Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 2, June, 2026, pp. 155-172
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 156

A hybrid cnn with vision transformer technique for effective iris recognition

Nahla Abdulnabee Sameer [1*], Bashar M. Nema [2]
[1] Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics, Baghdad, Iraq.
[2] Department of Computer Science, Faculty of Sciences, Mustansiriyah University Baghdad, Iraq.

Article Info

Article history:

Received Jan.,23, 2026
Revised Mar.,17, 2026
Accepted Apr.,6, 2026

Keywords:

Iris recognition
Convolutional neural networks (CNN)
Vision transformer (ViT)
Semantic segmentation
Biometrics

ABSTRACT (10 PT)

Biometric reputation has experienced an extensive increase, with many businesses integrating biometric technology into their structures. Among those, iris recognition stands out due to its effectiveness in preventing identity fraud through disposing of the risks of collisions or fake suits, even with large datasets. While Convolutional Neural Networks (CNNs) have shown excessive accuracy, they require great datasets and computational assets. To cope with this, this observer proposes a hybrid version that mixes CNNs and Vision Transformer (ViT) for efficient iris picture identification and authentication. By optimizing the knowledge of price, the hybrid model achieves an impressive accuracy of 99.67% in iris recognition. The use of a cross-entropy loss function minimizes prediction mistakes and enhances class labeling accuracy. Additionally, the research introduces a progressive neural network-based total prediction model, Interleaved Residual (IRU-Net), for semantic segmentation and iris mask generation, alongside the Predicted ID model for precise identity predictions. The proposed version undergoes rigorous testing on three extensively available iris databases, demonstrating robust reputation performance. Moreover, the model indicates ability packages in different biometric fields, which include face and retinal reputation.

Corresponding Author:

Nahla Abdulnabee Sameer
Informatics Institute for Postgraduate Studies, University of Information Technology and Communications (UOITC), Baghdad, Iraq.
Baghdad, Iraq
Email: nahlaphd1973@gmail.com

1. INTRODUCTION

Recently, there has been a developing interest in iris recognition and identification as a type of biometric verification. One of the most dependable and powerful biometric techniques for identity and authentication is iris fame [1]. Iris's reputation has made it one of the maximum dependable biometric identification technologies because of its accuracy, stability, and dependability. Consequently, it is used in many exceptional domains, which include border manipulation, forensics, healthcare, and sensible unlocking [2]. The iris has emerged as a completely informative and dependable object for identity in biometrics due to the eye's distinctive characteristics that continue to be steady at some stage in time [3, 4]. Age has no impact on the iris, and it doesn't want to come into direct contact with the scanning device [5]. The public's attention has been absolutely focused on contactless biometrics because of their reputation for being clean and consumer-pleasant [6]. The accuracy of iris scans, or the diploma to which biometric devices can differentiate among individuals, is distinctly excessive. In addition, iris recognition's dependability is considerably higher than that of other biometric techniques, suggesting that it's miles dependable for recognition responsibilities [7]. Furthermore, each iris is hugely exceptional from the others, which makes it a major assignment for studies aimed at particular non-public identities [8, 9]. Several technologies have been used in the development of the iris reputation machine. Convolutional Neural Networks (CNN), which are now the most used

approach in photography, are used the most [10]. But systems that rely mostly on self-attention, like Vision Transformers (ViT), have also shown similar accuracy in tasks involving the classification and recognition of photos. However, unlike neural networks, which incorporate image-unique biases, improving accuracy with ViT necessitates an exponential upward push within the vast array of parameters. ViT, which was first introduced in 2018 for device translation, has since evolved into the main architecture for tasks related to Natural Language Processing (NLP), such as speech recognition and text transcription. ViT has proven essential to recent advancements in natural language processing, as evidenced by models such as OpenAI's GPT-three and Google's BERT. Transformers have been studied for picture reputation obligations; however lightweight CNN have been generally used for mobile imaginative and prescient obligations. For example, the authors of [11] experimented with transformers in the picture category and produced results that were similar to those of contemporary CNN [12]. Convolutional networks are typically reworked utilizing transformers or precise module replacements in these attempts. The approach of dividing an input photo into a few patches, which were subsequently fed into the transformer as usual, was one innovation of this study. Together with words (tokens), the picture patches are handled in the context of natural language processing. According to Ref. [11], this approach interprets a 16x16 input photograph into 16x16 words. According to the findings of these investigations, this transformer-based structure performs better than contemporary CNN, which is sometimes regarded as the best community arrangement for photo reputation. However, the repeated nature of down-sampling makes it difficult to capture unique localization facts, and ViT's focus on low-decision characteristics makes it significantly less suitable for capturing images with exceptional-grained information. Many studies suggest hybrid models that combine CNN and ViT for improved performance in computer vision tasks, even though it has made significant advancements [13]. Inspired by [14] paintings, which proposed a Progressive Multi-scale Vision Transformer (PMVT) version combining convolutional feature maps and a transformer encoder, this study focuses on developing and implementing a hybrid model combining CNN and ViT technologies for iris photo recognition. CNN will extract capabilities in the early stages, using the strengths of each technology, while the transformer model will handle the following processing. The key contributions of the research are as follows:

- Advanced techniques like Hough Circle Transform (HCT) and linked component analysis to accurately segment pupils and iris, resulting in exact border detection.
- Used Local Binary Patterns (LBP) and histogram-based algorithms to extract particular iris styles and enhance type performance.
- Developed the Interleaved Residual U-Net (IRU-Net) version to improve iris segmentation accuracy and address the vanishing gradient issue with the use of residual connections.
- Proposed a hybrid model that combines CNN and ViT architectures to enhance iris picture categorization by way of utilizing their strengths in function extraction and context awareness
- Novelty prediction ID model uses an iris prediction ID to categorize authenticated IDs.

2. RELATED WORK

This phase affords a top-level view of famous deep getting-to-know (DL) strategies utilized in biometric recognition, alongside an overview of studies specializing in the usage of Vision Transformers (ViT) in picture recognition. DL stands proud of conventional system getting-to-know Machine learning (ML) techniques as they allow for impartial characteristic getting to know, which significantly reduces effort and time[15]. Additionally, DL fashions excel at robotically identifying hidden patterns in records [16]. One extraordinary technique introduces an iris reputation method that gets rid of the need for an iris segmentation segment. This technique complements the popular skills of a pre-trained Convolutional Neural Network (CNN) at the beginning evolved for the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [17]. To prevent overfitting, the model is high-quality-tuned using Bayesian optimization and information augmentation techniques. This advancement improves the model's adaptability to new duties, showcasing the effectiveness of DL in biometric recognition. In any other observe [18], researchers developed a hybrid network based totally on EfficientNet-b0 to pick out iris photos. This unmarried network integrates key processes like segmentation, normalization, feature extraction, and matching. Tests on blended datasets (MMU2 and CASIA Thousand) have proven that this composite community is more correct and efficient than previous designs, imparting terrific pace and parameter performance. Early deep-getting-to-know models for facial recognition were additionally evaluated on low-ability devices to measure accuracy, length, and processing time [19]. These experiments used various datasets, each public and personal, to simulate the complexities of mobile eventualities. To ensure equity, publicly available models and conventional approaches were examined alongside. The implementation, built with the use of Java and Android programming languages, was integrated right into a cellular gadget to evaluate its general overall performance. Another place of exploration focuses on multimodal big facts. A survey of contemporary deep mastering fashions highlights strategies for fusing multimodal records [20]. While this field offers huge capability, demanding situations remain. The survey

emphasizes teaching researchers and fostering the development of revolutionary multimodal information fusion strategies, regardless of disciplinary obstacles. A precise condensed 2-channel CNN for iris reputation and verification is proposed in [21]. This model calls for much less schooling facts while preserving high overall performance. It includes 3 online augmentation techniques and radial attention layers, with weight distribution analyses used to prune needless branches and channels. Tests on the CASIA-V4-Thousand dataset showed the model's robustness in difficult situations. Lastly, advancements in iris localization, inclusive of defining pupillary and limbic barriers, stay underutilized in enhancing CNN-based segmentation for iris reputation [22]. Table 1 summarizes related works, evaluating representative architectures that contribute to the knowledge of multimodal deep studying. The evaluation additionally discusses modern-day facts fusion models, current demanding situations, and the capacity for future instructions on this evolving subject.

