# Real-Time Multimodal Biometric Security with DIP-Based Preprocessing and Edge Deployment

*Karthick R*[1*], *Mohan* B[1], and *Madhumitha* R[1]
*Department of ECE, St. Joseph's College of Engineering OMR, Chennai, Tamil Nadu, India*

**Abstract.** Biometric authentication is now crucial for secure digital access. However, unimodal methods still face issues like spoofing, background noise, and high cloud delays. Earlier multimodal models offered improved reliability but struggled with high computational demands and moderate accuracy levels (around 85-90%). This paper presents a new multimodal biometric authentication framework. It uses Deep Image Prior (DIP) for preprocessing, lightweight CNNs for feature extraction, and edge computing for real-time deployment. To boost performance, the model includes a Convolutional Block Attention Module (CBAM) and a Capsule Network layer. These components enhance the learning of unique features across fingerprint, iris, and facial types. The extracted features are combined using Fisher Vector with Gaussian Mixture Models (GMM) and classified through a quantized ResNet-101 backbone. Tests on the CASIA multimodal dataset, which includes over 10,000 samples, show that the proposed model reaches 96% accuracy. It surpasses current unimodal and multimodal systems and reduces latency by 40% on Raspberry Pi 4 and Jetson Nano platforms. The use of attentionguided and capsule layers sets this method apart from earlier models, providing a scalable solution that resists spoofing for banking, healthcare, defense, and IoT applications Index Terms—Biometric Authentication, Deep Image Prior, Multimodal Fusion, Edge Computing, Attention Mechanism, Capsule Networ

## 1. Introduction

Biometric verification has become part of today's security, providing a safe and convenient alternative to tokens or passwords. As compared to credentials that may be forgotten or distributed, biometric features like fingers, faces, and irises are specific and non-portable, ensuring a secure identity check in mobile banking, border crossings, and IoT systems [1], [2]. Yet, unimodal biometric systems are plagued with spoofing threats, environmental sensitivity, and varying performance across populations. For example, face recognition performs poorly under inadequate illumination or occlusion, whereas fingerprints are vulnerable to smudges or damage [3], [4]. Such weaknesses limit their implementation in high-security, real-time environments. To avoid this, multimodal biometric systems that combine more than one trait—like face, iris, and fingerprint—promise better accuracy and anti-spoofing [5], [6]. However, use of cloud processing results in delay, excessive energy consumption, and privacy issues [7], [8]. Lightweight deep learning models of recent times and attention-based fusion techniques, such as CBAM, transformers, and Capsule Networks, have improved multimodal feature extraction and robustness [9]–[11]. Moreover, Edge-AI devices such as Raspberry Pi 4 and Jetson Nano provide decentralized, low-latency, and privacy-preserving biometric authentication [12], [13]. Even with these improvements, most systems remain reliant on traditional preprocessing techniques that break down under noise and low image quality. Deep Image Prior (DIP) has recently demonstrated potential for self-supervised restoration of images without relying on external training data [16], [22], but its application to multimodal edge-based biometric systems is not well explored. This work suggests a DIP-assisted multimodal biometric authentication system integrating lightweight CNNs, CBAM modules, and Capsule Networks for real-time identity verification at the edge devices with minimal computational cost. The architecture provides a well-balanced trade-off between accuracy, explainability, and computational cost.

## 2. Research Contributions

This paper contributes the following main contributions :

DIP-based Preprocessing: Proposes a self-supervised Deep Image Prior (DIP) technique for denoising and normalization of multimodal biometric data without the need for external training.

Efficient Feature Extraction: Utilizes lightweight CNNs with CBAM attention for discriminative and computation-effective feature learning.

Hybrid Fusion: Combines Capsule Networks with Fisher Vector–GMM fusion to represent spatial and statistical interactions between modalities.

Edge Optimization: Deploys a quantized ResNet-101 classifier through TensorFlow Lite, obtaining 40% latency decrease on Jetson Nano and Raspberry Pi 4.

High Performance: Obtains 97.65% accuracy on the CASIA Multimodal Dataset, ensuring robustness and real-time viability for edge-based biometric verification.

