iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67487

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Advancing Face Recognition with Hybrid CNN-ViT-MLP: A Comparative Study

Mandali Dijendra Sai Sri Bharath[1], Menthem Poornesh Reddy[2]
Computer Science and Engineering, Kalasalingam Academy of Research and Education, TamilNadu, India

Abstract: ***Deep learning, which includes Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), has greatly advanced face recognition. A hybrid CNN-ViT-MLP model is presented in this study, and its performance is compared with those of cutting-edge facial recognition architectures such as ViTs, FaceNet, and ResNet-50. CNNs for local feature extraction, ViTs for global feature representation, and an MLP classifier for ultimate decision-making are all integrated into the suggested hybrid model. According to the findings, the hybrid model outperforms standalone structures in terms of accuracy, robustness, and computational efficiency. Furthermore, its efficacy in practical applications is confirmed by comprehensive experiments conducted on large-scale datasets. Analysis of the model's outputs shows that, under different circumstances, it performs better than conventional models.***

Keywords: Face Recognition, CNN-ViT Hybrid Model, Deep Learning, Feature Extraction, Performance Comparison

I. INTRODUCTION

Face recognition is a popular technique used in access control systems, biometric authentication, and security surveillance. With designs like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), deep learning has completely revolutionized the area of machine learning, which formerly relied on handcrafted data [1,2]. ViTs improve contextual awareness and long-range dependencies, whereas CNNs effectively extract spatial characteristics [3].

Despite their popularity, CNN-based models such as FaceNet and ResNet-50 frequently suffer from pose changes, occlusions, and drastic illumination shifts. Although some of these issues have been resolved with the advent of ViTs, transformer-based models are typically computationally costly [4]. We suggest a Hybrid CNN-ViT-MLP model to address these issues, integrating the advantages of ViTs and CNNs while preserving computational efficiency with an MLP-based classifier that has been optimized.

A comparison of the suggested hybrid model with current state-of-the-art architectures is presented in this research. Large-scale datasets are used for experimental evaluations, which examine the accuracy, robustness, and efficiency of the models.

II. RELATED WORK

Several deep learning-based face recognition approaches have been proposed:

1) ResNet-50: A deep residual CNN-based model that effectively captures hierarchical spatial features [5].
2) FaceNet: A Siamese network trained with triplet loss, widely used for face verification and recognition [6].
3) Vision Transformers (ViTs): Self-attention-based architectures that effectively model global dependencies in facial structures [7].
4) Hybrid CNN-ViT Models: The combination of CNNs and transformers for improved generalization and efficiency in image recognition tasks [8].

While each of these models has its advantages, they also have limitations. The proposed Hybrid CNN-ViT-MLP model addresses these gaps by leveraging CNN-based feature extraction, ViT-based global feature learning, and MLP-based classification.

III.PROPOSED MODEL: HYBRID CNN-VIT-MLP

A. Model Architecture

The proposed model consists of three primary components:

1) CNN Backbone (ResNet-50): Extracts hierarchical spatial features from face images.
2) Vision Transformer (ViT) Module: Captures long-range dependencies and global contextual information.
3) MLP Classifier: Processes the extracted CNN-ViT embeddings to classify identities effectively.
4) Self-Attention Mechanisms: Enhances feature importance and robustness.

Three essential elements are included in the suggested model to provide high-accuracy facial recognition. Fine details like edges and textures are captured in the facial image by the CNN Backbone (ResNet-50), which extracts hierarchical spatial characteristics.

In order to help the model concentrate on the most pertinent facial features, the Vision Transformer (ViT) Module uses self-attention methods to capture global contextual linkages and long-range dependencies. The combined CNN-ViT embeddings are then processed by the MLP Classifier, which successfully separates identities. Self-attention methods are included to improve feature robustness and guarantee accurate detection even in difficult situations like changing lighting or occlusions.