Table 1. Literature benchmark-related comparison.

Author	Method(s)	Performance	Limitation
[17]	Studying Transfer (AlexNet & DenseNet201)	Accuracy; AlexNet: 0.97.22%, DenseNet201: 0.98.81%	Complexity of the version and lengthy processing time for algorithms and computations
[18]	EfficientNet-b0	Accuracy: 0.98%	Iris area deviated the usage of pre-trained version
[20]	Convolutional Neural Network for multimodal	Error Equal Rate: 0.60%	Multi-modal popularity
[21]	Deep Convolutional Neural Network; 2-Channels	EER Verification: CASIA V1: 0.33%, CASIA V3: 0.76%.	Lower floating-factor operations and parameters.
Proposed Model	Hybrid CNN with ViT Technique	Accuracy: 0.99.61%.	The hybrid CNN + ViT model is well-suited for a variety of complex image identification applications because it combines the advantages of both architectures to achieve reliable feature extraction and improved accuracy.

3. METHODS AND EXPERIMENTS

Iris recognition is a biometric method that relies on the unique patterns of the iris for identification and verification. The preprocessing stage is crucial for reinforcing the images' excellence and ensuring accurate characteristic extraction. As proven in Figure 1, the procedure begins with picture acquisition, where left and proper eye pix are captured. This is accompanied by the aid of preprocessing and scholar detection, which converts the photograph to grayscale and locates the student. Next, iris segmentation isolates the iris area, and normalization transforms it right into a square layout. Finally, function extraction and matching examine iris patterns with a stored database, leading to the very last recognition or verification choice. The dataset contains an Iris image obtained from Kaggle 2024. We obtained a dataset consisting of two sub-datasets: live photographs and false images, which were split into training 80% and testing data 20%. The versions is a Windows 11, with "11th Gen Intel (R) Core (TM)" i5-11400F @ 2.60GHz processor, 16GB of RAM, and an AMD Radeon (TM) R9 370 GPU. The models are set up in Python 3.11.5 with the Keras API, which is integrated with the TensorFlow version 2.12 backend and includes CUDA/CuDNN support for GPU acceleration.

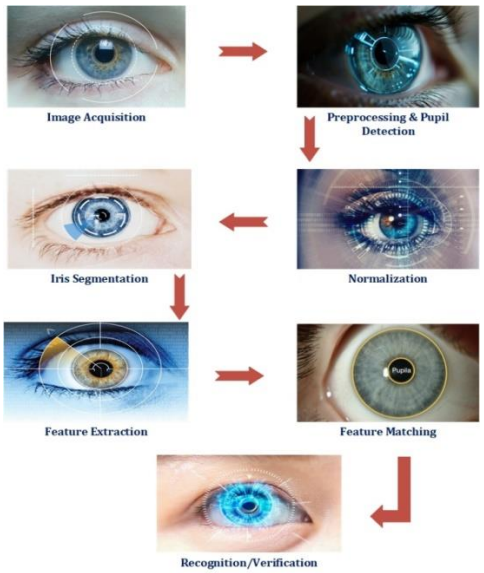


Figure 1. Iris recognition process diagram.

3.1. Image Preprocessing

The image preprocessing stage seeks to identify the critical iris texture region while removing undesirable parts, such as eyelids and eyelashes, that are unnecessary to feature extraction. This technique focuses on localizing the iris and pupil to prepare the image for segmentation and analysis. To begin, the input image is enhanced using an average blur filter, which lowers high-frequency noise and eliminates undesired details. Next, a binary thresholding procedure is used on the blurred image to separate the relevant elements from the background. The resulting binary image makes it easier to isolate critical regions such as the pupil and iris. The binary image is evaluated with connected components to discover probable locations of interest. The pupil, which is usually the second largest circular vicinity in length, is extracted by selecting the second biggest bounding element. This step helps to decide the scholar's middle and radius, which are essential for proper iris segmentation. To section the iris, the authentic image is despatched through a mean filter, which reduces noise while keeping essential edge data. The "Hough Circle Transform (HCT)" is then implemented to determine the internal boundary between the pupil and the iris, and to decide the center and radius of the iris. To finds on the outer boundary of the iris, the model makes use of the HCT circle remodel, which is used after outlining, to identify and become conscious of the outer iris boundary, which corresponds to wherein the iris meets the sclera surface.

3.2. Structure of Iris Recognition System

Iris's popularity is a reliable biometric approach for identifying people based totally on their distinct iris patterns. The device includes several important phases, including image acquisition, preprocessing, pupil and iris reputation, function extraction, normalization, and matching. Figure 2 depicts how this device integrates superior techniques inclusive of the Hough Circle Transform and connected element evaluation to create a reliable and secure approach for biometric identity.

3.2.1. Image Acquisition

The technique begins by way of acquiring grayscale pics of the left and proper eyes, which are commonly in “bmp layout”.

3.2.2. Pupil detection and preprocessing

The image is handled to limit noise with a blur filter out. Connected thing evaluation identifies the scholar and calculates its center and radius to make sure correct iris segmentation.

- **Average Blur Filter**

To flatten the image and lower the noise, apply the average blur filter. The mathematical representation of the filter's convolution operation is as follows Equation (1).

$$avg(x,y) = \sum_{i=-2}^{2}\sum_{j=-2}^{2}\frac{1}{25}\cdot src(x+i,y+j) \quad (1)$$

Where each pixel is averaged and the kernel is a 5x5 matrix.

- **Connected Components**

This approach finds the binary image's connected areas, or components. This stage entails counting the number of interconnected zones and classifying them according to size. Each connected component's size is determined by Equation (2).

$$size(i) = \sum_{(x,y)} I_{bin}(x,y) \quad (2)$$

Where uses the component's centroid to determine the centre, $c_x, c_y = centroid\ of\ component$

3.2.3. Iris segmentation

The approach extracts the iris by the use of a median filter and utilizing the HCT to determine its boundaries. Median Filter, to maintain edges and further smooth the image, a median filter is used to define the mathematical operation, $smoothed(x,y) = median(src(x',y'))for\ (x',y') \in neighborhood\ of\ (x,y)$. Where the local area is typically a square region surrounding each pixel.

- **Hough Circle Transform (HCT)**

This approach detects circles in an image. It converts the image into a parametric space, searching for maxima that correspond to potential circles. The circle in polar coordinates as Equation (3).

$$(x-c_x)^2 + \left(y-c_y\right)^2 = r^2 \quad (3)$$

Where (c_x, c_y) is the centre of the circle, and r is its radius. The HCT determines the parameters c_x, c_y, and r that best fit the observed iris.

3.2.4. Normalization

The iris region is normalized to a standard size to ensure consistency. This entails transforming the circular iris shape to a rectangle one and increasing the contrast via histogram equalization.

- **Transformation to a Rectangular Shape**

A rectangular shape is created from the circular iris region. In the normalized rectangular image, every pixel in the circular iris is mapped to a point. The procedure entails converting polar coordinates to Cartesian coordinates using the following Equation (4).

$$x = r \cdot \cos\theta + centrer_x$$
$$y = r \cdot \sin\theta + centrer_y \quad (4)$$

where the matching pixels in the new rectangular space are transferred to the normalized polar coordinates, $r, and\ \theta$.

- **Histogram Equalization**

Histogram equalization is used to enhance contrast. The following is an expression for the transformation of the pixel intensities as Equation (5).

$$s = T(r) = \frac{(r - \text{min\_intensity})}{\text{max\_}intensity - \text{min\_intensity}} \cdot new\_max - new\_min + new\_min \quad (5)$$

where $T(r)$ is the converted intensity and r is the original intensity.

3.2.5. Feature Extraction and Classification

Unique patterns from the iris are retrieved and compared. These characteristics are encoded to facilitate classification and matching.

3.2.6. Hamming Distance Matching

The retrieved traits are compared to a database to confirm or identify the individual, ensuring precise recognition as Equation (6).

$$D = \sum_i \left| F_{query}(i) - F_{database}(i) \right| \quad (6)$$

Where i is the index of every feature bit and F is the feature vector.

