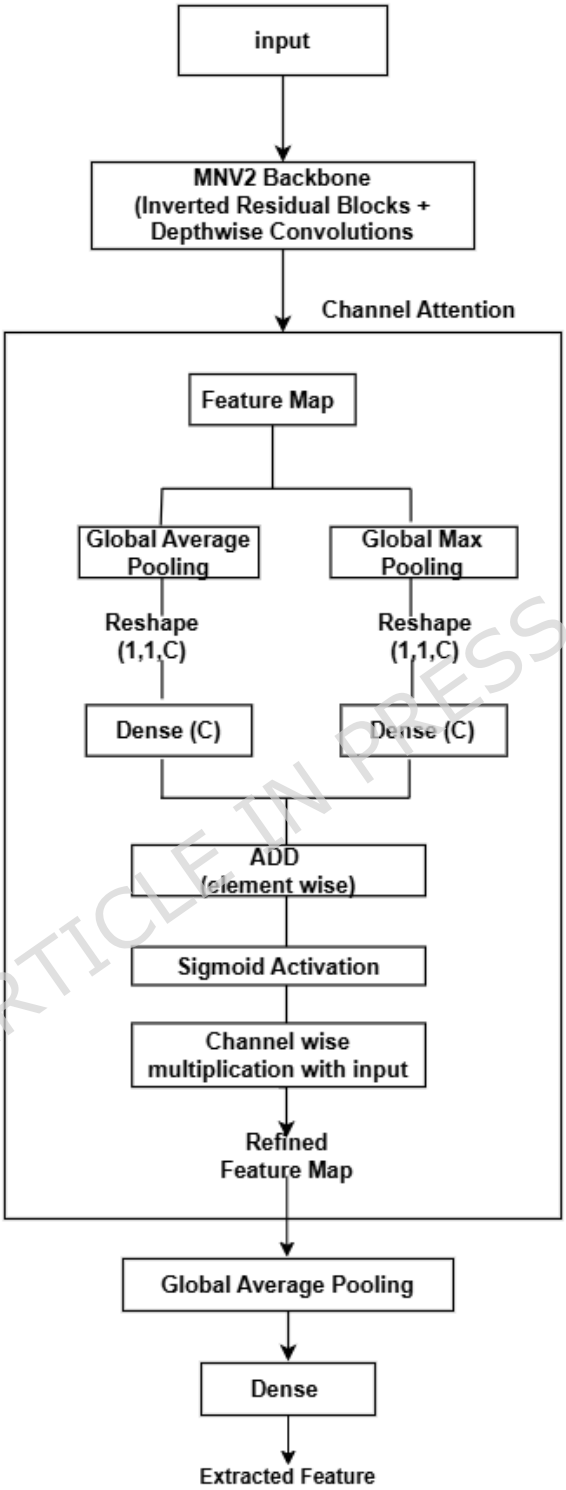


Figure 4. Feature extractor architecture combining MobileNetV2 and channel attention.

Table 4. Layer-wise configuration and parameter details of the proposed feature extractor with MobileNetV2 backbone and Channel Attention (CA) block.

Layer / Block	No. of Filters / Units	Output Feature Map Size	Kernel Size / Stride
Input Image	–	$224 \times 224 \times 3$	–
Conv2D + BN + ReLU6	32	$112 \times 112 \times 32$	3×3 / 2
Bottleneck Blocks (Inverted Residuals, repeated)	$16 \rightarrow 320$	$112 \times 112 \rightarrow 7 \times 7$	Depthwise 3×3 / {1,2}
Final Conv2D	1280	$7 \times 7 \times 1280$	1×1 / 1
Channel Attention (CA) Block			
Global Average Pooling (GAP)	–	$1 \times 1 \times 1280$	–
Global Max Pooling (GMP)	–	$1 \times 1 \times 1280$	–
Shared MLP (FC $\rightarrow$ FC, $r = 8$)	$1280/8 \rightarrow 1280$	$1 \times 1 \times 1280$	1×1 / 1
Add (GAP path + GMP path)	–	$1 \times 1 \times 1280$	–
Sigmoid Activation	–	$1 \times 1 \times 1280$	–
Multiply (with input feature maps)	–	$7 \times 7 \times 1280$	–
Global Average Pooling (GAP)	–	$1 \times 1 \times 1280$	–
Dense + ReLU	256	$1 \times 1 \times 256$	–

signify that the model is very sure in what it predicts, but the confidence level does not always indicate reliability especially when the model is put in degraded scenarios like the presence of a blurred fingerprint or low-light face images. In order to overcome this drawback, a trust-adaptive fusion (TAF) strategy is adopted to dynamically determine the contribution of each modality according to its reliability. The confidence scores of face C_f and fingerprint C_p are normalized to get the trust scores, which indicate the comparative reliability of each modality using the following equation

$$T_f = \frac{C_f}{C_f + C_p}, \quad T_p = \frac{C_p}{C_f + C_p} \tag{1}$$

where

C_f = Face Confidence Score, C_p = Fingerprint Confidence Score

T_f = Face Trust Score, T_p = Fingerprint Trust Score

Using these trust scores, the final fused feature vector is computed using the eq.2

$$F_{\text{fused}} = T_f \cdot F_{\text{face}} + T_p \cdot F_{\text{finger}} \tag{2}$$

where

F_{fused} is Final Fused Vector, F_{face} is normalized face feature vector, F_{finger} is normalized fingerprint feature vector

This formulation is such that modalities of greater trust make more contribution to the fused representation, whereas less reliable modalities are down-weighted. The face and fingerprint modalities are extracted using a modified MobileNet architecture with Channel Attention that uses 256 dimensions each per modality as a feature. MobileNet is used to achieve computational efficiency, and Channel Attention enhances the processes of extracting discriminative and modality-specific features. The feature dimension is selected to ensure a desired level of discriminative strength and at minimal computational cost. The resultant fused feature vector is then run through fully connected layers that model nonlinear inter-modal relations

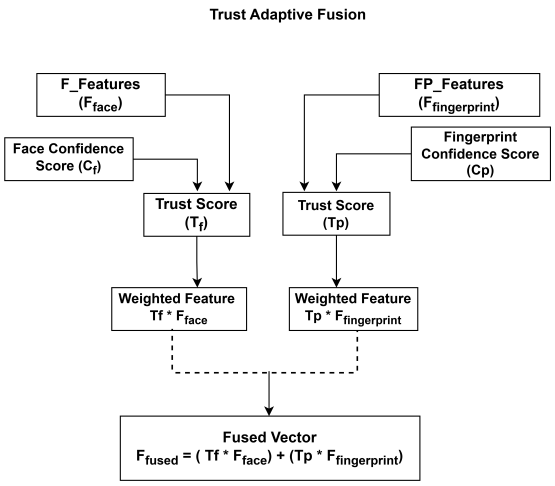


Figure 5. Trust-Adaptive Fusion - Integrates multi-modal biometric scores dynamically based on trust levels, ensuring reliable and context-aware decision-making.

and in effect increases the overall discriminative ability of the system. The proposed process of trust-adaptive fusion is shown in Fig.5 The proposed TAF mechanism gives the multimodal authentication a better robustness since it dynamically adapts to the quality of each input mode, and hence, it is specifically applicable in real-life situations where a single modality can be noisy, degraded, or even unavailable.

3.4 Privacy-Preserving Matching with CKKS Encryption

After computing the fused feature vector F_{fused}, which is 256 in size, it is sensitive biometric data of both the face and the fingerprint. In order to safeguard this information during storage and transmission, we use the CKKS homomorphic encryption scheme that enables arithmetic operations to be done on encrypted data directly. The encryption of the fused feature vector is denoted using the eq.3

$$\tilde{F}_{\text{fused}} = \text{Encrypt}_{\text{CKKS}}(F_{\text{fused}}), \tag{3}$$

where $\tilde{F}_{\text{fused}}$ represents the encrypted feature vector, which is stored as a reference template in the cloud server. During verification, the similarity between the stored reference template and a probe template $\tilde{F}_{\text{query}}$ can be computed entirely in the encrypted domain using cosine similarity:

$$\text{CosineSim}_{\text{encrypted}} = \frac{\tilde{F}_{\text{fused}} \cdot \tilde{F}_{\text{query}}}{\|\tilde{F}_{\text{fused}}\| \, \|\tilde{F}_{\text{query}}\|} \tag{4}$$

which allows the system to obtain similarity scores without ever revealing the raw biometric features. So that original biometric data (basically face or fingerprint features) is never exposed to the outside world through the secure system, thus securing the privacy of the users and enabling accurate user verification.

There are two stages involved in the computation of the cosine similarity. Firstly, all arithmetic is operated on ciphertexts in the encrypted domain: element-wise multiplication $\tilde{F}_{\text{fused}} * \tilde{F}_{\text{query}}$ produces an encrypted vector of pairwise products, and summing all elements of an encrypted vector is the encrypted dot product $\tilde{F}_{\text{fused}} \cdot \tilde{F}_{\text{query}}$; and computation of squared norms is as $(\tilde{F}_{\text{fused}} * \tilde{F}_{\text{fused}}).sum()$ and $(\tilde{F}_{\text{query}} * \tilde{F}_{\text{query}}).sum()$ provides the encrypted squared magnitudes $\|\tilde{F}_{\text{fused}}\|^2$ and $\|\tilde{F}_{\text{query}}\|^2$. Second, after these homomorphic operations, only the aggregated results are decrypted in plaintext: the dot product and squared norms are decrypted, the norms are square-rooted to obtain $\|F_{\text{fused}}\|$ and $\|F_{\text{query}}\|$, and the final cosine similarity is computed by dividing the dot product by the product of norms. This design ensures that the raw biometric features never leave the encrypted domain while still allowing accurate similarity computation. The CKKS scheme is configured to support effective and accurate computation with a degree of 8192 modulus polynomials, coefficient modulus sizes of $[60, 40, 40, 60]$ bits and a global scale of 2^{40}. The polynomial modulus degree allows all 256 features to be packed into a single ciphertext while maintaining approximately 128-bit security. The scale of the selected modulus coefficients guarantees the precision required in the homomorphic multiplications and summations, limiting the number of errors in the calculations, and the global

scale maintains the accuracy of the result obtained after each operation. This parameterization is a tradeoff between security, numerical stability and computational efficiency that allows real-time privacy-preserving cosine similarity computations to be possible in multimodal biometric authentication.