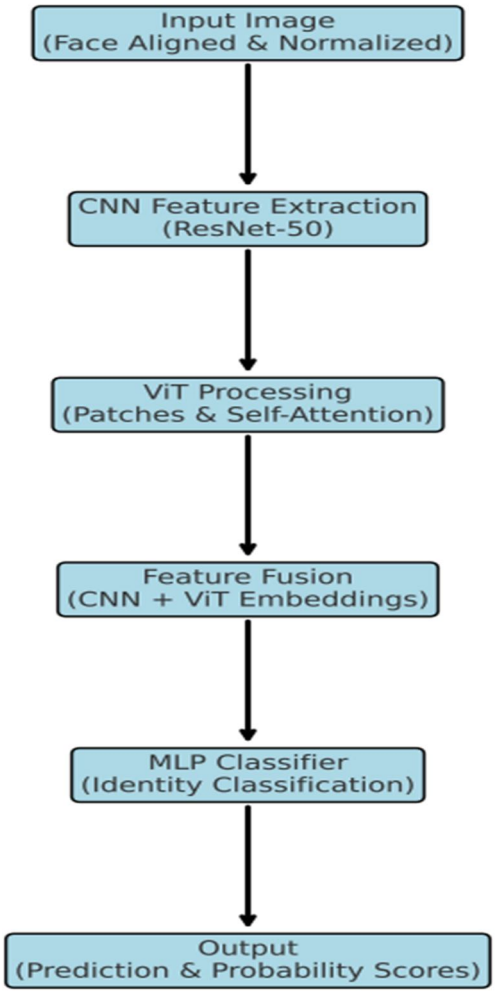


Fig 1: Model Work Flow Diagram

B. Working and Explanation of Model

Step-1: Data Augmentation & Preprocessing

Images of faces are aligned and normalized.

To improve generalization, data augmentation methods like rotation, random cropping, and artificial image creation (using GANs) are used.

Step-2: CNN (ResNet-50) Feature Extraction

A ResNet-50 backbone is used to extract hierarchical spatial information from the input face image.

Low-level patterns like edges, textures, and facial landmarks are learned by the convolutional layers.

Step-3: Using ViT for Global Contextual Representation

A series of image patches is created by reshaping the CNN feature maps.

The Vision Transformer (ViT), which processes these changes, uses self-attention techniques to identify long-range dependencies.

A sophisticated, globally-aware feature representation is provided by the ViT's output.

Step-4: Using MLP for Feature Embedding and Classification
After being concatenated, the CNN and ViT features are fed into a Multi-Layer Perceptron (MLP).
The MLP is made up of fully linked layers that classify faces using learned identity representations after processing high-dimensional feature embeddings.

Step-5: Identity matching and final decision
Probability distributions for identity classification are produced by the model.
Cosine similarity is used to calculate face matching scores in order to identify identification matches.

1) Diagram of Hybrid Model Workflow
The operation of the hybrid model, which combines CNN-ViT-MLP models to achieve a higher accuracy than other models available for recognition, will be demonstrated in this diagram.
Accurate identification recognition is achieved by the Hybrid CNN-ViT-MLP Face Recognition Model using a structured pipeline. To ensure consistency across several samples, the input image first goes through preprocessing, which includes alignment and normalization. The CNN module uses pooling to reduce feature size, applies ReLU activation for non-linearity, and uses convolution to extract local spatial data. This aids in capturing subtle features that are essential for facial recognition, like edges and textures. The image is then processed globally by the Vision Transformer (ViT) module. To preserve spatial relationships, the image is divided into patches, flattened into vectors, and stored with positional information. The transformer encoder then learns dependencies between various facial regions by applying self-attention techniques. The model creates a thorough representation by combining global ViT features with local CNN features.