3.3. Novel prediction model based on neural networks Interleaved Residual U-Net (IRU-Net)

This paper provides Interleaved Residual U-Net (IRU-Net), a progressive neural network-based exhaustive prediction version created especially for iris mask advent and semantic segmentation. The version makes use of the U-Net architecture, which achieves green segmentation, specifically for biomedical duties, utilizing an encoder-decoder shape with bypass links. Residual connections are incorporated into the network to enhance training performance and deal with the vanishing gradient problem, allowing for deeper topologies without sacrificing performance. The version additionally uses IRU-Net characteristic maps, which beautify record float and segmentation accuracy employing combining high-level and low-stage statistics. The model's iris mask era feature is designed to precisely find the iris boundary in eye photos, which is a crucial challenge for biometric packages along with iris recognition. In phrases of accuracy and efficiency, the IRU-Net performs better than conventional strategies. This novel technique expands the potential of semantic segmentation across a couple of domains and makes a vast contribution to iris segmentation.

3.3.1. Segmenting Iris using U-Net

The U-Net architecture is used on this have a look at to phase the iris location from eye pix. A convolutional neural community (CNN) known as U-Net was created specifically for semantic segmentation responsibilities, inclusive of biomedical image segmentation. The system for iris segmentation and the use of this model is defined below.

- **Preprocessing of Data**

Normalizing the pixel values to fall inside [0, 1] is the first step in preprocessing the attention snapshots, which can be supplied as grayscale BMP photographs. In particular, every picture is broken up by using 255.0 as the subsequent Equation (7).

$$I_{norm} = \frac{I}{255.0} \quad (7)$$

where I represent the original image and I_{norm} represents the normalized image. Images with "5" in their filenames are put aside for testing, and the dataset is divided into training and testing sets.

- **Architecture of Models**

The contracting path (encoder) and the expansive path (decoder) are the two primary parts of the U-Net architecture. The most abstract features are represented by a bottleneck layer that connects these.

1) Contracting Path (Encoder)

Convolutional layers and max-pooling operations are used by the encoder to gradually extract features. To capture higher-level abstractions, the number of filters rises with depth. To lessen overfitting and regularize the model, dropout layers are incorporated.

2) Bottleneck Layer

This layer, which represents the most abstract feature information, sits between the encoder and decoder.

3) Expanded Path (Decoder)

The decoder recovers spatial resolution by upscaling the feature maps using transposed convolutions. To retain spatial features lost during downsampling, skip connections from the encoder are concatenated with the appropriate layers in the decoder.

4) Output Layer

A binary mask with pixel values of 1 denoting the anticipated iris region and 0 denoting the backdrop is produced by the last layer, a 1x1 convolution.

3.3.2. Formulation of model

The following are the network's main functions;

- Convolution

Every layer undergoes a convolution procedure as the Equation (8).

$$\mathrm{Y}(i,j) = \sum_{m=1}^{M}\sum_{n=1}^{N} X(i+m-1, j+n-1)\cdot W(m,n) \quad (8)$$

where W is the convolutional kernel, X is the input feature map, and Y is the output feature map.

- Max-Pooling

This technique chooses the highest value in a 2x2 patch to minimize spatial dimensions as the Equation (9).

$$\mathrm{Y}(i,j) = max\big(X(2i-1,2j-1), X(2i-1,2j), X(2i,2j-1), X(2i,2j)\big) \quad (9)$$

- Transposed Convolution (Upsampling)

The feature maps are upsampled using transposed convolutions as the Equation (10).

$$\mathrm{Y}(i,j) = \sum_{m=1}^{M}\sum_{n=1}^{N} X(i+m-1, j+n-1)\cdot W^{T}(m,n) \quad (10)$$

where the transposed convolution kernel is denoted by W^T.

- Sigmoid Activation

A sigmoid activation function is applied to create the output mask as the Equation (11).

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (11)$$

where x is the last convolution layer's raw output.

3.3.3. Loss Function

Binary cross-entropy loss, which measures the variation between the ground truth and predicted masks, is used to train the model as Equation (12).

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log 1 - p_i) \quad (12)$$

Where, y_i and p_i indicate the ground truth and anticipated values for pixel i, respectively.

3.4. Hybrid CNN and Vision Transformer (ViT) Model

The model uses CNN and Vision Transformers (ViT) to categorize iris pictures. CNN extracts spatial characteristics, which ViT analyses utilizing transformer layers for global context awareness. Both outputs are combined to increase categorization accuracy as show in Figure 2.

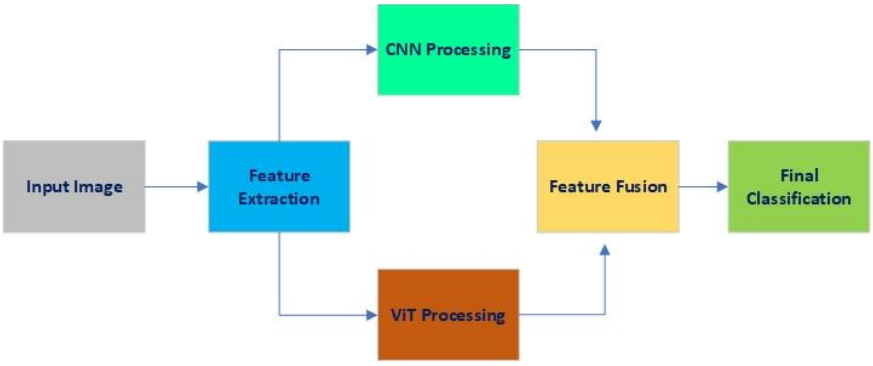


Figure 2. Hybrid model CNN+ViT digram.

1) Feature extraction and image preprocessing

- Local Binary Pattern (LBP)

$$\text{LBP}(x,y) = sgn(I(x,y) - I_c) \quad (13)$$

The comparing pixel intensities with relatives, LBP determines texture properties.
LBP produces texture characteristics by comparing pixel intensities to their relatives.

- Histogram

$$\text{H}(x) = \sum_{i=0}^{255} count\ (I(x,y) = i) \quad (14)$$

For texture analysis, histograms are created by computing the image's pixel intensities.

2) CNN Model Construction

Equation (15) describes the design of the CNN:

$$Conv2D \rightarrow MaxPooling \rightarrow Flatten \rightarrow Dense \quad (15)$$

- The Conv2D Layer extracts local patterns.
- MaxPooling reduces spatial dimensions.
- The dense Layer is a fully connected layer for categorization.

The CNN produces a 128-unit fully linked layer, followed by a Softmax for 45 classes.

3) Vision Transformer (ViT) Construction

ViT employs self-attention processes as Equation (16):

$$Attention\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (16)$$

- Multi-Head-Attention combines several attention heads to learn varied feature representations.
- Global average pooling reduces the output sequence to a fixed-size vector.
- The ViT utilizes a dense layer to generate class probabilities.

4) Hybrid Model CNN + ViT

The hybrid model combines the CNN and ViT outputs as Equation (17):

$$Hybrid\ Output = Concat\left(CNN_{output},\ ViT_{output}\right) \quad (17)$$

A final dense layer combines the two outputs for categorization.

Model Training

- Cross-entropy loss for multi-class classification.
- Adam optimizes to reduce loss.

Evaluation and Metrics

- Accuracy: Determines the percentage of correctly classified instances.
- AUC-ROC evaluates model performance across many thresholds as the following Equation (18).
-

$$AUC = \int_0^1 TRP(x)\, d\left(FPR(x)\right) \quad (18)$$

In this case, FPR stands for False Positive Rate and TPR for True Positive Rate.

4. RESULTS

4.1. Iris Segmentation

The iris segmentation and extraction procedure produces many key images for each input image. These include the original grayscale input image, the average blur image acquired after applying a smoothing filter to minimize noise, and the binary image created by thresholding. The pupil circle image displays the original image with a red circle representing the detected pupil. The iris region is removed and saved separately using a mask, and the iris bounding box displays a cropped representation of the iris region within a bounding box. The combined iris and pupil circles image shows the original image with the pupil and iris indicated in red and blue circles, respectively. Finally, the normalized iris picture is created using polar coordinate translation and histogram equalization, resulting in a standardized depiction of the iris. The output photographs are divided into subfolders for each subject, with distinct directories for "left" and "right" images as shown in Figure 3.