3.5 Explainability

DL models can be considered to be black boxes since it is hard to comprehend how such models make decisions. To overcome this, the framework proposed has explainability by including Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM analyzes the areas of the image that contribute most to the prediction made by the model, thus giving it an interpretative connection between the input and the decision. The generated heatmaps are warmer colors (red/yellow) that are of high importance and cooler colors (blue) that are not as important. Grad- cam process starts with locating the gradients of the class score to the last convolutional layer. These gradients are applied in assigning weight to the feature maps. The weight of feature map (or the k th element) of the class c is determined as in eq.5:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

Here,

α_k^c : contribution of the k-th feature map to class c,
y^c : score corresponding to class c,
Z : total number of pixels in the feature map.

Once the weights are obtained, the final Grad-CAM heatmap is generated as shown in eq.6:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{6}$$

In terms of weighted feature maps, the eq.6, is used to eliminate all the features that do not contribute positively to the decision. With such an implementation of Grad-CAM, the proposed system can be considered not only for high recognition accuracy but also for transparency and interpretability to ensure that the predictions offered by the model are easier to understand and trust

4 Experimental Results and Analysis

4.1 Datasets Used

Face Dataset: The CASIA-FaceV5[43] dataset provided by the Institute of Automation, Chinese Academy of Sciences (CASIA) was used for training and evaluation. The CASIA Face Image Database Version 5.0 (CASIA-FaceV5) consists of 2,500 color facial images from 500 subjects. These images were captured in a single session using a Logitech USB camera, with participants including graduate students, workers, waiters, and others. Each image is stored as a 16-bit color BMP file with a resolution of 640 × 480 pixels. The database reflects typical intra-class variations such as changes in illumination, pose, facial expression, eyeglasses presence, and imaging distance, making it suitable for robust face recognition research.

Fingerprint Dataset: The CASIA-FingerprintV5[44] dataset provided by the Institute of Automation, Chinese Academy of Sciences (CASIA) was used for training and evaluation. The CASIA-FingerprintV5 database consists of 20,000 grayscale fingerprint images from 500 subjects, collected in a single session using a URU4000 fingerprint sensor. Each subject contributed 40 images from eight fingers (left and right thumb, index, middle, and ring), with five samples per finger. To introduce intra-class variations, participants were instructed to apply different finger rotations and pressures during acquisition. All images are 8-bit BMP files with a resolution of 328 × 356 pixels.

Multimodal Fusion and Dataset Correspondence: Although both CASIA-FaceV5 and CASIA-FingerprintV5 datasets contain 500 classes labeled from 0 to 499, CASIA does not provide official confirmation that the subjects in the two datasets are identical. In the proposed system, multimodal fusion is performed at the feature level[37,45,46], without assuming direct subject-level correspondence between face and fingerprint samples. The identical class labels were used solely to maintain consistency during performance evaluation and analysis, rather than to enforce paired identity mapping across modalities. This design allows the fusion system to effectively combine information from face and fingerprint features, while keeping the methodology clear and avoiding incorrect assumptions about whether the two datasets contain the same individuals.

4.2 Experimental Setup

All experiments were conducted on a dedicated workstation to ensure consistent and reliable performance evaluation. The experiments were executed on a Dell Precision 3660 machine running Windows 10 Pro (64-bit), equipped with an Intel Core i7-12700 (12th generation) processor and 32 GB of RAM. For accelerated computation, the system utilized an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM, along with an Intel UHD Graphics 770 integrated GPU. The implementation was developed using Python 8, and the DL models were built and executed using TensorFlow version 2.12.0. All encryption-related operations were performed using the TenSEAL library. This controlled computational environment ensured stable training, testing, and evaluation of the proposed framework, allowing fair assessment of its performance across all experimental stages.

4.3 Model Training

Each user in the CASIA-FingerprintV5 dataset contributed images of eight distinct fingers:

- **Left hand:** thumb (L0), index (L1), middle (L2), ring (L3)
- **Right hand:** thumb (R0), index (R1), middle (R2), ring (R3)

To evaluate per-finger performance, the dataset was explicitly divided into eight finger-specific subsets (one subset per finger L0-L3 and R0-R3). For each subset finger, five samples were available, resulting in 40 fingerprint images per user. Since this limited number of samples may restrict model generalization, data augmentation was applied exclusively to the training set, increasing the number of samples per finger from 5 to 20. Augmentation strategies included small geometric transformations and intensity variations to simulate real-world acquisition conditions. The applied transformations were random rotations (up to 10°), slight width and height shifts (0.05), shear (0.05), zoom (0.1), and nearest-neighbor filling. Horizontal flipping was not applied to fingerprints to preserve the ridge-flow directionality. The CASIA-FaceV5 dataset was also included, with five images per subject. To address the limited sample size, augmentation was similarly applied to the training set, expanding each subject's set to 20 images. The augmentation pipeline for the face modality mirrored that for fingerprints, but also included horizontal flipping to better capture pose variability.

For both modalities, the data was divided into 80% for training and 20% for testing, and only the training set was augmented. For fingerprints, this split was applied separately to each finger-specific subset. This ensured that the model was trained on diverse data while being evaluated on the original samples, providing a fair estimate of real-world performance. The model is trained for 50 epochs with a batch size of 32. The Adam optimizer was used with an initial learning rate of 1×10^{-4}, and categorical cross-entropy was employed as the loss function. The input image size was set to 224×224 pixels for both fingerprint and face modalities.

4.4 Evaluation Metrics

The proposed biometric system is evaluated using standard performance metrics, namely False Acceptance Rate (FAR), False Rejection Rate (FRR), Equal Error Rate (EER), accuracy, and the Receiver Operating Characteristic (ROC) with its Area Under Curve (AUC). These metrics collectively assess the system's accuracy, robustness, and discriminative capability.

- **False Acceptance Rate (FAR):** The percentage of legitimate users (genuine) who are mistakenly denied access by the system.

$$\text{FAR} = \frac{\text{Number of impostor scores above threshold}}{\text{Total impostor scores}}. \tag{7}$$

- **False Rejection Rate (FRR):** The percentage of legitimate users (genuine) who are mistakenly denied access by the system.

$$\text{FRR} = \frac{\text{Number of genuine scores below threshold}}{\text{Total genuine scores}}. \tag{8}$$

- **Equal Error Rate (EER):** The point where FAR and FRR are equal represents the optimal balance between security and usability. A lower EER indicates better system performance.

- **Accuracy:** The percentage of correctly classified genuine and impostor attempts.

$$\text{Accuracy} = \frac{\text{Correctly classified scores}}{\text{Total scores}}. \tag{9}$$

- **ROC and AUC:** The ROC curve illustrates the trade-off between the True Positive Rate (TPR), also called the True Accept Rate (TAR), and the False Positive Rate (FPR, equivalent to FAR) at different thresholds:

$$\text{TPR}(t) = \frac{TP(t)}{TP(t)+FN(t)}, \quad \text{FPR}(t) = \frac{FP(t)}{FP(t)+TN(t)}. \tag{10}$$

The Area Under the Curve (AUC) quantifies overall discriminative ability:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{11}$$

A higher AUC value indicates superior separation between genuine and impostor classes, with 1.0 denoting perfect classification.

where,

$TP(t)$: number of true positives at threshold t,
$FN(t)$: number of false negatives at threshold t,
$FP(t)$: number of false positives at threshold t,
$TN(t)$: number of true negatives at threshold t,
$\text{TPR}(t)$: true positive rate at threshold t,
$\text{FPR}(t)$: false positive rate at threshold t,
AUC : area under the ROC curve.

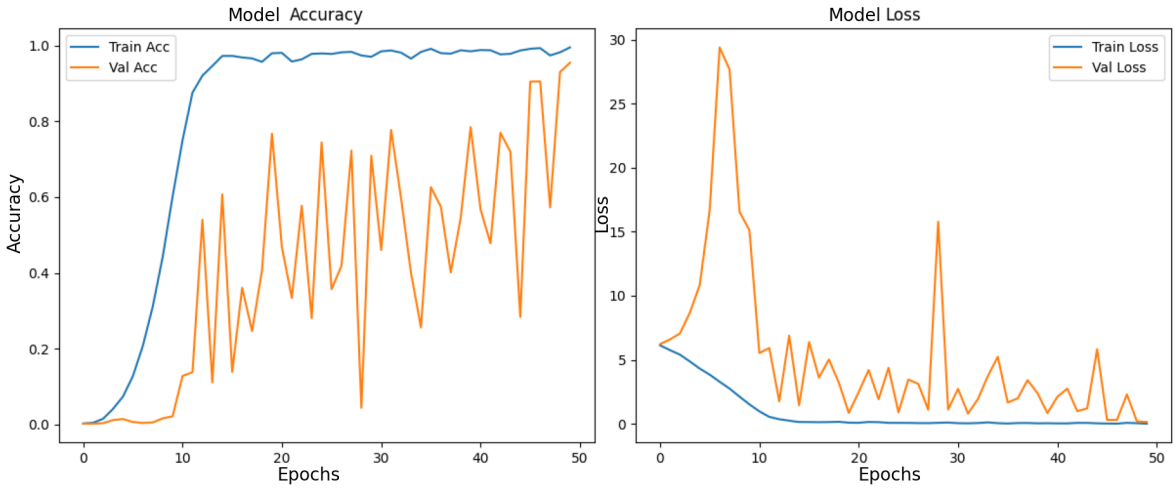
4.5 Experimental Results

Table 5 The accuracies obtained for CASIA-FaceV5 and CASIA-FingerprintV5.