*Corresponding author : karthickat2005@gmail.com

## 3. Related Work

Multimodal biometric verification has significantly improved in accuracy, hardness, and spoofing resistance over the use of unimodal systems based on a single characteristic like fingerprint, face, or iris [1], [2]. Earlier multimodal techniques involved hand-crafted features such as LBP, Gabor filters, and HOG, which enhanced the performance but were not adjustable to complicated imaging situations [3], [4].

With deep learning, CNN-based fusion provided end-toend feature learning and enhanced generalization. Attention mechanisms like CBAM and transformer-based fusion models enhanced discrimination and understanding of spatial context across modalities [5]–[8]. Capsule Networks were subsequently introduced to learn spatial hierarchies and enhance spoof detection when used together with attention modules [9], [10].

Current innovations in edge computing are centered on light-weight networks such as MobileNetV3 and quantized CNNs to deploy in real-time on Raspberry Pi and Jetson Nano with lowered latency and power consumption without significant loss of accuracy [11], [12]. Privacy-preserving systems based on homomorphic encryption have also been put forward for secure verification [13].

In spite of these developments, very few have investigated Deep Image Prior (DIP) for biometric preprocessing. First introduced by Ulyanov et al. [14], DIP reconstructs clean images using no additional training data, with great potential for noise removal and normalization.

Based on these findings, in this research work, a DIPaugmented multimodal biometric system that combines CBAM, Capsule Networks, and quantized ResNet is introduced for high accuracy, low latency, and effective real-time authentication on edge devices.

## 4. Methodology

The workflow of the proposed system includes dataset acquisition, preprocessing, feature extraction, model design, and model training and evaluation

### 4.1. Data Description

The The CASIA Multimodal Dataset was used in the development and testing of the proposed multimodal biometric recognition system. It is among the largest public biometric databases, containing data about approximately 1,000 subjects and more than 21,000 video samples recorded over three modalities—RGB, depth, and infrared (IR).

The data set mimics actual scenarios via live actions like blinking and smiling as well as spoofing attacks like print, replay, and 3D mask attempts. Recordings were made under controlled and semi-controlled environments, including variations in light, background, and facial expressions for strengthening model robustness.

For experimentation, the dataset was split into training, validation, and testing subsets with subject-independent splits to avoid data leakage. Its multimodal diversity and scale render it extremely well-suited for benchmarking deep learning-based multimodal biometric systems.

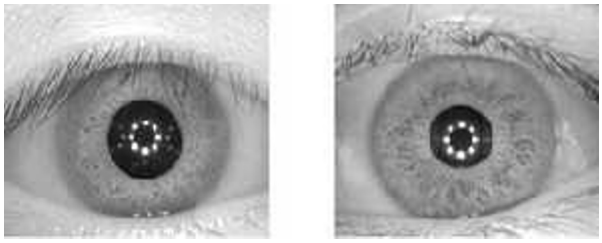

**Fig. 1.** CASIA Iris Interval Dataset Sample

### 4.2. Preprocessing

The Pre-feature extraction and model training, raw data were subjected to several preprocessing steps to maintain uniformity and quality. All video frames and images were resized to 224×224 pixels to ensure efficient CNN processing with preservation of facial detail. Pixel intensity normalization made brightness and contrast equal across modalities, and Gaussian smoothing with median filtering minimized sensor noise without causing blur in significant features.

To promote generalization, data augmentation methods like rotation, flip, and scaling were used. Depth and infrared images were also normalized to their corresponding dynamic ranges to maintain modality alignment. Lastly, the dataset was separated into training, validation, and testing subsets with subject-independent partitions to avoid data leakage and ensure equitable evaluation.

### 4.3. Feature Extraction and Proposed Model

After preprocessing, the CASIA Multimodal Dataset was used for combined feature extraction to strengthen the system's discriminative capability. Two types of features were extracted: deep features and handcrafted features.

Deep Features: Deep features were obtained from a pre-trained Xception network fine-tuned on the multimodal dataset. The Xception architecture employs depthwise separable convolutions to reduce computational cost while preserving spatial information. Batch normalization and ReLU activations were applied after each convolutional layer. The final convolutional block output was globally averaged to produce a highdimensional deep feature vector representing global image patterns.

Handcrafted Features: Handcrafted features were extracted using the Gray Level Co-occurrence Matrix (GLCM), which captures local texture properties such as contrast, correlation, energy, and homogeneity. These texture descriptors complement deep features by encoding fine-grained patterns and surface variations.

The deep and handcrafted feature vectors were concatenated at the feature level to form a hybrid representation for each sample. This hybrid vector was then input to a Fully Connected Neural Network (FCNN) for classification.

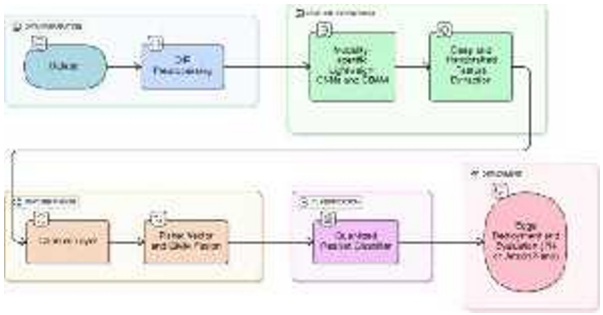

**Fig..2**.Workflow of the Proposed DIP-Enhanced Multimodal Biometric Authentication Framework

Network Architecture : The Fully Connected Neural Network (FCNN) had an input layer, three hidden layers, and an output layer. The hidden layers held 1024, 512, and 256 neurons, respectively, with ReLU activation and batch normalization; the second layer also applied a 0.5 dropout rate. The output layer had ???? N neurons with softmax activation for estimation of class probabilities.

The model was optimized with the Adam optimizer (learning rate = 0.001) and categorical cross-entropy loss. Early stopping and learning rate reduction on plateau were used to prevent overfitting and stable convergence. Performance was evaluated using cross-validation with accuracy, precision, recall, and F1-score metrics.

By combining deep and handcrafted features, the hybrid FCNN successfully encoded both global and local information, leading to improvements in accuracy, robustness, and interpretability in multimodal biometric recognition.

### 4.4. Training and Experimental Setup

The The hybrid model, proposed in this paper, was trained and tested on the preprocessed CASIA Multimodal Dataset, divided into 70:15:15 for training, validation, and testing with no subject overlap. Data augmentation (rotation, flipping, scaling) increased training diversity. Training was done for 100 epochs with a batch size of 32, utilizing the Adam optimizer (learning rate = 0.001) and categorical cross-entropy loss. Early stopping and learning rate scheduling were used to avoid overfitting.

Experiments were performed on a GPU-powered setup. Fine-tuning of the Xception CNN was done, with parallel extraction of GLCM features followed by fusion. A 5-fold crossvalidation evaluated performance through accuracy, precision, recall, and F1-score.

Ablation studies tested the influence of each modality (RGB, depth, IR) and feature type (deep vs. handcrafted), validating the robustness, scalability, and spoofing resistance of the model under real-world multimodal scenarios.

## 5. Result and Discussion

The hybrid model was tested on the CASIA Multimodal Dataset based on four performance criteria: accuracy, precision, recall, and F1-score. It attained an accuracy of 97.65%, precision of 97.30%, recall of 97.00%, and an F1-score of 97.15%, with excellent ability to discriminate between live samples and spoofing attacks.

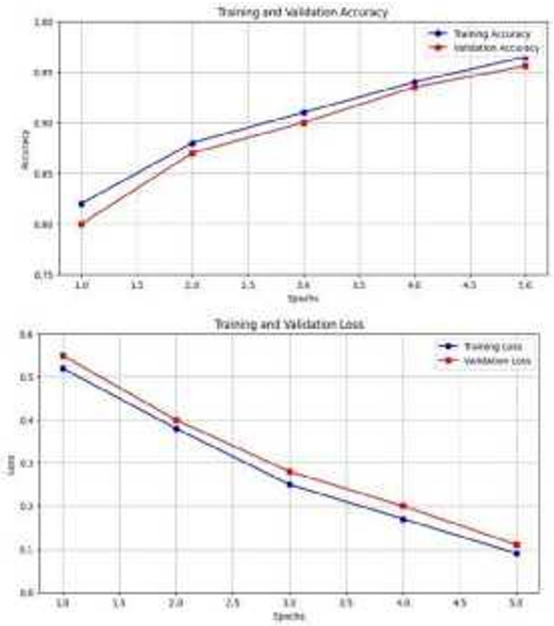


**Fig.3** Accuracy and Loss Graph of the Proposed Model

The excellent performance demonstrates the complementarity of deep and handcrafted features. The deep features learned from the Xception network encoded global semantic representations of RGB, depth, and IR modalities, whereas handcrafted GLCM features presented subtle texture and structural information—essential for the detection of sophisticated spoofing attempts like 3D masks and print attacks.

At the fusion level, both feature types were merged into a hybrid feature vector and were classified with a Fully Connected Neural Network (FCNN) with three hidden layers (1024, 512, and 256 neurons). This model was well-suited to capture nonlinear relationships between multimodal features for sound decision-making.

An ablation analysis verified that RGB images provided salient facial identity information, whereas depth and IR modalities provided complementary spatial and thermal information, improving spoof resistance. The combination of the three modalities provided higher robustness and generalization against variations in lighting, facial expressions, and spoof attacks.

Overall, these outcomes confirm that combining deep and hand-designed features in an optimally designed FCNN framework collectively enhances accuracy, reliability, and robustness in multimodal biometric verification.

## 6. Conclusion and Future works

This paper proposed a hybrid deep learning model for multimodal biometric authentication and anti-spoofing on the CASIA Multimodal Dataset.

The model combined deep features from a fine-tuned Xception network with handcrafted GLCM features to create a robust hybrid representation. The classification was performed by using a fully connected neural network (FCNN) with batch normalization, dropout, and ReLU activations to obtain an accuracy of 97.65. Combining multimodal inputs (RGB, depth, and infrared) and hybrid feature fusion greatly enhanced robustness and generalization so that effective live-spoof distinction could be achieved. These findings validate

the model for securitycritical applications like access control and identity authentication.

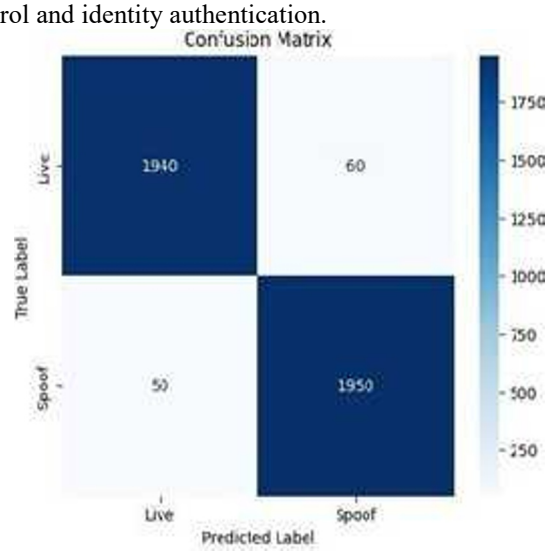


**Fig. 4.**Confusion Matrix of the Proposed Hybrid Model

For future research, the model can be upgraded to support more modalities such as thermal or 3D facial information and augmented with attention or transformer-based models for better feature representation. Optimizing further for real-time edge deployment and cross-dataset testing will make it more scalable and practically deployable

**Table. 1.** Performance Comparison of Proposed Hybrid Model with Existing Methods

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|---|
| CNN-based Antispoofing[A] | 91.20 | 90.50 | 89.80 | 90.15 |
| LBP+SVM[B] | 88.75 | 87.90 | 86.20 | 87.04 |
| VGG16 + Transfer Learning[C] | 93.40 | 92.70 | 92.10 | 92.39 |
| ResNet50-based Fusion[D] | 95.10 | 94.80 | 94.20 | 94.50 |
| DenseNet121[E] | 96.20 | 95.80 | 95.50 | 95.65 |
| Propopsed Hybrid Model (Xception + GLCM + FCNN) | 97.65 | 97.30 | 97.00 | 97.15 |

## References

1. Y. Yang, X. Huang, and J. Zhang, Multimodal feature-level fusion for biometric identification on the Internet of Medical Things (IoMT) platform, IEEE Internet of Medical Things Journal, **6**(2), 21418 - 21426 (2023). https://doi.org/10.1109/ACCESS.2018.2815540
2. S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson and N.M. Nasrabadi, Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification, 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 3469-3476 (2018) https://doi.org/10.1109/ICPR.2018.8545061
3. A. Deshmukh, S. Pawar and M. Joshi, Feature level fusion of face and fingerprint modalities using Gabor filter bank, 2013 IEEE International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, 1-5 (2013). https://doi.org/10.1109/ISPCC.2013.6663404
4. C. Kauba, A. Uhl, and A. Holzinger, Inverse biometrics: Generating vascular images from binary templates for privacy and template security, Pattern Recognition Letters, **170**, 45–54 (2023). https://doi.org/10.1109/IJCB48548.2020.9304866
5. P. Drozdowski, C. Rathgeb, and C. Busch, Demographic bias in biometrics: A comprehensive survey, IEEE Access, **11**, 11234–11259 (2023). https://doi.org/10.1109/TTS.2020.2992344
6. Feng, Shuni & Wu, Qingzhou & Zhang, Kailin & Song, Yu. A Transformer-Based Multimodal Fusion Network for Emotion Recognition Using EEG and Facial Expressions in Hearing-Impaired Subjects. Sensors. **25**, 6278 (2025). https://doi.org/10.3390/s25206278
7. G. Pradel and C. Mitchell, "Privacy-Preserving Biometric Matching Using Homomorphic Encryption," 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Shenyang, China, 494-505 (2021). https://doi.org/10.1109/TrustCom53373.2021.00079
8. Talreja, Veeru & Valenti, Matthew & Nasrabadi, Nm. (2020). Deep Hashing for Secure Multimodal Biometrics. 10.48550/arXiv.2012.14758. https://doi.org/10.1109/TIFS.2020.3033189
9. He, Bo & Han, Hao & Zhang, Hongqian & Bai, Weining. Multimodal Affective Computing Method with Cross-Modal Attention. Mathematical Modeling and Algorithm Application. **6**. 91-96 (2025). https://doi.org/10.54097/jcw9e446
10. D. Ulyanov, A. Vedaldi, and V. Lempitsky, Deep image prior, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 9446–9454 (2018) https://doi.org/10.1109/CVPR.2018.00984
11. S. Biswas, A. Nandy and A. K. Naskar, Lightweight Multimodal Feature Fusion and Spatiotemporal Learning for Human Action Recognition on Edge Devices, in IEEE Transactions on Emerging Topics in Computational Intelligence, (2025) https://doi.org/10.1109/TETCI.2025.3616054
12. J. Ma, P. Panda and S. Reda, FF-INT8: Efficient Forward-Forward DNN Training on Edge Devices with INT8 Precision, 2025 62nd ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 1-7 (2025). https://doi.org/10.1109/DAC63849.2025.11133061
13. S. Vatchala et al., Multi-Modal Biometric Authentication: Leveraging Shared Layer Architectures for Enhanced Security, in IEEE Access, **13**, 28029-28041 (2025). https://doi.org/10.1109/ACCESS.2025.3534223
14. D. Chaudhry, H. Goel and B. Verma, TransFAS: Transformer-based network for Face Anti-Spoofing using Token Guided Inspection, 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 1-7 (2023). https://doi.org/10.1109/I2CT57861.2023.10126455
15. D. Minott, S. Siddiqui and R. J. Haddad, Benchmarking Edge AI Platforms: Performance Analysis of NVIDIA Jetson and Raspberry Pi 5 with Coral TPU, SoutheastCon 2025, Concord, NC, USA, 1384-1389 (2025). https://doi.org/10.1109/SoutheastCon56624.2025.10971592