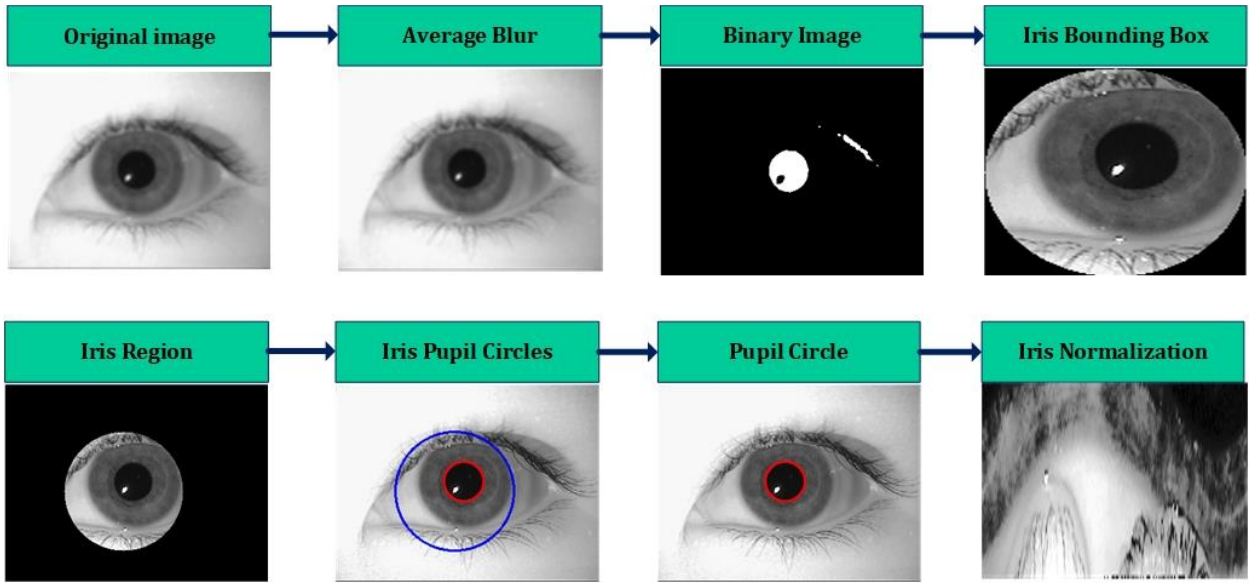
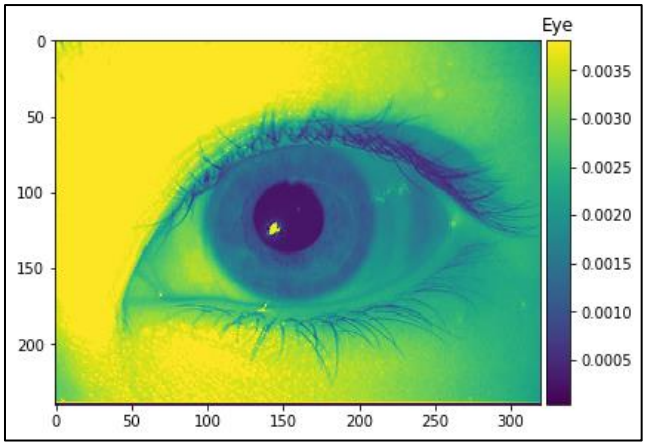


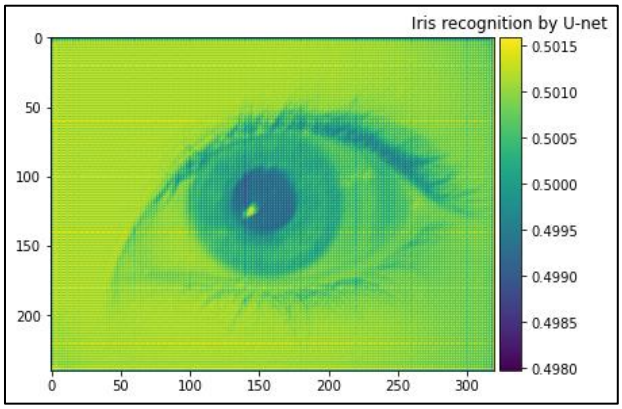
Figure 3. Iris segmentation results.

4.2. IRU-Net a novel prediction model

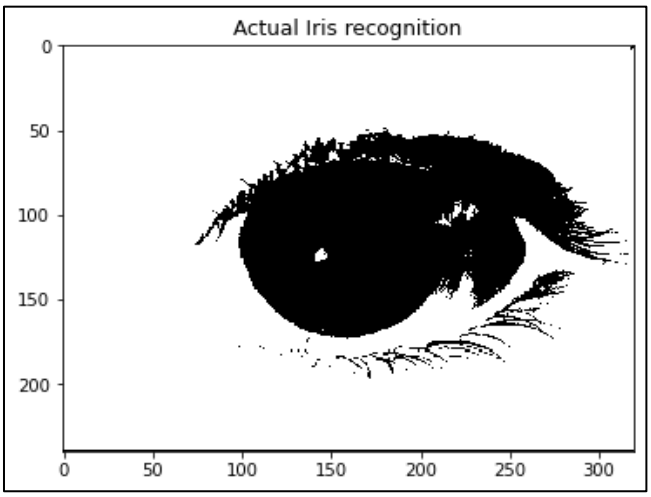
The results of the iris segmentation model utilizing U-Net. The model first processes the input eye photos by extracting the iris region with a convolutional neural network (CNN) architecture, especially U-Net. The trained model predicts iris segmentation masks, which are then compared to the actual ground truth masks (which are manually labelled). For visual comparison, as shown in Figure 4.



A



B



C

Figure 4. Iris mask prediction results in a) Eye Image, b) Iris Recognition by U-Net, c) Actual Iris Recognition.

Figure 4 shows the following firstly the eye image, The original input image of the eye. The second, is iris recognition by U-Net, the predicted segmentation mask generated by the U-Net model, highlighting the detected iris region. The third, is actual iris recognition, the ground truth mask, representing the actual iris region as labelled in the dataset. Figure 4 and Table 2, are shown to demonstrate how well the model can identify and separate the iris from the eye images. The accuracy is tracked during the training process, and the model employs a binary cross-entropy loss function. By contrasting the expected and real masks, the visual output enables the segmentation quality to be assessed.

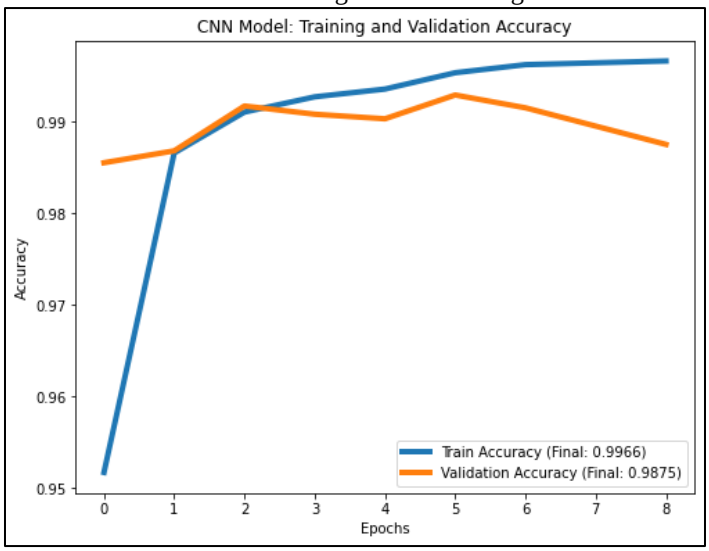
Table 2. The U-Net model with layer detail's structure.