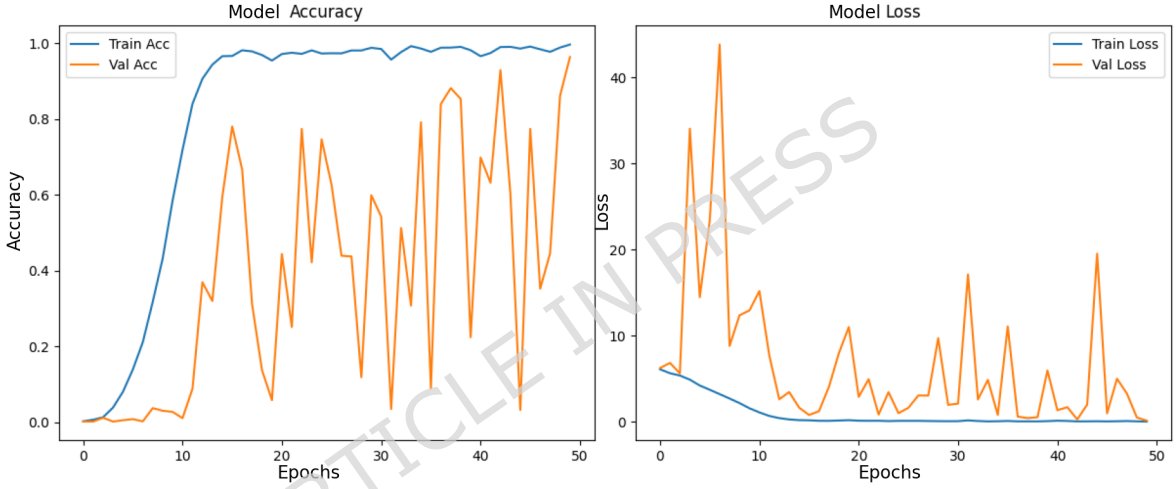
Table 5. Finger-wise and Face accuracy on CASIA-FaceV5 and CASIA-FingerprintV5

Modality	Label	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Left Thumb finger	L0	95.50	93.29	95.34	94.30
Left Index finger	L1	96.35	94.06	96.22	95.13
Left Middle finger	L2	91.70	91.73	91.40	91.56
Left Ring finger	L3	82.40	80.74	81.76	81.25
Right Thumb finger	R0	80.05	80.63	79.32	79.97
Right Index finger	R1	93.75	92.34	93.52	92.93
Right Middle finger	R2	90.50	91.50	90.16	90.83
Right Ring finger	R3	94.50	92.68	94.30	93.48
CASIA-FaceV5	Face	98.75	98.24	98.70	98.47

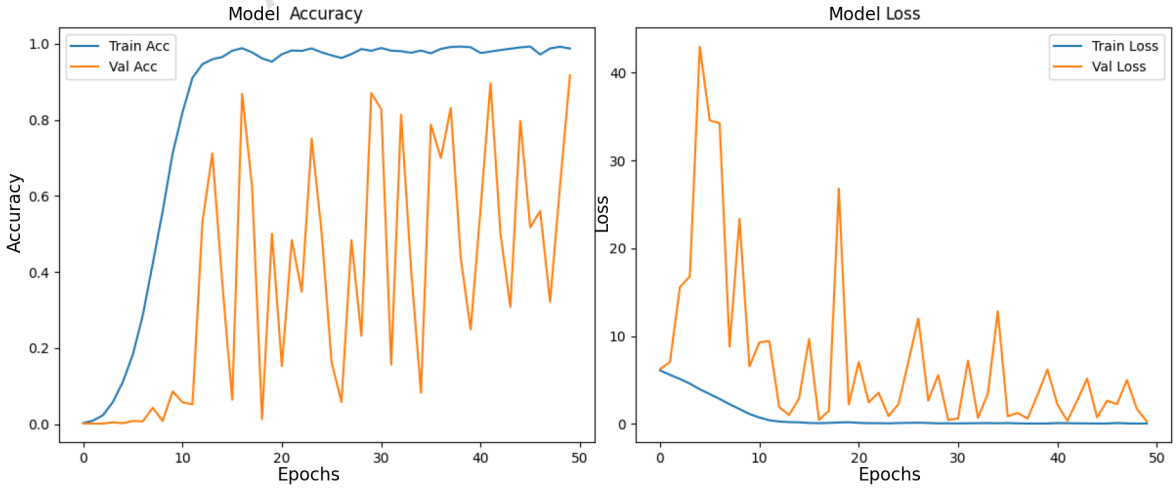
Figure. 6 (a) to (h) represent the accuracy curve and loss curve of the MobileNet model with channel attention. Based on the plots, it is evident that the attention mechanism allows the model to concentrate on the most useful features, resulting in gradual improvements as the training progresses. The loss is kept to a minimum as accuracy increases, indicating that the model is learning successfully. The curve of accuracy and the loss are close, indicating that the model is highly generalizable and it is not overfitted. These findings indicate that channel attention can be added to MobileNet to ensure it is more trusted and suitable to finger-based biometric recognition. On the same note, in the face dataset, accuracy was obtained under augmented sample training. The achievement of the same is presented in the Table.5 and in the Accuracy plot in the Figure.6 (i).



(a)

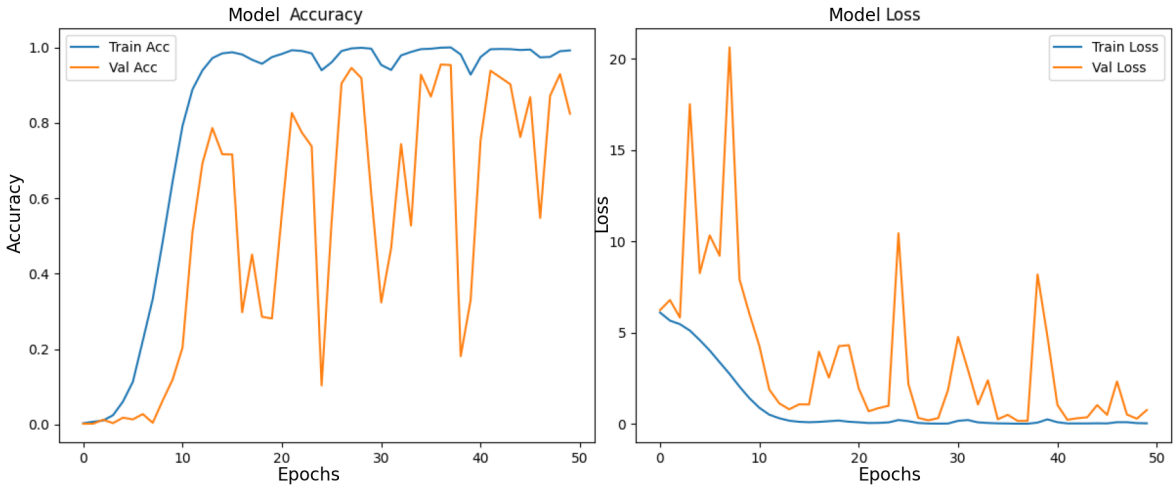


(b)

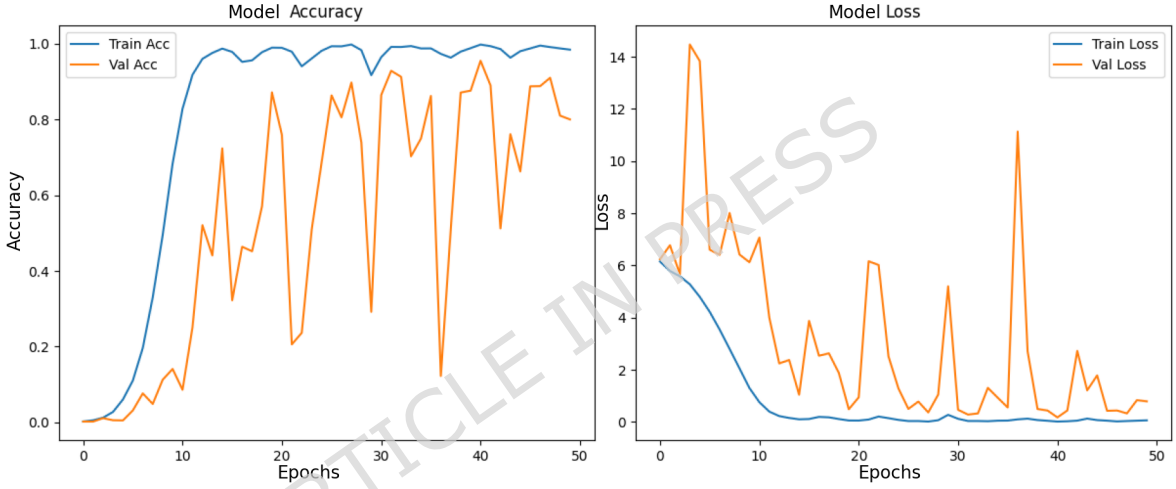


(c)

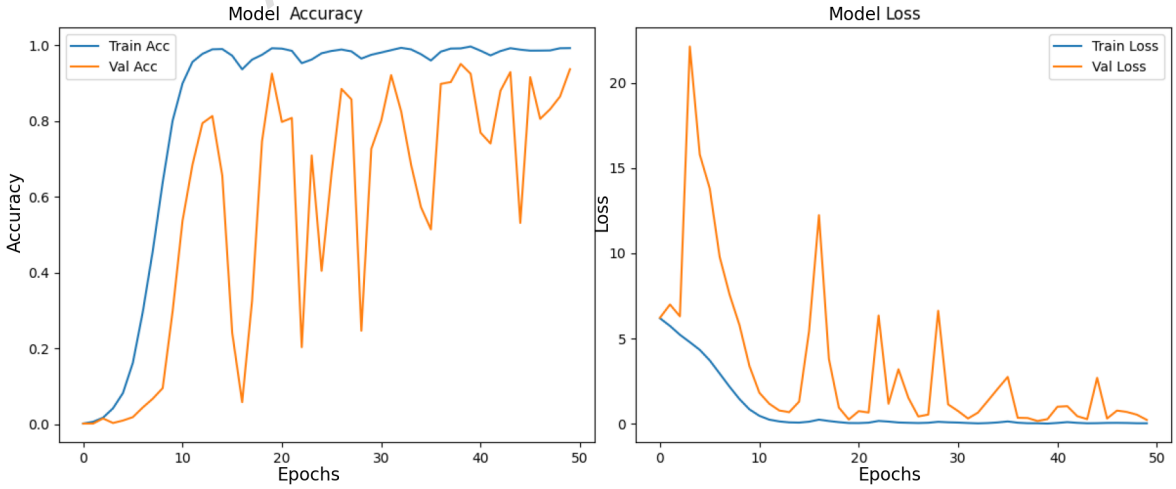
/



(d)