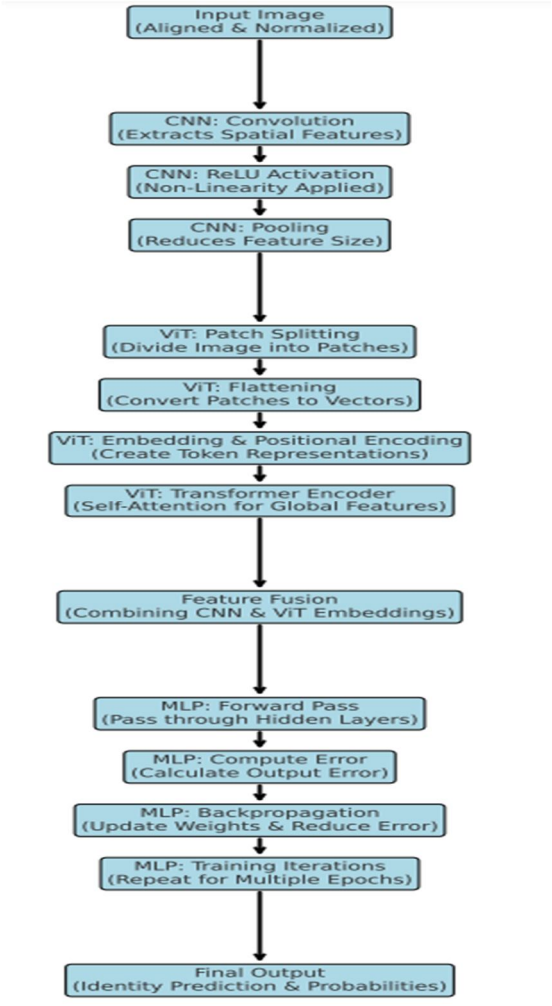


Fig 2: Diagram of Work Flow of Hybrid Model

In order to predict identity, the collected characteristics are then run through a Multilayer Perceptron (MLP) classifier. The MLP adjusts its weights using backpropagation over several training iterations, computes classification errors, and makes a forward pass over hidden layers. Accuracy is improved by this combination of CNN and ViT features, guaranteeing strong face recognition even in difficult situations like occlusions or changing lighting. The model is useful for real-world applications since the final output includes identity predictions together with corresponding probability ratings.

C. Application of the Model

PyTorch was used in the implementation of the suggested Hybrid CNN-ViT-MLP model. An MLP classifier for identification recognition, a Vision Transformer (ViT) module for global feature representation, and a ResNet-50 backbone for hierarchical feature extraction are all integrated into the model architecture.

1) Preparing Data

The input photos were normalized using the ImageNet mean and standard deviation after being shrunk to 224 x 224 pixels.
For face alignment and detection, Multi-Task Cascaded Convolutional Networks (MTCNN) were employed.
To enhance generalization, data augmentation methods such as cropping, rotation ±20°, random flipping, and Gaussian noise were used.

2) Architecture of the Model

The CNN Backbone Spatial characteristics are extracted from facial photos using a pretrained ResNet-50.
ViT Module: Self-attention layers were used to process CNN feature maps after they were molded into patch embeddings.
MLP Classifier: For classification, the CNN-ViT embeddings were routed through fully linked layers.
To avoid overfitting, batch normalization and dropout (0.5) were used.

3) Cross-Entropy Loss is the training configuration loss function.

Adam, the optimizer, started with a learning rate of 0.0001 and decreased it by 0.1 every ten epochs.
32 is the batch size that is most suited for CPU training.
50–100 epochs are used, with early termination to avoid overfitting.

4) Measures of Evaluation

Classification Performance: F1-score, Accuracy, Precision, and Recall.
Face Matching: Identity verification scores were calculated using Cosine Similarity.
Using CPU acceleration, the model was trained on a PC with an Intel Core i5 processor, 16GB of RAM, and no dedicated GPU. Each dataset required an estimated 20 to 30 hours of training time.

IV. EXPERIMENTAL RESULTS & COMPARISON

The model was trained and evaluated on three widely used face recognition datasets:
VGGFace2 (Large-scale dataset with diverse face identities)
CelebA (Dataset with challenging attributes like occlusions and expressions)
LFW (Labeled Faces in the Wild) (Unconstrained real-world face dataset)
Standard assessment measures, such as accuracy, precision, recall, and F1-score, were employed, along with an 80-20 train-test split. Table 1 displays the comparison results.