Layer Type	Shape of Output	Parameter Count	Connections
input_1 (Input Layer)	(None, 240, 320, 1)	zero	None
conv2d (Convolutional Layer)	(None, 240, 320, 16)	160	input_1[0][0]
dropout (Dropout Layer)	(None, 240, 320, 16)	zero	conv2d[0][0]
conv2d_1 (Convolutional Layer)	(None, 240, 320, 16)	2320	dropout[0][0]
max_pooling2d (MaxPooling)	(None, 120, 160, 16)	zero	conv2d_1[0][0]
conv2d_2 (Convolutional Layer)	(None, 120, 160, 32)	4640	max_pooling2d[0][0]
dropout_1 (Dropout Layer)	(None, 120, a 160, 32)	zero	conv2d_2[0][0]
conv2d_3 (Convolutional Layer)	(None, 120, a 160, 32)	9248	dropout_1[0][0]
max_pooling2d_1 (MaxPooling)	(None, 60, 80, 32)	zero	conv2d_3[0][0]
dropout_2 (Dropout Layer)	(None, 60, 80, 64)	zero	conv2d_4[0][0]
conv2d_5 (Convolutional Layer)	(None, 60, 80, 64)	36928	dropout_2[0][0]
max_pooling2d_2 (MaxPooling)	(None, 30, 40, 64)	zero	conv2d_5[0][0]
conv2d_6 (Convolutional Layer)	(None, 30, 40, 128)	73856	max_pooling2d_2[0][0]
dropout_3 (Dropout Layer)	(None, 30, 40, 128)	zero	conv2d_6[0][0]
conv2d_7 (Convolutional Layer)	(None, 30, 40, 128)	147584	dropout_3[0][0]
max_pooling2d_3 (MaxPooling)	(None, 15, 20, 128)	zero	conv2d_7[0][0]
conv2d_8 (Convolutional Layer)	(None, 15, 20, 256)	295168	max_pooling2d_3[0][0]
dropout_4 (Dropout Layer)	(None, 15, 20, 256)	zero	conv2d_8[0][0]
conv2d_9 (Convolutional Layer)	(None, 15, 20, 256)	590080	dropout_4[0][0]
conv2d_transpose (Upsampling)	(None, 30, 40, 128)	131200	conv2d_9[0][0]
concatenate (Concatenate Layer)	(None, 30, 40, 256)	zero	conv2d_transpose [0][0], conv2d_7[0][0]
conv2d_10 (Convolutional Layer)	(None, 30, 40, 128)	295040	concatenate [0][0]
dropout_5 (Dropout Layer)	(None, 30, 40, 128)	zero	conv2d_10[0][0]
conv2d_11 (Convolutional Layer)	(None, 30, 40, 128)	147584	dropout_5[0][0]
conv2d_transpose_1 (Upsampling)	(None, 60, 80, 64)	32832	conv2d_11[0][0]

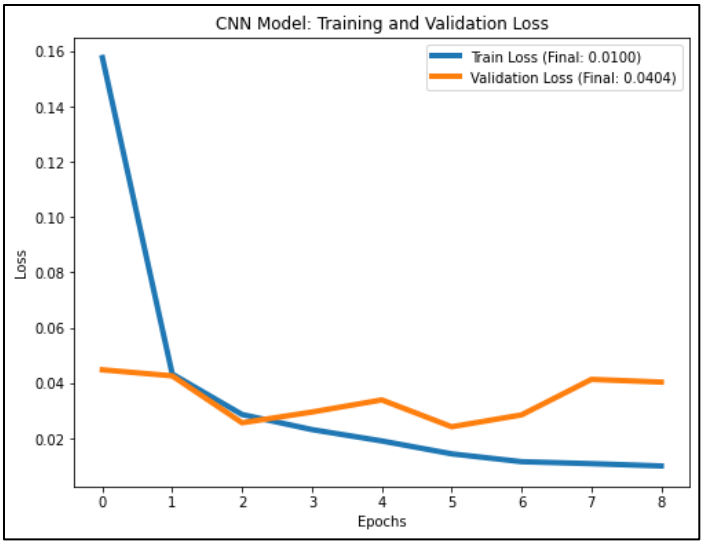
concatenate_1 (Concatenate Layer)	(None, 60, 80, 128)	zero	conv2d_transpose_1[0][0], conv2d_5[0][0]
conv2d_12 (Convolutional Layer)	(None, 60, 80, 64)	73792	concatenate_1[0][0]
dropout_6 (Dropout Layer)	(None, 60, 80, 64)	zero	conv2d_12[0][0]
conv2d_13 (Convolutional Layer)	(None, 60, 80, 64)	36928	dropout_6[0][0]
conv2d_transpose_2 (Upsampling)	(None, 120, 160, 32)	8224	conv2d_13[0][0]
concatenate_2 (Concatenate Layer)	(None, 120, 160, 64)	zero	conv2d_transpose_2[0][0], conv2d_3[0][0]
conv2d_14 (Convolutional Layer)	(None, 120, 160, 32)	18464	concatenate_2[0][0]
dropout_7 (Dropout Layer)	(None, 120, 160, 32)	zero	conv2d_14[0][0]
conv2d_15 (Convolutional Layer)	(None, 120, 160, 32)	9248	dropout_7[0][0]
conv2d_transpose_3 (Upsampling)	(None, 240, 320, 16)	2064	conv2d_15[0][0]
concatenate_3 (Concatenate Layer)	(None, 240, 320, 32)	zero	conv2d_transpose_3[0][0], conv2d_1[0][0]
conv2d_16 (Convolutional Layer)	(None, 240, 320, 16)	4624	concatenate_3[0][0]
dropout_8 (Dropout Layer)	(None, 240, 320, 16)	zero	conv2d_16[0][0]
conv2d_17 (Convolutional Layer)	(None, 240, 320, 16)	2320	dropout_8[0][0]
conv2d_18 (Final Conv Layer)	(None, 240, 320, 1)	17	conv2d_17[0][0]
Trainable Parameters	1,940,817	-	-
Non-Trainable Parameters	zero	-	-

4.3. CNN Model results

The CNN model performs well, with high accuracy and low loss values during training and validation. A training accuracy of 99.66% implies that the model correctly predicted nearly all of the data in the training set, with a minimal training loss of 0.0100 indicating good error minimization. Similarly, a validation accuracy of 98.75% confirms the model's great generalization to new data, while a validation loss of 0.0404, somewhat larger than the training loss, is still low and indicates minimal overfitting as shown in Figure 5.



A

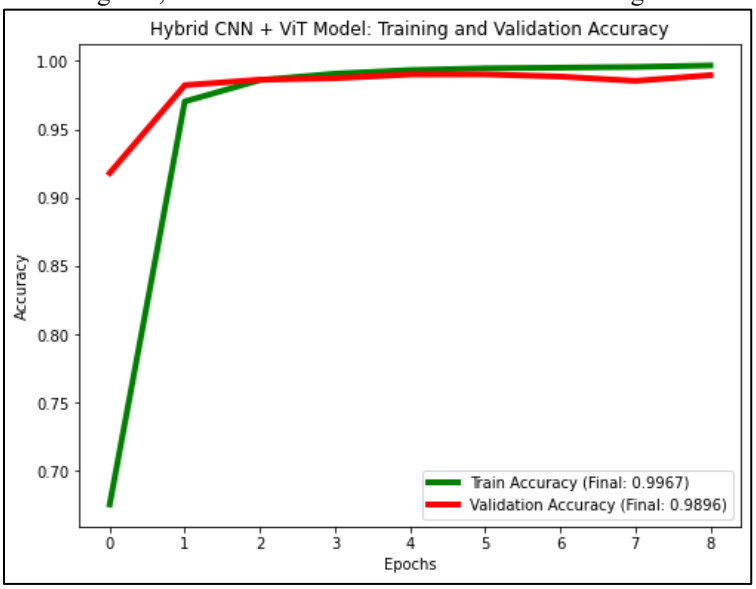


B

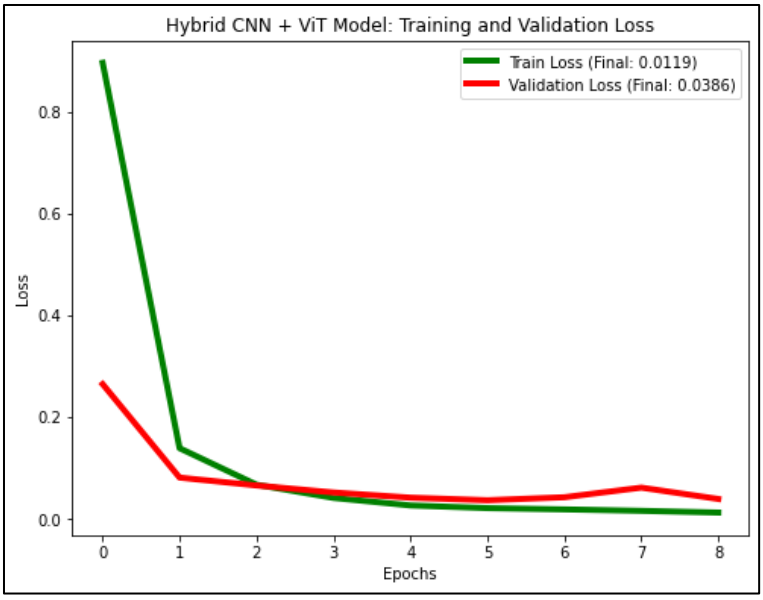
Figure 5. CNN Model; a) Training and validation accuracy, b) Training and validation loss results.