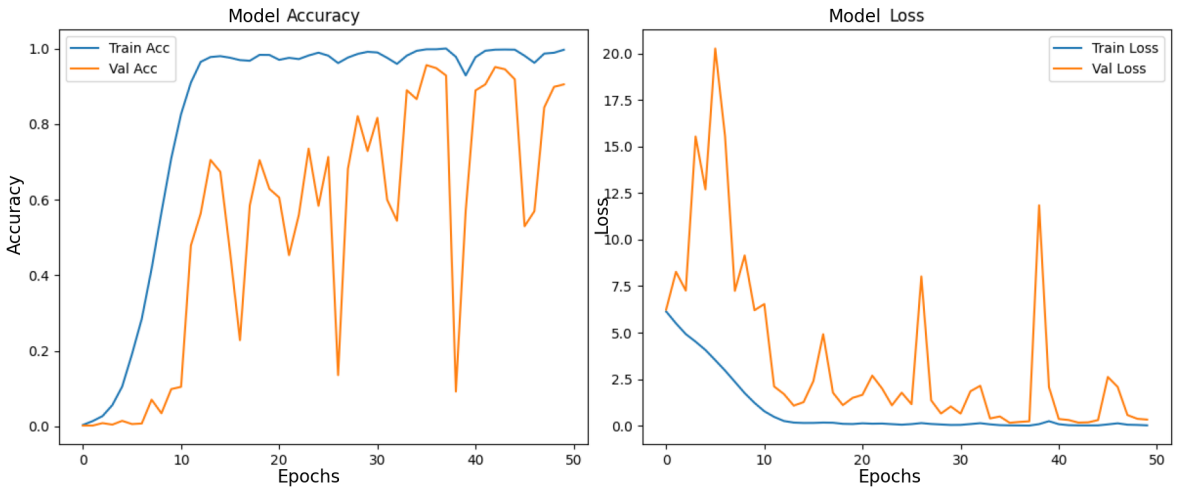


(e)

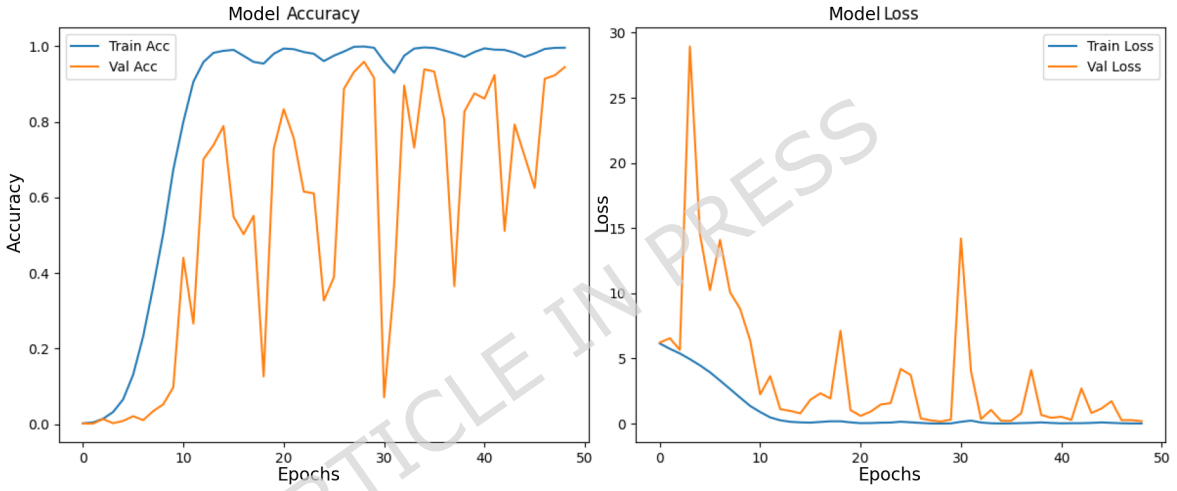


(f)

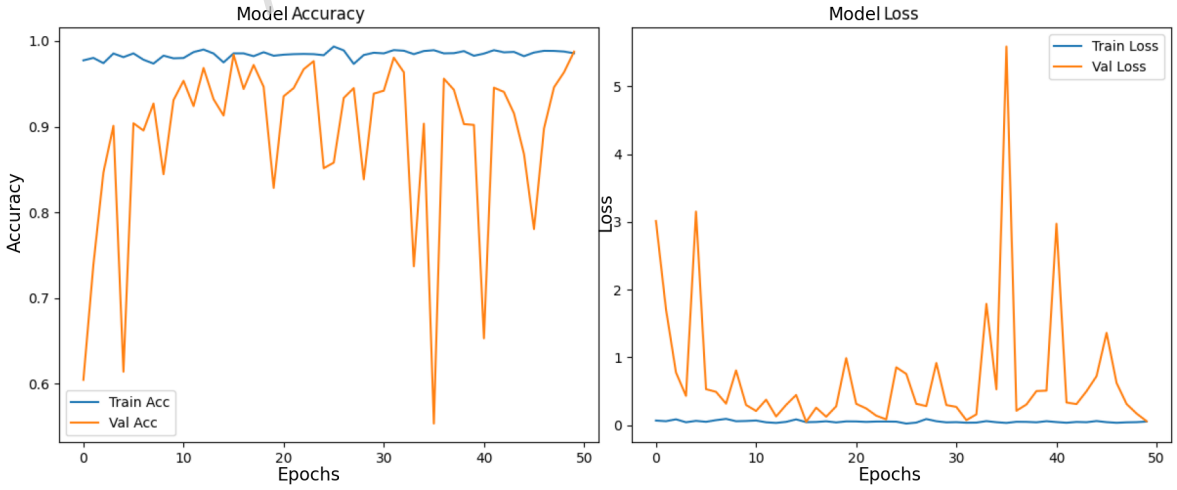
/



(g)



(h)



(i).Face

Figure 6. (a) to (i) Accuracy and loss curves of the MobileNet model with channel attention for Face and different finger classes of CASIA FingerprintV5 dataset, showing how the model learns and generalizes.

/

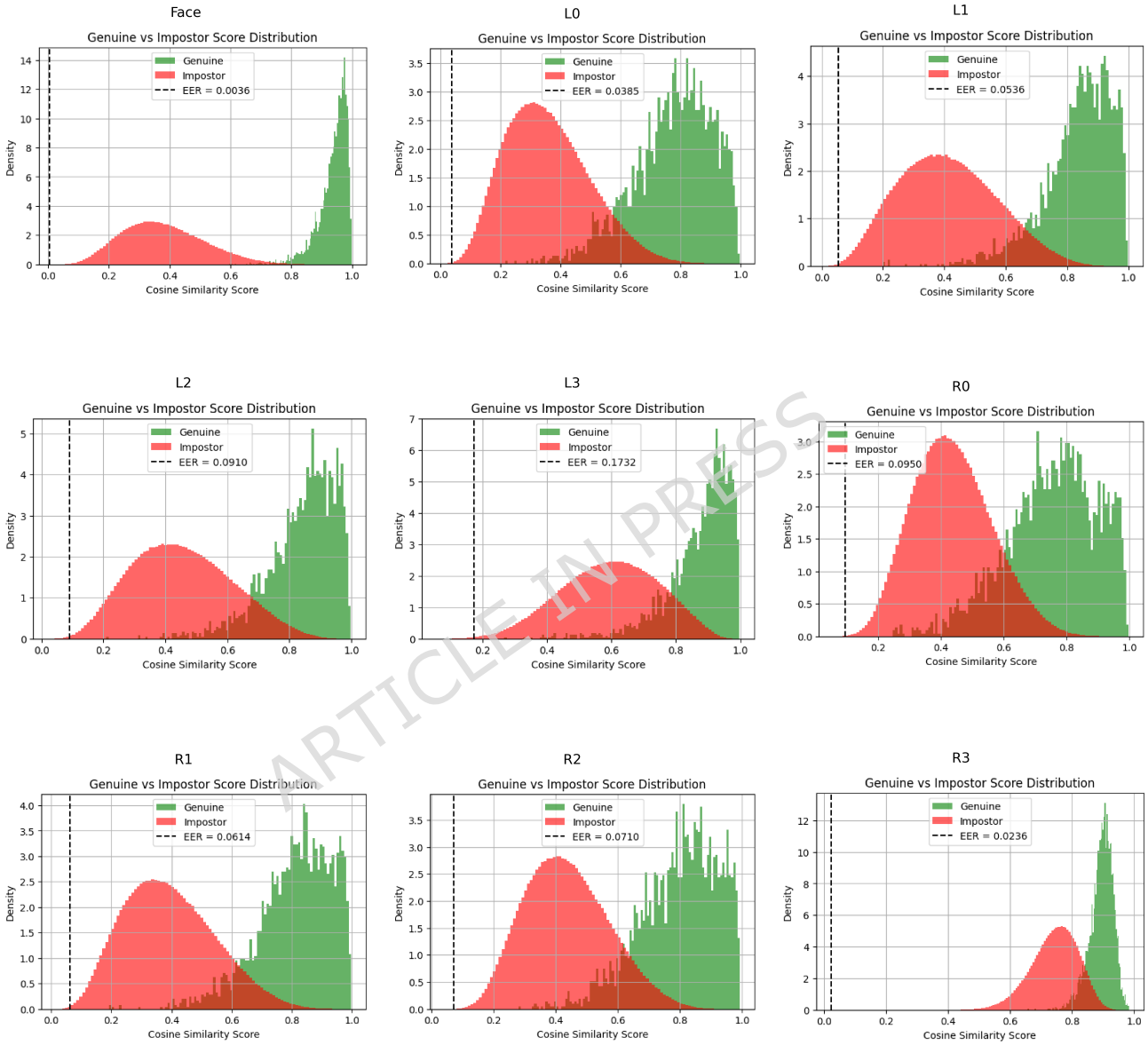


Figure 7. Genuine and impostor score distributions for face and fingerprint instances L0, L1, L2, L3, R0, R1, R2, and R3, showing the EER operating threshold.

Figure. 7 shows the genuine and impostor score distributions for the evaluated face and fingerprint instances, providing a qualitative view of score separability. The decision threshold corresponding to the Equal Error Rate (EER), obtained from the FAR–FRR versus threshold analysis, is indicated to illustrate its position within the score space. This visualization complements the quantitative FAR–FRR and EER evaluation by relating the EER operating point to the underlying score distributions.

Table 6. Biometric Verification performance of the proposed multimodal face-fingerprint fusion under different feature configurations.

Modality	EER	FAR@EER	FRR@EER	TAR@FAR	Accuracy	Precision	Recall	F1 score	ROC
Face + L0	0.0041	0.0041	0.0042	0.9886	0.9959	0.9920	0.9958	0.9939	0.9999
Face + L1	0.0038	0.0039	0.0038	0.9906	0.9961	0.9935	0.9962	0.9948	0.9999
Face + L2	0.0042	0.0042	0.0042	0.9800	0.9958	0.9910	0.9958	0.9934	0.9998
Face + L3	0.0934	0.0930	0.0938	0.4328	0.9070	0.8950	0.9062	0.9006	0.9689
Face + R0	0.0042	0.0042	0.0042	0.9870	0.9958	0.9905	0.9958	0.9931	0.9999
Face + R1	0.0053	0.0053	0.0054	0.9858	0.9947	0.9880	0.9946	0.9913	0.9999
Face + R2	0.0066	0.0065	0.0066	0.9816	0.9935	0.9895	0.9934	0.9914	0.9997
Face + R3	0.0053	0.0053	0.0054	0.9858	0.9947	0.9875	0.9946	0.9910	0.9999

Table.6 shows the biometric verification performance of the proposed face-fingerprint fusion system with various feature configurations. The evaluation of performance is done on the basis of accuracy, precision, recall, F1-score, ROC-AUC, and EER. All the configurations except Face + L3 are highly accurate. The accuracy of most configurations has been reported as higher than 99, and this implies that the proposed system has a high discriminative capability. The values of ROC-AUC are close to 1.0 in all effective configurations, which validates excellent class separability. Face + L1 configuration is the lowest EER, and it depicts the best general verification performance. Face + L3, has a large EER and small AUC. This shows that such a combination of features is not as effective with biometric verification, and the significance of feature selection. The ROC and EER curves of each configuration are shown in Figure.8 and Figure.9. The ROC curves are increasing towards the top-left corner, which represents high true acceptance rates and low false acceptance rates. These findings verify the usefulness of the proposed multimodal face-fingerprint verification system. The values of EER are always low, which is indicative of the good balance between security and usability. Face + L1 and Face + L0 have the most consistent and precise results in all measures.

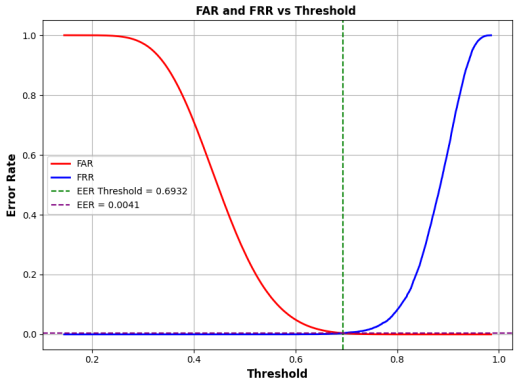
4.6 Bootstrap Statistical Analysis

To assess the statistical robustness of the proposed Multimodal biometric authentication system, bootstrap resampling was employed to estimate confidence intervals for the evaluation metrics. Bootstrap resampling with 1000 iterations was used to estimate 95% confidence intervals for the evaluation metrics. The proposed system achieved a mean EER of 0.0039, with a tight 95% confidence interval of [0.0028, 0.0048], indicating highly accurate and stable verification performance. The ROC analysis yields a mean AUC of 0.99993, with a 95% confidence interval of [0.99990, 0.99995], demonstrating near-perfect discriminative capability between genuine and impostor score distributions. Furthermore, under a strict security constraint of FAR = 10^{-3}, the proposed system attained a TAR of 99.05%, with a 95% confidence interval of [98.77%, 99.36%]. This result highlights the system's ability to maintain high usability while operating at very low false acceptance rates. The consistent low EER, near-perfect AUC, and high TAR at low FAR, together with narrow confidence intervals, confirm that the proposed multimodal biometric system exhibits robust, reliable, and statistically significant performance.

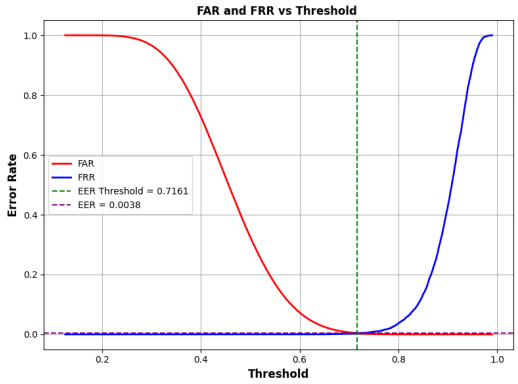
4.7 Comparison Analysis

The performance of the proposed system is analyzed against several recent works in the literature, as shown in the Table. 7. Recent multimodal biometric authentication systems employ different modality combinations and fusion strategies. Purohit et al.[22] explored pixel-, feature-, and score-level fusion using deep neural networks, but did not report EER values. Lee et al.[33] proposed a tokenless face–fingerprint system and achieved an EER of 0.24% on standard datasets. Dang et al.[21] introduced an AVET-based cancellable biometric framework using multiple modalities and reported an EER of 2.53%. Vallabhadas et al.[18] and Li et al.[19] focused on cancelable template generation for iris–fingerprint and fingerprint–finger vein systems, achieving EERs of 0.038% and 0.09%, respectively. Sasikala et al.[20] reported an EER of 0.12% using fingerprint and retina modalities. Higher EERs were observed in face–voice fusion systems, such as Jha et al.[37].

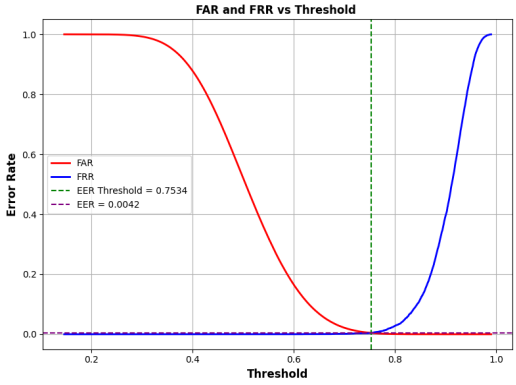
In comparison, the proposed system uses face and fingerprint modalities with feature-level fusion based on TAF and is evaluated on CASIA-FaceV5 and CASIA-FingerprintV5 datasets. It achieves an EER of 0.0038%, a ROC-AUC of 0.9999, and a verification accuracy of 99.61%, indicating highly reliable discrimination. In addition, the system incurs a low computational overhead of 59.4 ms.



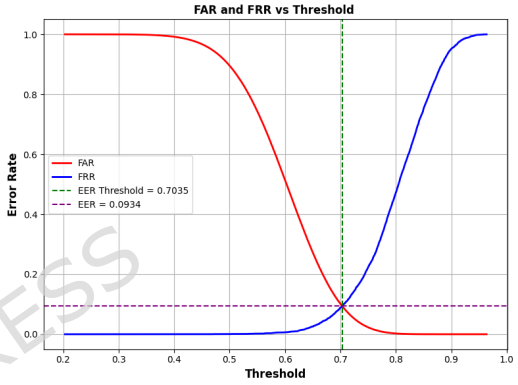
(a) Face+L0



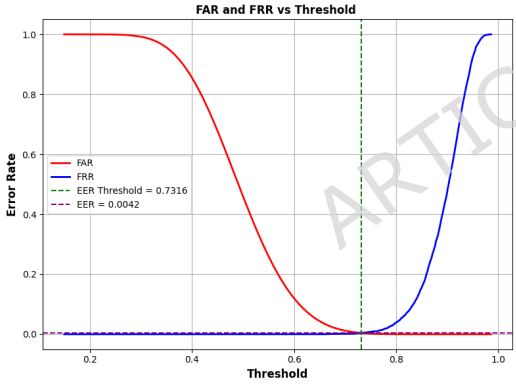
(b) Face+L1



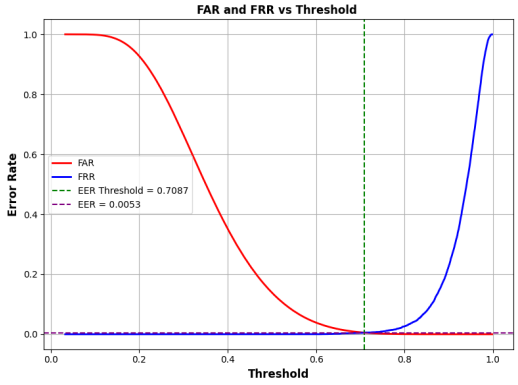
(c) Face+L2



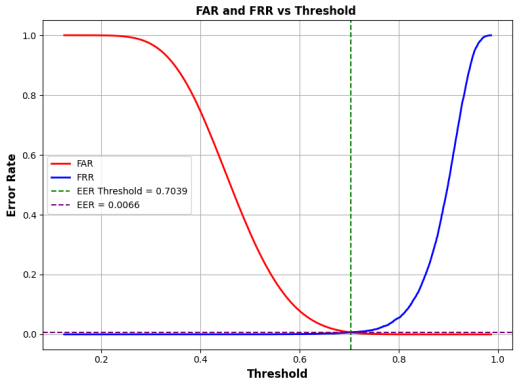
(d) Face+L3



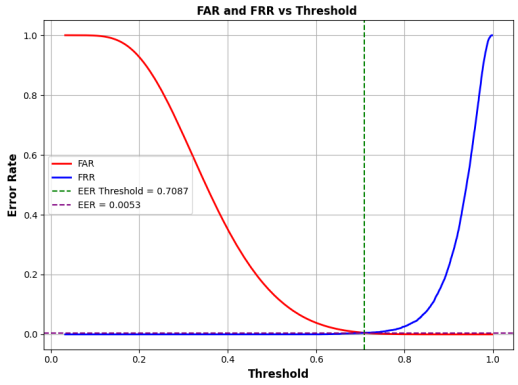
(e) Face+R0



(f) Face+R1



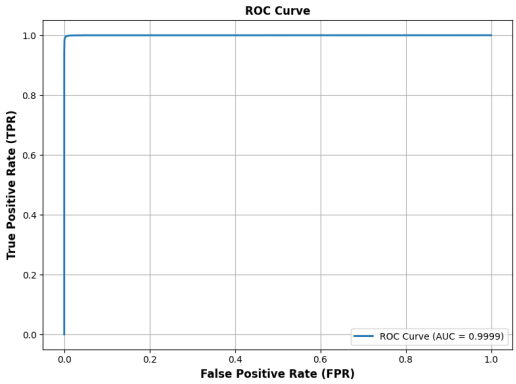
(g) Face+R2



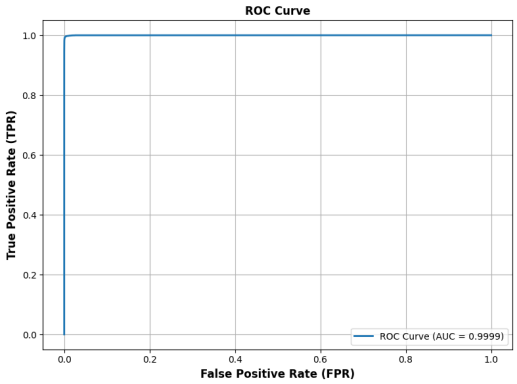
(h) Face+R3

Figure 8. FAR–FRR and EER curves for the proposed face–fingerprint verification system under different feature configurations.

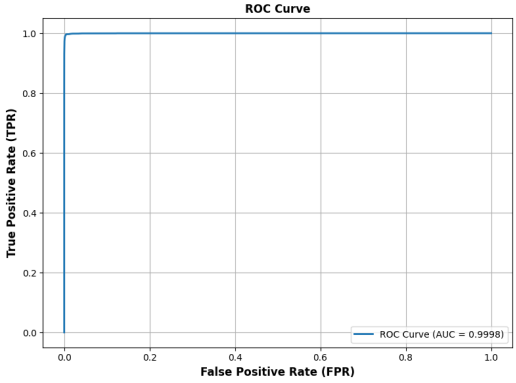
/



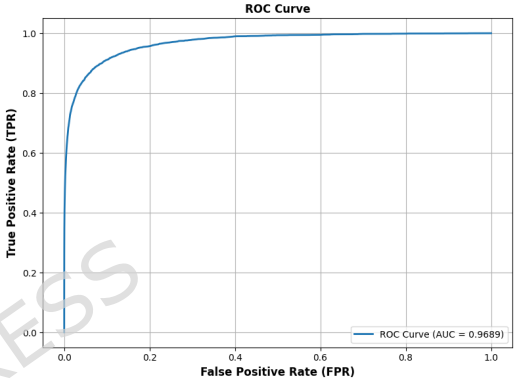
(a) Face+L0



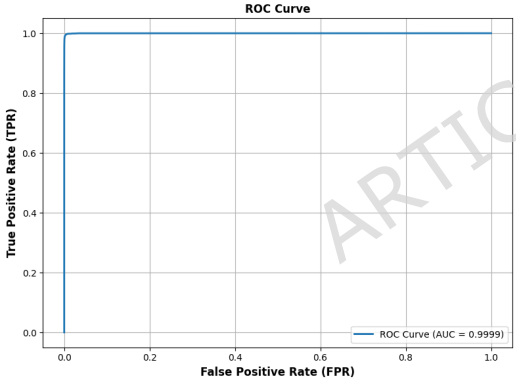
(b) Face+L1



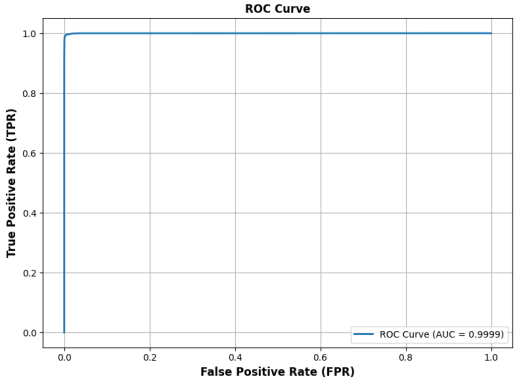
(c) Face+L2



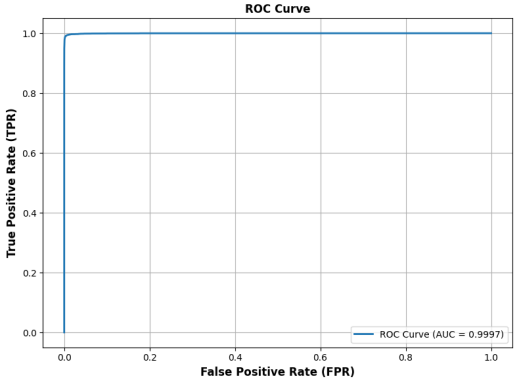
(d) Face+L3



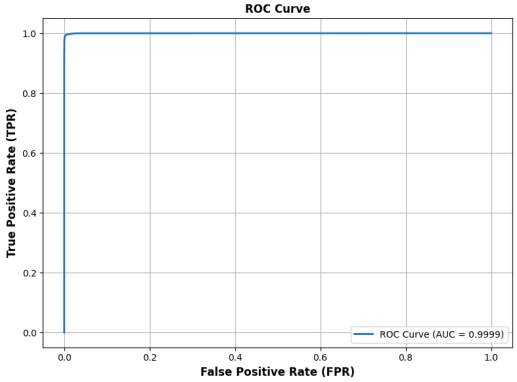
(e) Face+R0



(f) Face+R1



(g) Face+R2



(h) Face+R3

Figure 9. ROC curves of the proposed face–fingerprint verification system under different feature configurations.

/

Table 7. Comparison of multimodal biometric authentication systems in terms of modalities, fusion strategies, datasets, and reported EER.

Paper / Year	Modalities	Fusion Level	Datasets	EER (%)	Accuracy (%)
Aleem et al.[17] (2020)	Face, Fingerprint	Score-level	ORL, FVC2000, YALE	EER	99.59
Lee et al.[33] (2021)	Face, Fingerprint	Feature-level	LFW, FVC2002	0.24	–
Dang et al.[21] (2022)	Face, Fingerprint, Iris, Palmprint	Feature-level	LFW, CFPW,CASIA-FaceV5, IITD-E, Gait	2.53	–
Vallabhadas et al.[35] (2023)	Iris, Fingerprint	Feature-level	Casia v1+FVC2004 and Casia v3+FVC2006 and Children multimodal biometric dataset	0.032, 0.09, 0.015	–
Morampudi et al.[36] (2023)	Iris, Fingerprint	Feature-level	CASIA-Iris, FVC, CMDB	0.16 and 0.24	–
Vallabhadas et al.[18] (2024)	Fingerprint, Iris	Feature-level	CASIA-Iris, FVC, CMDB	0.038 and 0.073	
Kazi et al.[26] (2024)	Face, Fingerprint, Signature	Score- and decision-level fusion	YALE, FVC2002 and KVKR	–	99.03
Batouche et al.[27] (2025)	Face, Fingerprint	Feature-level	Face Recognition dataset, Sokoto Conventry	–	98.92
Li et al.[19] (2025)	Fingerprint, Fingervein	Feature-level	FVC2002, FVC2004, Saint Deem FingerVein dataset	0.01	–
Sasikala et al.[20] (2025)	Fingerprint, Retina	Feature-level	DRIVE and FVC2004	0.12	99.94
Jha et al.[37] (2025)	Face, Voice	Feature-level	Voxceleb1 and VidTIMIT	3.23 and 3.62	98.23 and 97.92
Ours	Face, Fingerprint	Feature-level using TAF	CASIA-FaceV5, CASIA-FingerprintV5	0.0038	99.61

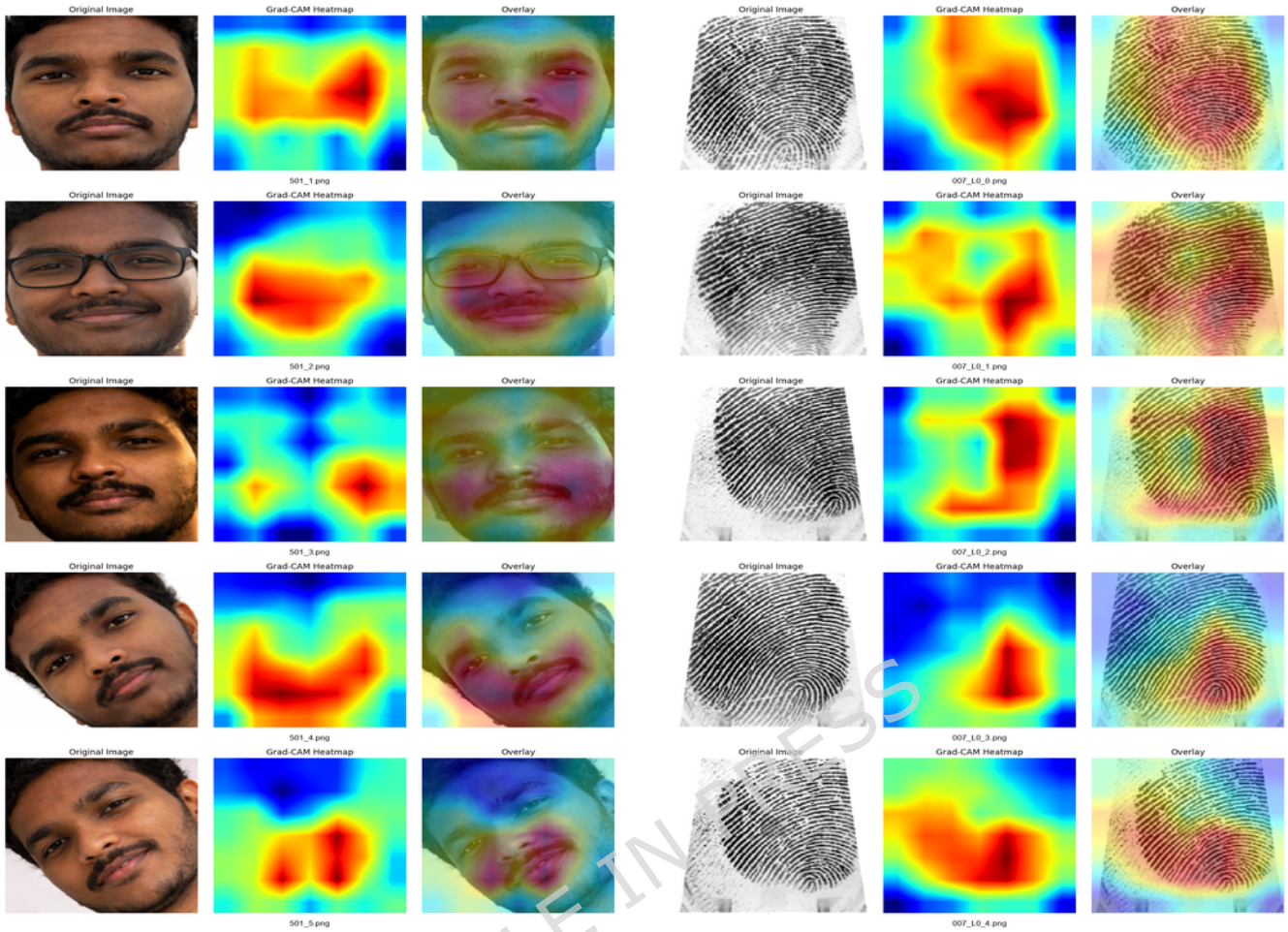
4.8 Grad-CAM Based Results

To evaluate the interpretability of the proposed system, Grad-CAM visualizations were generated for both face and fingerprint modalities. Figure 10 illustrate the image regions that contribute most significantly to the model's decisions.

For the face modality, the activated regions are consistently concentrated around the periocular area (eyes), nose bridge, and mouth region. It is known that these features can be helpful in identifying a person. The model does not pay much attention to hair and backgrounds. This demonstrates that it is not based on irrelevant information. In the case of fingerprint images, the model only concentrates on ridges and centre regions. It takes care of ridge ends and bifurcations. These are common features in fingerprint identification. The model treats background noise and image artefacts as irrelevant.

4.9 Ablation Study

In a typical MobileNetV2 network, every feature channel is treated equally, even though not all of them are equally useful. Some channels may capture apparent and discriminative features, while others may carry weak or noisy information. This limitation is reflected in the baseline configuration, which yields relatively high EERs for face (27.56%) and fingerprint (16.64%). The results of the ablation study for all configurations—including baseline, channel attention, feature fusion, proposed TAF, and encrypted matching—are summarized in Table. 8. When channel attention (CA) is applied, the network learns to emphasize informative channels and suppress less relevant ones. This is achieved by passing the feature maps through a lightweight module consisting of global average pooling, a couple of fully connected layers with non-linear activations, and a final sigmoid layer that produces channel-wise weights. The CA module allows the model to automatically pay attention to discriminative features. Consequently, there is a large decrease in the EER value upon attention to individual modalities: 0.61% face and 9.91% fingerprint. The results are obtained by simply fusing the two modalities, resulting in an EER of 0.3105%. This shows that modality complementation information should be used to improve recognition accuracy, but it does not explain the relative



(a) Face modality (b) Fingerprint modality

Figure 10. Grad-CAM visualizations illustrating discriminative regions for face and fingerprint modalities in the proposed multimodal biometric recognition system.

significance of the modalities.

Table 8. Ablation study results in terms of EER (%). Lower EER indicates better performance.

Modality	Face FE	Fingerprint FE	TAF	CKKS	EER (%)	A (%)	P(%)	R(%)	F1(%)
Face	MN	–	No	No	27.56	72.44	68.20	75.10	71.48
Face	MN	–	No	Yes	27.56	72.44	68.20	75.10	71.48
Fingerprint	–	MN	No	No	16.64	83.34	79.60	86.10	82.72
Fingerprint	–	MN	No	Yes	16.64	83.34	79.60	86.10	82.72
Face	MN+CA	–	No	No	0.61	98.75	96.80	99.10	97.94
Face	MN+CA	–	No	Yes	0.61	98.75	96.80	99.10	97.94
Fingerprint	–	MN+CA	No	No	9.91	96.35	94.10	97.80	95.91
Fingerprint	–	MN+CA	No	Yes	9.91	96.35	94.10	97.80	95.91
Face+Fingerprint	**MN+CA**	**MN+CA**	**Yes**	**No**	**0.0038**	**99.61**	**99.40**	**99.80**	**99.09**
Face+Fingerprint	**MN+CA**	**MN+CA**	**Yes**	**Yes**	**0.0038**	**99.61**	**99.40**	**99.80**	**99.09**

FE: Feature Extractor; *TAF*: Trust-Adaptive Fusion; *CKKS*: Cheon–Kim–Kim–Song homomorphic encryption scheme;
A: Accuracy; *P*:Precision; *R*:Recall; *F*:F1-Score

The proposed TAF, which combines channel attention and TAF strategy, attains an extremely low EER of 0.0038. This substantiates the fact that the most discriminative features of both modalities is captured by attention-guided features that

undergo adaptive fusion and significantly superior to other fusion methods based on a baseline or simple fusion. Lastly, the fused feature vector is encrypted in a CKKS homomorphic encryption to secure the privacy. All similarity calculations are done in the encrypted domain with the EER of 0.0038. This demonstrates the feasibility of privacy-preserving biometric matching without affecting the accuracy, and hence the framework is applicable in providing secure applications.

4.10 Security Analysis

The suggested multimodal biometric authentication system is specified with the understanding of security and privacy as the key concepts and a client-server architecture. The system is compliant to the ISO/IEC 24745 biometric information protection principles, where the acquisition, transmission, storage, and processing of the biometric data during the enrollment and verification processes are secure. Multi-stage protection provides template irreversibility. Face and fingerprint features of a discriminatory nature, obtained with the MN+CA model, are fused using a trust-adaptive feature-level fusion strategy that incorporates modality-specific confidence scores to create a concise and trustworthy representation. A fused feature vector is then encrypted by Fully Homomorphic Encryption (CKKS scheme) and stored so that no original biometric characteristics can be obtained from the secured templates.

Unlinkability is also facilitated by the fact that fused templates are stored in the Data server as encrypted versions and no direct connection is maintained between the stored templates and the actual biometric data. As the fusion is affected by the confidence and trust scores that can differ in different authentication cases, the resulting representations contribute to even a low potential of cross-matching between applications or databases. The template renewability is facilitated by the possibility to regenerate fused templates by changing fusion parameters or cryptographic keys, without re-enrolling the biometric data of the user.

The integrity of processing is also guaranteed by carrying out all similar operations wholly in the encrypted domain in the TA. Raw biometric samples, unencrypted features and decrypted templates are never revealed at any of the stages of communication or comparison. The design has gone a long way in averting threats like template leakage, replay attacks, and interception in transmission. System robustness-wise, the Trust-Adaptive Fusion mechanism helps in maximizing security by dynamically downplaying the impact of untrustworthy or undoubtedly unreliable modalities, restricting the impact of partial modality breaches or sensor-level issues. Besides, the centralized Trusted Authenticator offers controlled access to templates and decision logic, which minimizes the overall attack space. Trust-aware fusion and encrypted-domain matching connectemplates a security-guaranteed and privacy-assured multimodal biometric authentication system.

4.11 Computational Cost and Efficiency

The proposed system was tested on the performance in terms of computational speed to cover all key processes such as preprocessing, feature extraction, and encrypted-domain matching and is illustrated in the Table. 9 The system combines MobileNetV2 with channel attention to extract lightweight and discriminative features and generate 256-dimensional embedded features, which can be used in encrypted matching. The preprocessing pipeline involves face detection, face normalization and fingerprint enhancement. Face detection took an average of 398.34 ms with the MTCNN detector, which comprises the major part of the preprocessing latency. Further processes like cropping, resizing, and normalization were very efficient and took 0.32ms. The preprocessing of fingerprints such as denoising, thresholding, morphological enhancement, and ROI extraction, took 0.67 ms. These findings verify that the face detector is the computational bottleneck of preprocessing with all other processing steps incurring insignificant overheads.

The MobileNetV2 + Channel Attention model has 3.12 million parameters, and it consumes about 615 million FLOPS (about 300 million MACs). The model had an inference time average of 49.02 ms with an input resolution of $224 \times 224 \times 3$ and this illustrates the ability of the model to process biometrics in real-time. The lightweight system is such that most of the computation is done when detecting the face, as opposed to feature extraction. The 256-dimensional feature vectors were homomorphically encrypted by using the CKKS homomorphic encryption scheme to maintain the privacy. The encrypted-domain evaluation consisted of 3 stages: encryption, homomorphic computation, and decryption. The key and context setup time 0.279 s was measured separately as a one-time cost. Across 100 measurement runs, encryption required 13.8 ms, homomorphic cosine similarity computation required 93.4 ms, and decryption required 2.7 ms. Thus, the total cost of computing the encrypted-domain similarity was approximately 110 ms. The overhead introduced by encrypted computation compared to plaintext inference was:

$$\text{Overhead} = 110\text{ms} - 50.57\text{ms} \approx 59.4\text{ms},$$

indicating a reasonable increase, considering the strong privacy guarantees provided by CKKS.

The complete end-to-end runtime for multimodal biometric processing was also measured. The total enrollment time was 462.15ms, which includes preprocessing, feature extraction, and encryption. For the verification, the total runtime of 558.25ms.

Table 9. Computational time of the proposed system

Processing Stage	Time (ms)
Face Detection (MTCNN)	398.34
Face Preprocessing (Crop, Resize, Normalize)	0.32
Fingerprint Preprocessing	0.67
Feature Extraction (MobileNetV2 + CA)	49.02
Encryption (CKKS)	13.8
Homomorphic Computation	93.4
Decryption	2.7
Total Enrollment Time	**462.15 ms**
Total Verification Time	**558.25 ms**

Both enrollment and verification are completed well within one second, demonstrating that the system is suitable for practical deployment in secure real-time applications. The computational time is illustrated in the Table. 9

5 Conclusion

The proposed system is a secure, privacy-preserving, and explainable multimodal biometric authentication system that integrates face and fingerprint modalities. It employs MobileNetV2 with channel attention for robust feature extraction, a trust-adaptive fusion module that dynamically weights the contributions of each modality, and CKKS homomorphic encryption to ensure privacy during matching. The interpretability and transparency of the proposed system are enhanced with the help of Grad-CAM visualisation. While experiments on CASIA-FaceV5 and CASIA-FingerprintV5 have good performance. The EER of the system is remarkably low (0.0038), which means that the system demonstrates high discriminative performance in terms of separation of the genuine and impostors. Encrypted-domain matching is stronger in the protection of the template confidentiality, with a lower biometric template privacy threat, and without loss in recognition accuracy. Despite these benefits, the defenses against presentation attacks (spoofing), which include printed face images or created fake fingerprints, are explicitly not used or tested in the current work. The proposed system demonstrates potential for real-world applications, though further validation in unconstrained environments is required. Further evaluation on larger and more diverse datasets is required to confirm generalizability. Although multimodal fusion might enhance resilience through the necessity of achieving successful spoofing across multiple traits, explicit presentation attack detection (PAD) mechanisms are also a promising field to study in the future. Biometric templates can also be implemented that can be cancelled and replaced by the user in case their biometric information is leaked. The possibility of adaptive encryption strategies can be examined to change the level of security depending on the capability of the device used or the necessity of the application. It can also be noted that future research also looks into the use of quantum-safe encryption techniques to ensure that the system is immune to future quantum threats. It is possible to improve the explainability of the system by integrating other XAI methods, including SHAP or LIME, to gain a clearer idea as to the decisions made by the system.

References

1. Gawande, U. & Golhar, Y. Biometric security system: a rigorous review of unimodal and multimodal biometrics techniques. *Int. J. Biom.* **10**, 142–175 (2018).
2. Wild, P., Radu, P., Chen, L. & Ferryman, J. Robust multimodal face and fingerprint fusion in the presence of spoofing attacks. *Pattern Recognit.* **50**, 17–25 (2016).
3. Trigueros, D. S., Meng, L. & Hartnett, M. Face recognition: From traditional to deep learning methods. *arXiv preprint arXiv:1811.00116* (2018).
4. Kumar, T. A. & Ilango, S. Robust forgery detection via ensemble methods using intuitionistic fuzzy lbp and sift features. *Int. J. Comput. Appl.* **0**, 1–13, DOI: 10.1080/1206212X.2025.2469909 (2025). https://doi.org/10.1080/1206212X.2025.2469909.
5. Ametefe, D. S. *et al.* Enhancing fingerprint authentication: a systematic review of liveness detection methods against presentation attacks. *J. The Inst. Eng. (India): Ser. B* **105**, 1451–1467 (2024).
6. Adjabi, I., Ouahabi, A., Benzaoui, A. & Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics* **9**, 1188 (2020).
7. Sundararajan, K. & Woodard, D. L. Deep learning for biometrics: A survey. *ACM Comput. Surv. (CSUR)* **51**, 1–34 (2018).

8. Zhao, Z. & Kumar, A. Improving periocular recognition by explicit attention to critical regions in deep neural network. *IEEE Transactions on Inf. Forensics Secur.* **13**, 2937–2952 (2018).

9. Acar, A., Aksu, H., Uluagac, A. S. & Conti, M. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv. (Csur)* **51**, 1–35 (2018).

10. Rivest, R. L. Cryptography and machine learning. In *International Conference on the Theory and Application of Cryptology*, 427–439 (Springer, 1991).

11. Tourky, D., ElKawkagy, M. & Keshk, A. Homomorphic encryption the "holy grail" of cryptography. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 196–201 (IEEE, 2016).

12. Chander, B., John, C., Warrier, L. & Gopalakrishnan, K. Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. *ACM Comput. Surv.* **57**, 1–49 (2025).

13. Barni, M., Droandi, G., Lazzeretti, R. & Pignata, T. Semba: secure multi-biometric authentication. *IET Biom.* **8**, 411–421 (2019).

14. Walia, G. S., Singh, T., Singh, K. & Verma, N. Robust multimodal biometric system based on optimal score level fusion model. *Expert. Syst. with Appl.* **116**, 364–376 (2019).

15. Dwivedi, R. & Dey, S. Score-level fusion for cancelable multi-biometric verification. *Pattern Recognit. Lett.* **126**, 58–67 (2019).

16. Rathgeb, C., Gomez-Barrero, M., Busch, C., Galbally, J. & Fierrez, J. Towards cancelable multi-biometrics based on bloom filters: a case study on feature level fusion of face and iris. In *3rd international workshop on biometrics and forensics (IWBF 2015)*, 1–6 (IEEE, 2015).

17. Aleem, S., Yang, P., Masood, S., Li, P. & Sheng, B. An accurate multi-modal biometric identification system for person identification via fusion of face and finger print. *World Wide Web* **23**, 1299–1317 (2020).

18. Vallabhadas, D. K., Sandhya, M., Reddy, S. D., Satwika, D. & Prashanth, G. L. Biometric template protection based on a cancelable convolutional neural network over iris and fingerprint. *Biomed. Signal Process. Control.* **91**, 106006 (2024).

19. Li, Y. *et al.* A cancelable multi-biometric system based on the feature-level fusion of fingerprint and finger vein. *Multimed. Tools Appl.* **84**, 24765–24787 (2025).

20. Sasikala, T. Multimodal secure biometrics using attention efficient-net hash compression framework. *Digit. Signal Process.* **160**, 105018 (2025).

21. Dang, T. M. *et al.* Avet: A novel transform function to improve cancellable biometrics security. *IEEE transactions on information forensics security* **18**, 758–772 (2022).

22. Purohit, H. & Ajmera, P. K. Optimal feature level fusion for secured human authentication in multimodal biometric system. *Mach. Vis. Appl.* **32**, 24, DOI: https://doi.org/10.1007/s00138-020-01146-6 (2021).

23. Vijay, M. & Indumathi, G. Deep belief network-based hybrid model for multimodal biometric system for futuristic security applications. *J. Inf. Secur. Appl.* **58**, 102707, DOI: https://doi.org/10.1016/j.jisa.2020.102707 (2021).

24. mehdi Cherrat, E., Alaoui, R. & Bouzahir, H. Convolutional neural networks approach for multimodal biometric identification system using the fusion of fingerprint, finger-vein and face images. *PeerJ Comput. Sci.* **6**, e248 (2020).

25. Mwaura, G. W., Mwangi, W. & Otieno, C. Multimodal biometric system: Fusion of face and fingerprint biometrics at match score fusion level. *Int. J. Sci. & Technol. Res.* (2017).

26. Kazi, M. *et al.* Face, fingerprint, and signature based multimodal biometric system using score level and decision level fusion approaches. *IETE J. Res.* **70**, 3703–3722, DOI: https://doi.org/10.1080/03772063.2023.2217784 (2024).

27. Batouche, A., Meshoul, S., Shaiba, H. & Batouche, M. A novel approach to enhanced cancelable multi-biometrics personal identification based on incremental deep learning. *Comput. Mater. & Continua* **83** (2025).

28. Zhou, Z., Liu, Y., Zhu, X., Zhang, S. & Liu, Z. Privacy-preserving cancelable multi-biometrics for identity information management. *Inf. Process. & Manag.* **62**, 103869 (2025).

29. Zhao, G., Jiang, Q., Wang, D., Ma, X. & Li, X. Deep hashing based cancelable multi-biometric template protection. *IEEE Transactions on Dependable Secur. Comput.* **21**, 3751–3767 (2023).

30. Naeem, E. A. *et al.* Efficient cancelable authentication system based on drpe and adaptive filter. *Multimed. Tools Appl.* **83**, 76131–76175 (2024).

31. Elsheikh, A. G. *et al.* Application of mace filter with drpe for cancelable biometric authentication. *J. Opt.* **53**, 101–116 (2024).

32. Wang, Y., Shi, D. & Zhou, W. Convolutional neural network approach based on multimodal biometric system with fusion of face and finger vein features. *Sensors* **22**, 6039 (2022).

33. Lee, M. J., Teoh, A. B. J., Uhl, A., Liang, S.-N. & Jin, Z. A tokenless cancellable scheme for multimodal biometric systems. *Comput. & Secur.* **108**, 102350 (2021).

34. Kim, J., Jung, Y. G. & Teoh, A. B. J. Multimodal biometric template protection based on a cancelable softmaxout fusion network. *Appl. Sci.* **12**, 2023 (2022).

35. Vallabhadas, D. K. & Sandhya, M. Cancelable bimodal shell using fingerprint and iris. *J. Electron. Imaging* **32**, 063027–063027 (2023).

36. Morampudi, M. K., Sandhya, M. & Dileep, M. Privacy-preserving bimodal authentication system using fan-vercauteren scheme. *Optik* **274**, 170515 (2023).

37. Jha, K., Jain, A. & Srivastava, S. Multimodal biometric authentication system leveraging optimally trained ensemble classifier using feature-level fusion. *Technol. Heal. Care* 09287329251363424 (2025).

38. A. El_Rahman, S. & Alluhaidan, A. S. Enhanced multimodal biometric recognition systems based on deep learning and traditional methods in smart environments. *Plos one* **19**, e0291084, DOI: 10.1371/journal.pone.0291084 (2024).

39. Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* **23**, 1499–1503 (2016).

40. Jalali, S., Boostani, R. & Mohammadi, M. Efficient fingerprint features for gender recognition. *Multidimens. Syst. Signal Process.* **33**, 81–97 (2022).

41. Ahmadyfard, A. A comprehensive survey of channel attention mechanisms in single image super-resolution. *J. Electr. Syst.* **20**, 9571–9583 (2024).

42. Guo, M.-H. *et al.* Attention mechanisms in computer vision: A survey. *Comput. visual media* **8**, 331–368 (2022).

43. Institute of Automation, Chinese Academy of Sciences (CASIA). Casia-facev5 dataset. http://www.idealtest.org/ (2010).

44. Institute of Automation, Chinese Academy of Sciences (CASIA). Casia-fingerprintv5 dataset. http://www.idealtest.org/ (2010).

45. Selvaraj, A., Russel, N. S. & Seenivasan, M. Robust penta-modal biometric identification through deep learning and weighted score fusion. *Iran J. Comput. Sci.* **8**, 553–569 (2025).

46. Es-Sobbahi, H., Radouane, M. & Nafil, K. Multimodal biometrics: A review of handcrafted and ai–based fusion approaches. *IET Biom.* **2025**, 5055434 (2025).

Declarations

Competing Interests

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.

Authors' Contribution Statement

This study was carried out in collaboration between the authors. **Pavani Chitrapu** was primarily responsible for the conception and design of the study, data collection, analysis, and manuscript drafting.**Hemantha Kumar Kalluri** provided critical revisions and intellectual guidance throughout the research process and supervised the findings of this work. **Mahesh Kumar Morampudi** provided critical revisions and intellectual guidance throughout the research process and supervised the findings of this work.

Data Availability Statement

The datasets used in this study are publicly available biometric datasets. The CASIA-Face V5 dataset is publicly available via Figshare at `https://doi.org/10.6084/m9.figshare.26509591`. The CASIA-Fingerprint V5 dataset is publicly available from the Institute of Automation, Chinese Academy of Sciences at `http://english.ia.cas.cn/rs/sd/201611/t20161123_170932.html`.

Funding

The Authors received NO FUNDING for this work

Ethical Approval

This study did not involve direct experiments on human participants; therefore, ethical approval was not required. All datasets used are publicly available.

Consent to Publish

All human images shown in this manuscript are taken from publicly available datasets. The respective dataset providers obtained consent for publication.