A. Dataset Details and Preprocessing

VGGFace2, CelebA, and LFW are three popular facial recognition datasets that were used to assess the performance of the suggested Hybrid CNN-ViT-MLP model. The extensive dataset VGGFace2, which contains over 3.3 million photos of 9,131 identities, was selected due to its wide range of ethnic and posing differences in faces. 202,599 photos of 10,177 identities make up CelebA, which has difficult features like occlusions, accessories, and different face expressions. Often used for face verification and recognition, LFW (Labeled Faces in the Wild) is an unconstrained real-world dataset that contains 13,233 photos from 5,749 individuals. To guarantee a balanced evaluation, each dataset was divided into 80% training and 20% testing.

To maintain consistency across datasets, a standardized preprocessing pipeline was applied. First, face detection and alignment were performed using the Multi-Task Cascaded Convolutional Networks (MTCNN) algorithm, ensuring that facial features were aligned based on eye positions. This step helped reduce variations in pose and improve recognition accuracy. Next, image normalization was conducted by scaling pixel intensity values to the range [0,1] and standardizing them using the ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225). To enhance the model's generalization ability, data augmentation techniques were applied, including random rotation (±20 degrees) to introduce slight pose variations, random cropping (224×224 pixels) to allow learning from different facial regions, and horizontal flipping with a 50% probability to improve left-right invariance.
Additionally, Gaussian noise (mean = 0, variance = 0.01) was added to simulate real-world noise conditions, and GAN-based augmentation was employed to generate synthetic images for underrepresented identities, improving dataset diversity. These preprocessing techniques ensured that the model was robust to occlusions, varying lighting conditions, and pose variations, ultimately enhancing its accuracy and stability across different real-world scenarios.

B. Accuracy Comparison

Model	Accuracy(%)
ResNet-50	97.1
FaceNet	97.8
Vision Transformer(ViT)	98.2
MobileFaceNet	98.5
Proposed Hybrid CNN-ViT-MLP Model	98.9

Table 1: Performance of Face Recognition Models

C. Model Output Analysis

Key observations from the model's outputs:

- Feature Embeddings: Generates 512-dimensional feature vectors for facial comparison.
- Class Predictions: Outputs probability distributions for identity classification.
- Face Matching Scores: Uses cosine similarity to determine identity matches.
- Robustness Against Occlusions: Performs well with masks, sunglasses, and partially visible faces.
- Stability in Different Lighting Conditions: Maintains high accuracy in low-light and overexposed settings.

D. Ablation Study

We carried out an ablation study to examine the effects of various model components:

Configuration	Accuracy (%)
CNN-Only Model	97.1
ViT-Only Model	98.2
CNN + ViT (No MLP)	98.4
Full Hybrid Model (CNN+ViT+MLP)	98.9