4.4. Hybrid Model CNN + ViT results

The hybrid Model CNN + ViT model performs well, with high accuracy and low loss values during training and validation. A training accuracy of 99.67% implies that the model correctly predicted nearly all of the data in the training set, with a minimal training loss of 0.0119 indicating good error minimization. Similarly, a validation accuracy of 98.96% confirms the model's great generalization to new data, while a validation loss of 0.0386, while somewhat larger than the training loss, is still low and indicates minimal overfitting as shown in Figure 6.



A



B

Figure 6. Hybrid CNN+Vit Model; a) Training and validation accuracy, b) Training and validation loss results.

4.5. Comparison of CNN vs Hybrid CNN+Vit Model; Training and validation accuracy results

Both the CNN model and the hybrid CNN + ViT model perform well in terms of training and validation accuracy, with the hybrid model slightly outperforming in generalization and loss minimization. The CNN model achieved a training accuracy of 99.66%, demonstrating its ability to properly classify nearly all training data, and a low training loss of 0.0100, showing successful error minimization. Its validation accuracy was 98.75%, showing good generalization to unseen data. The validation loss of 0.0404, while significantly greater than the training loss, is still low, indicating minimal overfitting. In comparison, the hybrid CNN + ViT model performed marginally better, reaching a training accuracy of 99.67%, which is comparable to the CNN model, but with a slightly larger training loss of 0.0119. Interestingly, the hybrid model surpassed the CNN model in validation, reaching a validation accuracy of 98.96%, an improvement of 0.21%. Furthermore, its validation loss of 0.0386 was less than that of the CNN model, indicating better generalization and less overfitting. The hybrid CNN + ViT model showed higher validation performance, demonstrating that the inclusion of Vision Transformer components improves the model's potential to generalize while maintaining high accuracy and little overfitting, making it a preferable alternative for complicated data settings as shown in Figure 7.

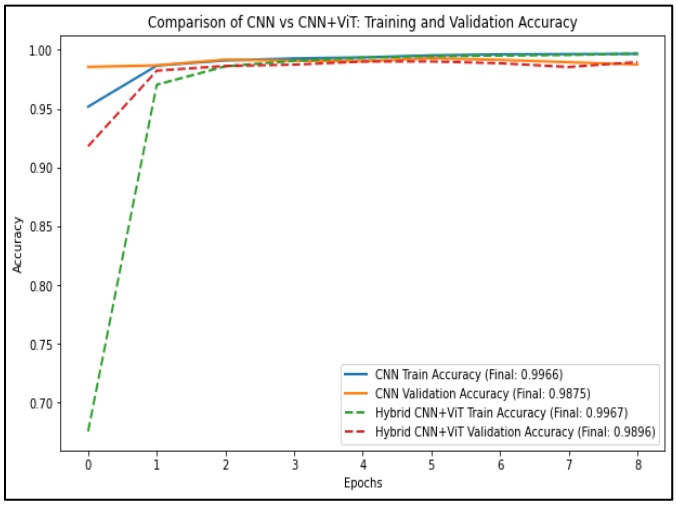


Figure 6. Comparison of CNN vs Hybrid CNN+Vit Model; Training and validation accuracy results.

4.6. CNN vs Hybrid CNN+Vit; AUC-ROC Curve Models result

The CNN and Hybrid CNN + ViT models use convolutional layers to extract spatial feature hierarchies, whereas max-pooling layers minimize spatial dimensions to improve computing efficiency and promote feature invariance. Fully connected layers complete the categorization operation. During training, the models showed constant growth in both training and validation accuracy before plateauing at high levels, demonstrating effective learning and generalization. The CNN model has an accuracy of 99.29%, with macro and weighted averages for precision, recall, and F1-score all approaching 0.993 (Table 3).

Table 3. Performance of CNN Model.

Class	Precision	Recall	F1-Score	Support
0	0.991870	0.995918	0.993890	980
1	0.993001	1.000000	0.996488	1135
2	0.994192	0.995155	0.994673	1032
3	0.996032	0.994059	0.995045	1010
4	0.991878	0.994908	0.993391	982
5	0.993228	0.986547	0.989876	892
6	0.996825	0.983299	0.990016	958
7	0.995112	0.990272	0.992686	1028
8	0.985743	0.993840	0.989775	974
9	0.991098	0.993062	0.992079	1009
Accuracy	0.992900	0.992900	0.992900	10000
Macro Avg	0.992898	0.992706	0.992792	10000
Weighted Avg	0.992914	0.992900	0.992897	10000

Hybrid CNN + ViT models' has a little lower accuracy of 99.02%, with macro and weighted averages of precision, recall, and F1-scores around 0.990 (Table 4). Despite having a marginally reduced overall accuracy, the hybrid model performed well across all categories, as shown by F1 scores of more than 0.985 in most classes. The hybrid models' balanced precision and recall measures indicate that it manages class variations and imbalances effectively.

Table 4. Performance of Hybrid CNN + ViT Model.

Class	Precision	Recall	F1-Score	Support
0	0.994893	0.993878	0.994385	980
1	0.987794	0.998238	0.992989	1135
2	0.999010	0.977713	0.988247	1032
3	0.989184	0.996040	0.992600	1010
4	0.987743	0.984725	0.986232	982
5	0.995490	0.989910	0.992693	892
6	0.993684	0.985386	0.989518	958
7	0.977969	0.993191	0.985521	1028
8	0.989785	0.994867	0.992320	974
9	0.988095	0.987116	0.987605	1009
Accuracy	0.990200	0.990200	0.990200	10000
Macro Avg	0.990365	0.990106	0.990211	10000
Weighted Avg	0.990248	0.990200	0.990198	10000

This performance demonstrates a balanced and robust classification across all fingerprint categories. For example, the precision and recall for most classes are above 0.99, indicating good confidence in predictions across diverse class distributions. Figure 7 illustrates these findings, with minor misclassifications. Precision, recall, and F1-score data show that both models are very accurate and perform well across all classes. The CNN model has a

modest advantage in accuracy and class-specific measures, but the Hybrid CNN + ViT model produces equivalent results, demonstrating its ability to generalize effectively across heterogeneous and imbalanced datasets.

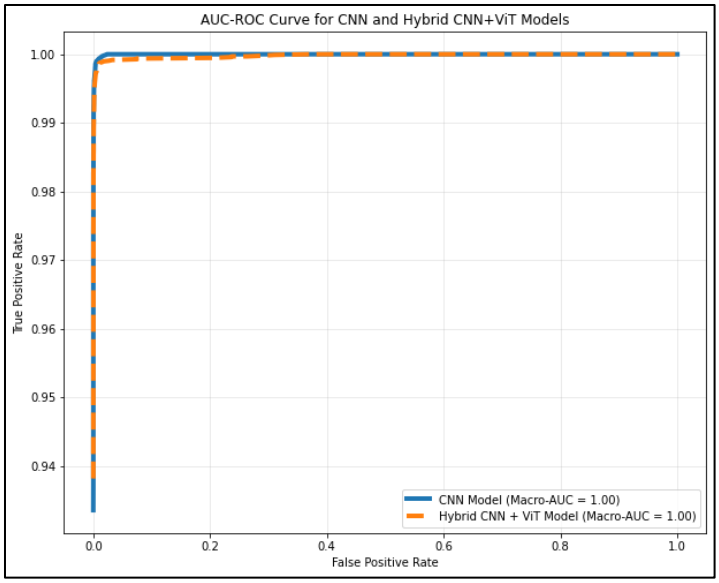
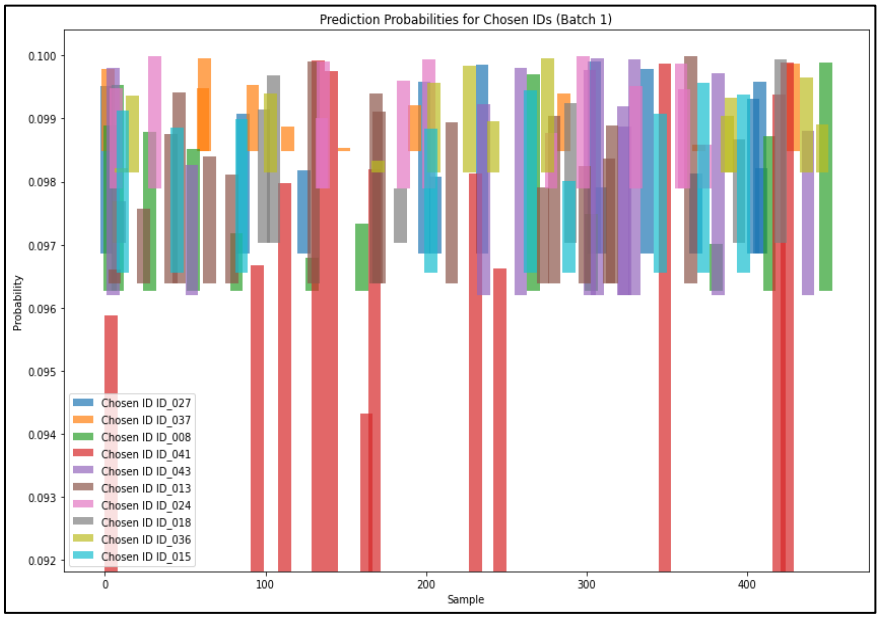


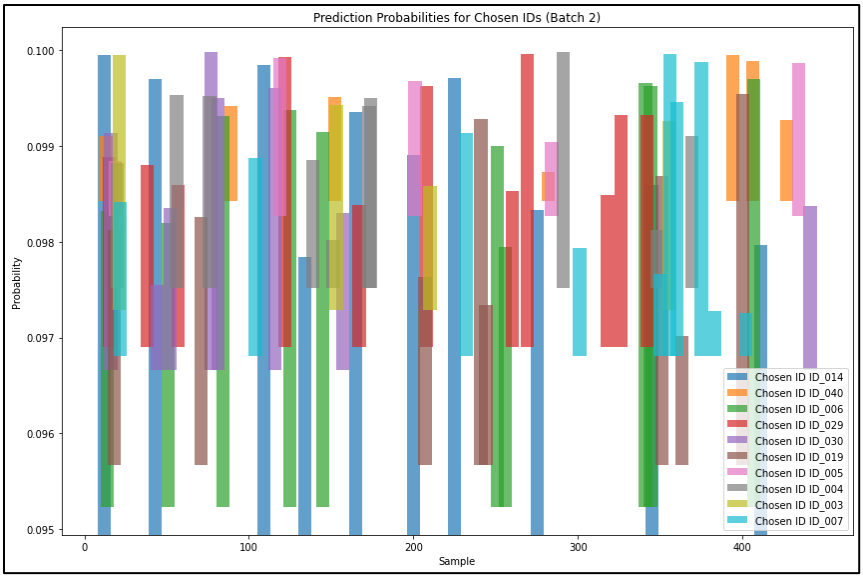
Figure 7. AUC-ROC Curve results for CNN vs Hybrid CNN+Vit Model.

4.7. Novelty Iris Prediction ID Result

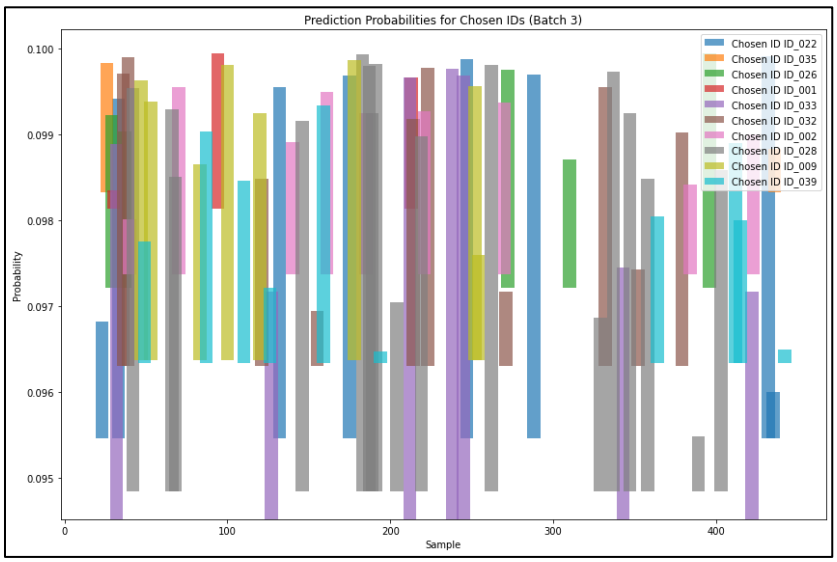
The novel Iris prediction ID model is shown to be effective in classifying authenticated irises using a Hybrid CNN+ViT architecture. The model analyses 45 distinct IDs, each of which is associated with 10 authenticated images 5 from the left eye and 5 from the right, as well as 450 tonal images for 45 classes. The program accurately distinguishes authenticated users by assigning the highest likelihood to the most likely ID. The findings for each ID type were examined to calculate probability, with authenticated IDs receiving greater values. Figure 8 (A, B, C, D, and F) shows a performance visualization that displays the model's accuracy over multiple samples, proving its classification capacity.



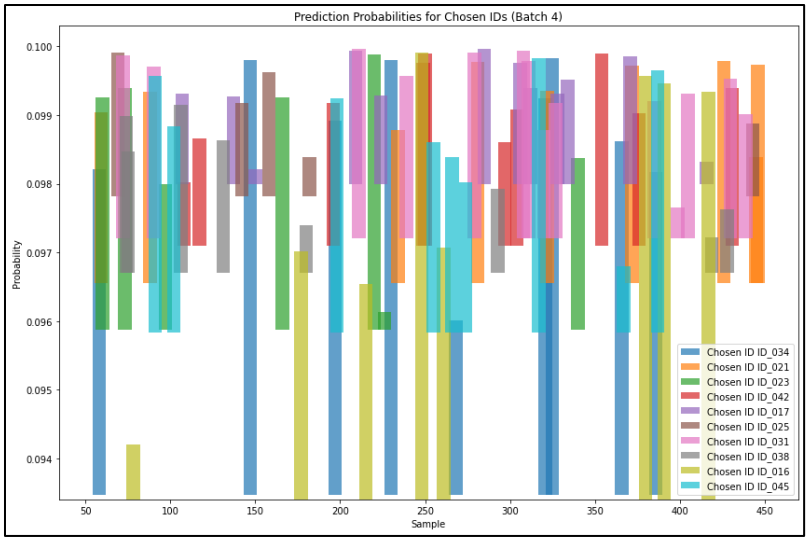
A


Prediction Probabilities for Chosen IDs (Batch 2)
Probability
Sample
0.100
0.099
0.098
0.097
0.096
0.095
0
100
200
300
400
Chosen ID ID_014
Chosen ID ID_040
Chosen ID ID_006
Chosen ID ID_029
Chosen ID ID_030
Chosen ID ID_019
Chosen ID ID_005
Chosen ID ID_004
Chosen ID ID_003
Chosen ID ID_007

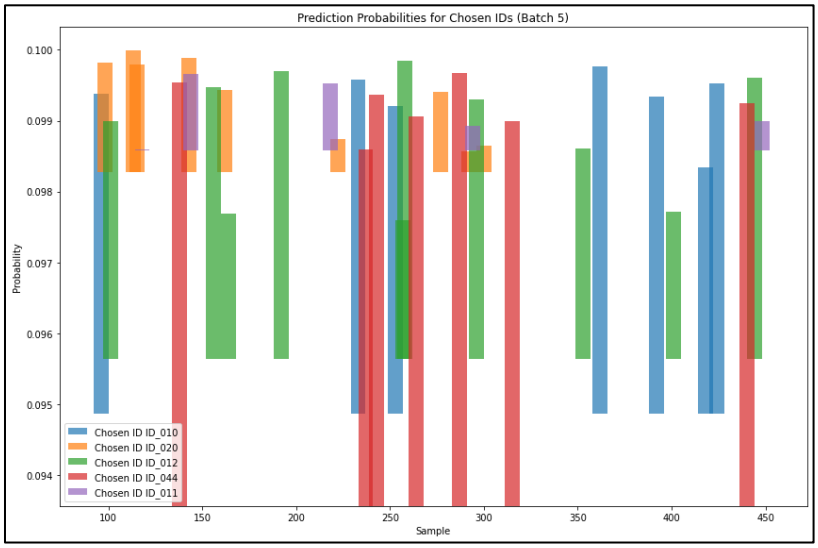
B


Prediction Probabilities for Chosen IDs (Batch 3)
Probability
Sample
0.100
0.099
0.098
0.097
0.096
0.095
0
100
200
300
400
Chosen ID ID_022
Chosen ID ID_035
Chosen ID ID_026
Chosen ID ID_001
Chosen ID ID_033
Chosen ID ID_032
Chosen ID ID_002
Chosen ID ID_028
Chosen ID ID_009
Chosen ID ID_039

C



D



F

Figure 8. Novelty iris prediction ID result; A) Prediction ID, Bach 1, B) Prediction ID, Bach 2, C) Prediction ID, Bach 3, D) Prediction ID, Bach 4, F) Prediction ID, Bach 5.

The figure illustrates the version's overall performance over five batches of iris samples. Batch One suggests how the model handles prediction IDs, giving the opportunity distribution for every and altering values relative to their baseline. Batch Two maintains this process with the aid of highlighting the converting prediction probabilities as the model actions via the subsequent set of samples. In Batch Three, the version's accuracy is illustrated by way of ranking the predicted IDs with the aid of chance, offering a comparison view of its overall performance. Batch Four emphasizes the version's consistency by assigning a possibility to a fresh batch of samples, demonstrating how the version performs with a higher range of authenticated IDs. Finally, Batch Five summarizes the version's overall accuracy and consistency in predicting accurate IDs all-around Five batches.

5. CONCLUSION

In this paper, we created a complete iris identity device with the use of sophisticated techniques for iris segmentation, characteristic extraction, and category. The device shows promising consequences employing the use of the U-Net architecture for particular iris segmentation, as well as the Hybrid CNN + ViT model for improved characteristic extraction and type. The consequences reveal that the hybrid model outperforms the standalone CNN model, with higher validation accuracy and higher generalization throughout numerous datasets. Furthermore, the

unique IRU-Net model supplied in this paper substantially improves segmentation accuracy, indicating tremendous progress in iris identification obligations. The experimental effects reveal that the models are useful in biometric applications, with high training and validation accuracies, little overfitting, and extraordinary generalization. For future work, improvements may be made utilizing integrating extra advanced transformers or hybrid architectures to enhance the version's robustness, particularly in instances of bad-first-rate pix or noisy environments. Additionally, expanding the dataset to consist of more diverse iris samples from one-of-a-kind demographic groups can enhance version generalization. Future studies could also discover the optimization of the fashions for actual-time iris reputation applications and check out the potential of integrating different biometric modalities to create a more complete identification device. With these upgrades, the proposed gadget may be extended to provide extra reliable and scalable iris recognition solutions for a variety of actual-world applications. Future work can increase the model's robustness by way of using more advanced transformers or hybrid designs, specifically in situations of low-excellent pics or noisy surroundings. Furthermore, growing the dataset's range with the aid of which includes iris samples from various demographic companies can increase model generalization. Future studies can also look into optimizing the models for actual-time iris recognition applications and the possibility of incorporating other biometric modalities to create a extra entire identifying device. With these upgrades, the counseled system can be accelerated to deliver extra reliable and scalable iris identification answers for several real-global applications.

REFERENCES

[1] Latif, S.A.; Sidek, K.A.; Hashim, A.H.A. An Efficient Iris Recognition Technique Using CNN and Vision Transformer. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2023**, *34*, 235–245.

[2] Wang, C.; Muhammad, J.; Wang, Y.; He, Z.; Sun, Z. Towards Complete and Accurate Iris Segmentation Using Deep Multi-Task Attention Network for Non-Cooperative Iris Recognition. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 2944–2959.

[3] Odinokikh, G.; Korobkin, M.; Solomatin, I.; Efimov, I.; Fartukov, A. Iris Feature Extraction and Matching Method for Mobile Biometric Applications. In *Proc. Int. Conf. Biometrics (ICB)*; IEEE: **2019**; pp. 1–6.

[4] Rana, H.K.; Azam, M.S.; Akhtar, M.R.; Quinn, J.M.; Moni, M.A. A Fast Iris Recognition System Through Optimum Feature Extraction. *PeerJ Comput. Sci.* **2019**, *5*, e184.

[5] Danlami, M.; Jamel, S.; Ramli, S.N.; Deris, M.M. A Framework for Iris Partial Recognition Based on Legendre Wavelet Filter. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*.

[6] Ahmadi, N.; Nilashi, M.; Samad, S.; Rashid, T.A.; Ahmadi, H. An Intelligent Method for Iris Recognition Using Supervised Machine Learning Techniques. *Opt. Laser Technol.* **2019**, *120*, 105701.

[7] Farouk, R.H.; Mohsen, H.; El-Latif, Y.M.A. A Proposed Biometric Technique for Improving Iris Recognition. *Int. J. Comput. Intell. Syst.* **2022**, *15*, 79.

[8] Szymkowski, M.; Jasiński, P.; Saeed, K. Iris-Based Human Identity Recognition with Machine Learning Methods and Discrete Fast Fourier Transform. *Innov. Syst. Softw. Eng.* **2021**, *17*, 309–317.

[9] Jenadeleh, M.; Pedersen, M.; Saupe, D. Blind Quality Assessment of Iris Images Acquired in Visible Light for Biometric Recognition. *Sensors* **2020**, *20*, 1308.

[10] Wei, Y.; Zhang, X.; Zeng, A.; Huang, H. Iris Recognition Method Based on Parallel Iris Localization Algorithm and Deep Learning Iris Verification. *Sensors* **2022**, *22*, 7723.

[11] Dosovitskiy, A. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.

[12] Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Inception Recurrent Convolutional Neural Network for Object Recognition. *Mach. Vis. Appl.* **2021**, *32*, 1–14.

[13] Adjei-Mensah, I.; et al. Investigating Vision Transformer Models for Low-Resolution Medical Image Recognition. In *Proc. 18th Int. Comput. Conf. Wavelet Active Media Technol. Inf. Process. (ICCWAMTIP)*; IEEE: **2021**; pp. 179–183.

[14] Wang, C.; Wang, Z. Progressive Multi-Scale Vision Transformer for Facial Action Unit Detection. *Front. Neurorobot.* **2022**, *15*, 824592.

[15] Jamil, A.H.; Yakub, F.; Azizan, A.; Roslan, S.A.; Zaki, S.A.; Ahmad, S.A. A Review on Deep Learning Application for Detection of Archaeological Structures. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2022**, *26*, 7–14.

[16] Khan, U.; Pao, W.; Sallih, N.; Hassan, F. Flow Regime Identification in Gas–Liquid Two-Phase Flow in Horizontal Pipe by Deep Learning. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2022**, *27*, 86–91.

[17] Yow, S.C.; Ali, A.N. Iris Recognition System (IRS) Using Deep Learning Technique. *J. Eng. Sci.* **2019**, *15*, 125–144.

[18] Hu, Q.; Yin, S.; Ni, H.; Huang, Y. An End-to-End Deep Neural Network for Iris Recognition. *Procedia Comput. Sci.* **2020**, *174*, 505–517.

[19] Ríos-Sánchez, B.; Silva, D.C.; Martín-Yuste, N.; Sánchez-Ávila, C. Deep Learning for Face Recognition on Mobile Devices. *IET Biom.* **2020**, *9*, 109–117.

[20] Li, H.; Zhang, L. Multi-Exposure Fusion with CNN Features. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*; IEEE: **2018**; pp. 1723–1727.

[21] Liu, G.; Zhou, W.; Tian, L.; Liu, W.; Liu, Y.; Xu, H. An Efficient and Accurate Iris Recognition Algorithm Based on a Novel Condensed 2-Channel Deep Convolutional Neural Network. *Sensors* **2021**, *21*, 3721.

[22] Wang, C.; Wang, Y.; Xu, B.; He, Y.; Dong, Z.; Sun, Z. A Lightweight Multi-Label Segmentation Network for Mobile Iris Biometrics. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*; IEEE: **2020**; pp. 1006–1010.