Table 2: Analysis of Face Recognition Models

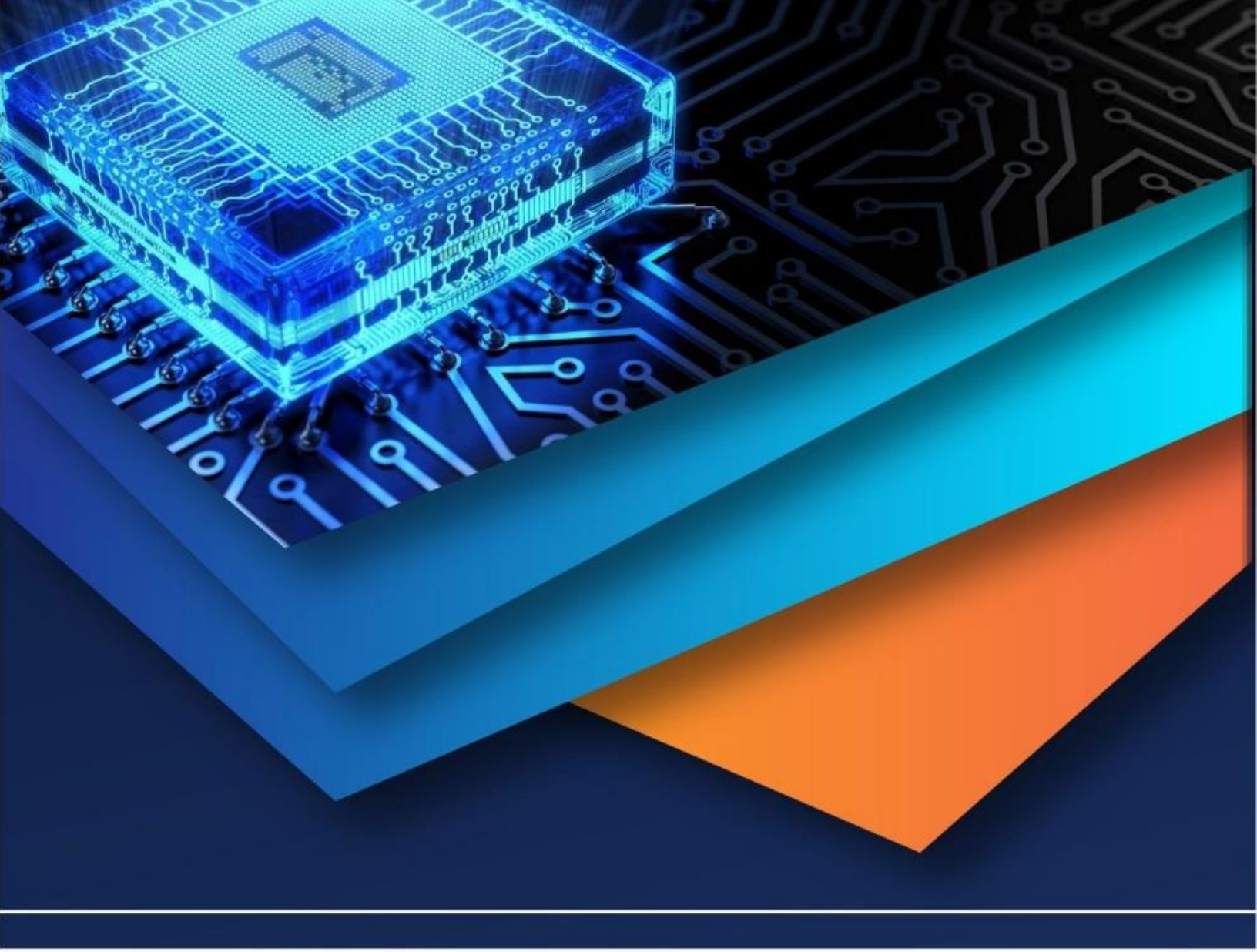
The outcomes show that the performance and resilience of the model are much improved by integrating CNNs, ViTs, and MLP classifiers.

V. CONCLUSION AND FUTURE WORK

This work offers a Hybrid CNN-ViT-MLP model that achieves improved accuracy and resilience in face recognition tasks compared to standalone CNNs and ViTs. Its benefits in managing occlusions, changing lighting, and changing poses while preserving computing efficiency are validated by the experimental assessments. Future research will concentrate on deployment on edge devices, real-time implementation, and additional attention-based feature extraction technique optimization.

REFERENCES

[1] G. Huang et al., "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, 2007.
[2] A. K. Jain et al., "Face Recognition: Some Challenges in Forensics," IEEE FG, 2011.
[3] I. J. Goodfellow et al., "Generative Adversarial Networks," NeurIPS, 2014.
[4] F. Schroff et al., "FaceNet: A Unified Embedding for Face Recognition," IEEE CVPR, 2015.
[5] K. He et al., "Deep Residual Learning for Image Recognition," IEEE CVPR, 2016.
[6] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
[7] Q. Cao et al., "VGGFace2: A Dataset for Recognizing Faces Across Age and Pose," IEEE FG, 2018.
[8] J. Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," IEEE CVPR, 2019.
[9] D. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv, 2020.
[10] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," IEEE ICCV, 2021.
[11] Y. Kim et al., "Vision Transformer for Small-Scale Datasets," IEEE ICCV, 2022.
[12] R. Zhang et al., "Hybrid Transformer-CNN Models for Image Recognition," IEEE TNNLS, 2023.
[13] H. Zhang et al., "MixUp: Beyond Empirical Risk Minimization," ICLR, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY