International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 344-365
https://doi.org/10.70917/ijcisim-2026-2081
Article

FACE RECOGNITION USING HYBRID QUANTUM CONVOLUTIONAL NEURAL NETWORK (QCNN) AND IMPROVED FEATURE EXTRACTION TECHNIQUES

Aswathy R[1], A. Sherin[2]

[1] Department of Computer Science, Nehru Arts and Science College. aswathyravi998@gmail.com
[2]Department of Computer Science, Nehru Arts and Science College. sherinfaizalrahiman@gmail.com
Corresponding Author: Aswathy R (aswathyravi998@gmail.com)**

Abstract: Face recognition is an advanced biometric technology that focuses on accurately identifying facial features to differentiate individuals, even when faces appear highly similar or identical. Modern face recognition systems are much more dependable and effective at managing difficult obstacles because of rapid growth of computer vision and deep learning, incorporating variations in lighting, position, facial expressions, and occlusions. Analysing minute variations in facial traits that may not be easily discernible to the human eye is the main objective of same face identification. Additionally, since the human face is a three-dimensional (3-D) entity, it may be illuminated unevenly and with a distorted perspective. Consequently, it may not be possible to identify a real face. This research work introduces a hybrid automatic detection and intelligent feature extraction model for face recognition that leverages two-dimensional (2D) facial images sourced from diverse origins to construct three-dimensional (3D) face meshes using 468 MediaPipe landmarks. Preprocessing pipeline includes use of Restormer to improve image quality, then use of segmentation based on deep learning algorithms like Mask R-CNN to isolate the foreground (face) from the background to decrease noise and increase detection accuracy. Extraction of features is done by use of Vision Transformer (ViT). Features are then classified through hybrid deep classifiers in the form of FaceNET (FN) with quantum convolutional neural network (QCNN). This is to ensure accuracy despite the presence of different facial poses and expressions..

Keywords: Face recognition, two-dimensional (2D) facial images, three-dimensional (3D), pre-processing, Segmentation, Feature Extraction, Vision Transformer (ViT), hybrid deep classifier as Face NET model (FN) with Quantum Convolutional neural network (QCNN).

1. INTRODUCTION

AI has made many advancements in the last few years, and its influence can be seen in the development of technologies like autonomous vehicles and automated supermarket checkouts. One of the areas related to artificial intelligence that has been widely recognized is Computer Vision. Humans utilize visual sensing to understand their surroundings, while computer vision aims at achieving the same with electronic devices capable of sensing, processing, and analyzing visual data to derive some meaning out of it. Computer vision is not only about detecting visual inputs passively but also about being able to react to the reasons for which these visual inputs occur. Computer vision is thus capable of detecting and analyzing visual inputs in a similar way that humans do. For example, if a person appears suddenly in front of a moving car, the driver needs to quickly understand the situation and act accordingly [3]. Here, human vision works in three basic stages, namely perception, cognition, and decision-making. It is an important goal for computer vision to carry out these procedures effectively. Vision, including skills such as coordination, recollection, recall, inference, estimation, and identification, is an integral part of intelligence. For a system to possess just one skill means that it does not comprise a total vision system [4]. Computer vision attempts to replicate these human skills combined into one. An important problem that exists because of this is that visual sensors capture only two dimensional images whereas the real world we are dealing with now is three dimensional.

In computer vision and artificial intelligence, face image recognition is one of the most significant fields of study. A lot of real-life applications, like security, human identification, surveillance, access control, and human-computer interaction, use automatic face image recognition due to the fast evolution of modern digital technology. Shallow machine learning techniques have several disadvantages when applied to face recognition tasks with varying pose, lighting conditions, and facial expressions. This method uses expert knowledge along with handcrafted features to acquire simple properties of an image. The deep learning approach allows for automated discovery of complex face properties. Deep learning has enabled a lot of development to be made in solving problems that previously limited the growth of artificial intelligence. The deep learning approach is applicable in many different areas because it has demonstrated remarkable skill in identifying complex patterns in high dimensional data. Through the ability to perform hierarchical learning using unified methods, it has performed exceptionally well in areas such as semantic segmentation, image recognition, natural language processing, and many more practical uses. Some of the deep learning models include the use of Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), and stacked autoencoders [7]. The CNN is usually used in problems related to face and image recognition. A neural network model known as CNN is an artificial neural network that employs convolutional operations to automatically learn the features in the input data.

By comparing a face to the facial database, a face recognition (FR) system may identify a face. Due to advancements in feature and face recognition model design and learning, it has made significant strides in recent years. Given that humans are highly efficient at identifying individuals regardless of their age, lighting, or facial expressions [8]. The development of a FR system is the primary objective of researchers capable of matching or exceeding human-level accuracy, which is approximately 97.5%. The choice of techniques in high-performance facial recognition systems typically depends on the specific application. Broadly, there are two primary categories into which facial recognition systems are classified:

By comparing a query image to a massive collection of face image images, one may identify a certain individual. Such systems return the corresponding identity or associated details of the matched person. Typically, for each individual, a single reference image is provided, and real-time processing is not always required.

Perform real-time identification of individuals, typically in access control systems that grant or restrict entry based on identity. These algorithms require several training images of each individual as well as real-time recognition capabilities. The system suggested in this paper belongs to this class, as its design is concerned with overcoming difficulties caused by the variety of faces, facial expressions, and viewing angles.

The four main stages of the conventional face recognition method are face detection, face alignment, feature extraction (facial representation), and classification [9]. The suggested method extracts face characteristics from input images and provides deep neural networks for training and classification, typically using a softmax layer. The network architecture is flexible, allowing layers to be adjusted to optimize performance. In recent years, various libraries, tools, and platforms have facilitated the design and customization of such models. A specific type of neural networks termed Convolutional Neural Networks (CNNs) is made to analyze input having an organized, grid-like architecture, like image data. These networks have demonstrated remarkable success across practical applications, comprising image data, which is modelled as two-dimensional pixel grids, and time-series data, which is represented as regularly sampled one-dimensional grids. CNNs are distinguished by the use of convolution operations in at least one layer rather than using standard matrix multiplication [10]. The term “convolutional neural network” reflects this reliance on the mathematical operation of convolution, which is a specialized form of linear transformation.

Several advanced deep learning models such as FaceNet and DeepFace have demonstrated remarkable accuracy in face recognition tasks [11]. By learning high-level feature embeddings, these models enable efficient comparison of facial images and improve the robustness of recognition systems. Therefore, deep learning-based facial image identification has emerged as a promising method for developing highly accurate and reliable biometric identification systems. This research proposes to examine deep learning methods for facial image recognition and assess how well they can enhance recognition performance and accuracy. This research work introduces a hybrid automatic detection and intelligent feature extraction model for face identification that uses 468 Media Pipe landmarks to create three-dimensional (3D) face models from two-dimensional (2D) facial images from various sources.

The rest of the research is scheduled as follows: Several of the recent methods for facial recognition employing various deep learning and feature extraction techniques are summarized in section 2. The proposed technique's procedure is shown in Section 3. The results and comments are presented in Section 4. The conclusion and future work are addressed in Section 5.

2. LITERATURE REVIEW

This section reviews some of the recent techniques for the recognition of face using different deep learning techniques and feature extraction methods.

Author and year	Methods	Advantages	Disadvantages
Pandey et al [2019]	Convolution Neural Network (CNN)	the proposed approach has improved the performance of face recognition with better results of recognition	Lack of transparency makes it difficult to trust model decisions.
Nachet et al [2022]	Multi-Task Convolutional Neural Network (MTCNN)	The proposed model has an accuracy rate of 97.50%.	Faces captured at different angles (side view, tilted, rotated) reduce recognition accuracy.
Guptaet al [2018]	deep neural network	This method providing the accuracy of 97.05% on Yale faces dataset.	Face datasets are often small and imbalanced due to privacy concerns and difficulty in data collection.
Hussain et al [2020]	convolutional neural network model	High precision rate	The variability affects model accuracy and generalization.
Benradi et al [2023]	SIFT+CNN combination	optimization algorithm used during training (adam, Adamax, RMSprop, and stochastic gradient descent (SGD)) for best results	Ground truth may not always be precise.
Ghasemzadeh et al [2018]	collaborative representation-based classifier (CRC)	proposed 3D-DWT methods is superior to alternative methods using spatio-spectral classification	Models may become biased toward majority classes.
Vishwakarma et al [2020]	Hybridization of fractional discrete Cosine transform (FrDCT) and DWT	very good performance outcome has been achieved as compared to the other state of art methods	Changes in lighting conditions (bright, dim, shadows) significantly affect facial features

Bendjillali et al [2019]	CNN network	The face recognition rate of the ORL face database and the AR face database based on this network achieved 99.85% and 99.80% respectively.	Makes it difficult to extract consistent features.
Almabdy et al [2019]	convolutional neural network (CNN)	The result showed that our model achieved a higher accuracy	Same face may appear very different under varying illumination.
Udawant et al [2024]	Convolutional Neural Network (CNN)	The proposed model with an accuracy rate of 97.50%,	Smiling, frowning, or other expressions alter facial geometry.
Tao et al [2019]	face recognition model	The accuracy rate obtained on the test set was 97.63%.	Deep models like FaceNet and Convolutional Neural Network act as black boxes.

Recent advancements in deep learning have significantly improved the performance of face recognition systems. Deep learning models, especially the Convolutional Neural Network, are capable of automatically learning complex and hierarchical facial features from large datasets. However, these models can detect fine-grained details in facial images, making them more effective for distinguishing between highly similar or identical faces.Modern deep learning frameworks such as FaceNet and DeepFace generate compact feature representations known as face embeddings. These embeddings allow the system to measure similarity between facial images using mathematical distance metrics, thereby enabling accurate identification even when the differences between faces are minimal. Therefore, identical face image recognition using advanced deep learning has become a promising approach for improving the reliability and accuracy of biometric systems. By leveraging quantum neural network architectures and large-scale training datasets, modern face recognition systems can effectively analyze subtle facial variations and provide robust identification in real-world applications.

3. PROPOSED METHODOLOGY

The existing research proposes a system that integrates an automatic face detection and intelligent feature extraction model for face recognition, which uses two-dimensional (2D) face images from various sources to develop three-dimensional (3D) 468 MediaPipe landmarks. In the pre-processing phase, the images are improved using Restormer, and then they undergo segmentation through deep learning algorithms mask R-CNN framework use to distinguish the foreground (face) from the background. Feature extraction is done through the Vision Transformer (ViT). These feature vectors are then trained through a hybrid deep classifier, called FaceNET model (FN) together with Quantum Convolutional Neural Network (QCNN). This model is specifically designed to retain its accuracy even in cases where there is a wide range of poses, expressions, and occlusions.

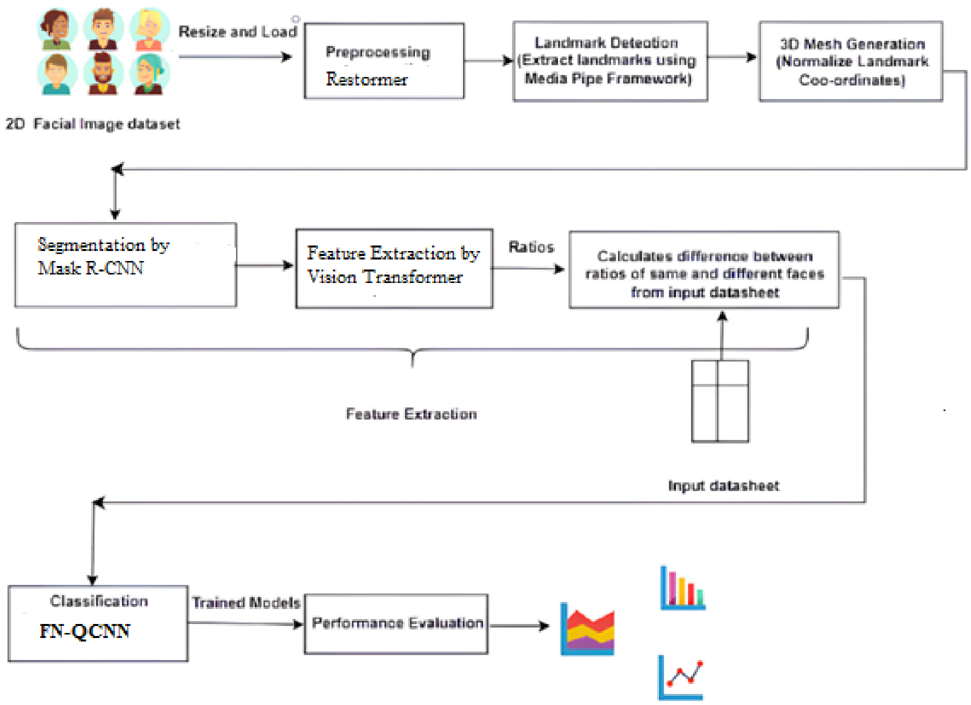


Figure 1. The proposed methodology's overall procedure

3.1. Data Collection

One of the most commonly used benchmark databases for exploring problems related to unconstrained facial recognition has been known as LFW. Researchers from Amherst progressed. Within 13,233 online photographs comprising 5,749 people, faces were detected and aligned using the Viola-Jones algorithm. Of these people, 1,680 have at least two photographs available. Different variations of this database exist; four original photographs and three different alignments of images are present. Importantly, deep funneled images have reportedly shown better results than other images.

3.2. Pre-processing

Face images are captured for obtaining the standardized inputs in order to perform detection and normalization processes. The proposed method provides a standard set of face images in the following way; firstly, selecting a face image is done, followed by its preprocessing through methods such as cropping and scaling. Noise removal is carried out by utilizing the Adaptive Median Filter (AMF) method that effectively eliminates any noise present in FERC images. AMF uses spatial algorithms to detect noisy pixels in the image [23]. They compare pixels with their neighbours for these determinations where both comparison criteria and neighbourhood sizes can be varied [24]. Noisy pixels can be defined as pixels that differ from majority of their neighbours and are not physically aligned with their comparable pixels. Median pixel values of nearby pixels that pass noise labelling tests are then used to replace noisy pixels. Therefore, they are applied to lower noises in FERC images and enhance image qualities.

Restormer

A transformer-based neural network termed restormer was designed specially to handle image restoring tasks. Unlike traditional convolution-based networks, restormer collects global context of sight and long-range interdependence using attention methods, which helps in restoring fine details in degraded images.

Based on the Transformer architecture, Restormer is a modern neural network architecture that improves the conventional feed-forward and self-attention processes with an encoder-decoder design [25]. In order to enhance performance, it uses a gradual learning strategy. One of the most significant strengths of the Restorer network is that it can generate multiple-level representations of high-resolution pictures directly into the global scope without needing window partitioning. It is feasible to preserve critical contextual information within the model due to such an architectural design, which is vital for effective image restoration. There are two key elements in this architecture;

these include the Multi-Dconv Head Transposed Attention (MDTA) and Gated Dconv Feed-Forward Network (GDFN). The MDTA module substitutes the standard multi-head self-attention model and facilitates the successful incorporation of both local and global dependencies among pixels by generating the transformed attention maps using the cross-covariance function between feature channels. In contrast, the GDFN module boosts the performance of the classic feed-forward neural network through the use of a gating module and depth-wise convolutions. Furthermore, Restorer uses a progressive training method where the study starts with smaller image patches and gradually moves to large areas at some point in the training. This sequential training technique makes it possible to acquire worldwide contextual facts in image extra effectively. Restores generate gigantic overall performance will increase at some point in the assessment using modern studies. All things considered, Restorer is a totally popular version of image enhancement that combines innovative education and solid international dependency modelling to offer progressive results on image restoration commitments.

3.3. Face Detection

Using patterns from learnt data, face detection determines if an image has a face. This task is inherently challenging due to variations in environmental conditions, illumination, motion, pose, and facial expressions. The MediaPipe framework simplifies this process by enabling efficient face detection and landmark extraction [26]. In the proposed approach, MediaPipe's face detection model is utilized to perform real-time detection from image or video inputs.

3.3.1. Landmark Detection

This method enables it possible to identify face landmarks with accuracy and dependability. A set of precise reference mappings is applied to individual facial images to obtain the required measurements, allowing automatic feature localization and extraction. In this work, 468 landmarks are extracted rather than the traditional 68 to use the MediaPipe library to increase accuracy. The MediaPipe framework also supports multi-face detection and estimates facial landmarks in a 3D space. Real-time landmark estimate is made possible by its face geometry solution, especially on systems with limited resources like mobile platforms.

3.3.2. 3D Face Mesh Generation

A face mesh, or 3D model of the face, was created by superimposing all of them that were taken from the 2D facial images over the 3D image. For every face that is recognized, a 468-point, edge, and triangular face mesh is produced using the face mesh detection API.

3.4. Segmentation using Mask R-CNN architectures

Instance segmentation aims to detect and distinguish each object present in an image by assigning unique pixel labels to individual instances. It operates by classifying every pixel and generating a corresponding segmentation mask for each object [27]. This approach provides a greater comprehension of the scenario material, as it preserves object boundaries while enabling precise recognition.

A branch for object recognition and pixel-wise segmented mask prediction is added to Faster R-CNN in Mask R-CNN. It evolved to segment instances, which involves localizing items in an image and properly identifying their pixels.

Mask R-CNN Architecture

The inclusion of a segmented prediction layer is the only difference in Mask R-CNN and Faster R-CNN. The second stage guesses the class, develops a boundary box, and generates binary masks for each RoI in parallel. Neither architecture has a different area proposal stage.

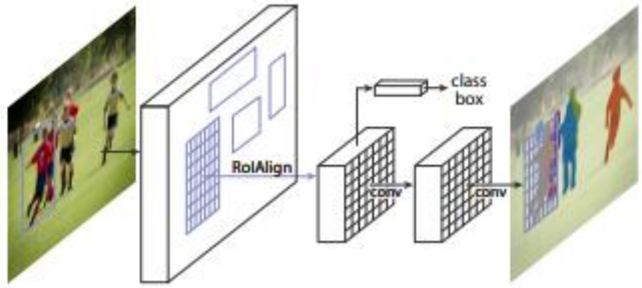


figure 2. Mask R-CNN

Backbone Network

A basic ResNet configuration (ResNet-C4) and a ResNet combined with a Feature Pyramid Network (FPN) were the two backbone network options that the developers of Mask R-CNN assessed. The ResNet-C4 setup closely follows the backbone used in Faster R-CNN, whereas the ResNet-FPN introduces architectural enhancements. Specifically, the FPN employs a multi-scale feature hierarchy for Region of Interest (RoI) generation, enabling the extraction of features at different resolutions. This multi-level representation improves detection and segmentation accuracy compared to the standard ResNet support. At each stage of the network, the number of channels is increased, but the feature maps' spatial resolution is reduced in half. Feature outputs are extracted from four stages (layers 1–4). To construct the final feature representations, starting with the highest-level feature map (w/32, h/32, 256), a top-down route is used and progressively upsampling to higher-resolution maps, thereby integrating multi-scale information. Prior to upsampling, a (1 * 1) convolution reduces channels to 256. Following that, element-wise addition is used to merge the resultant feature map, with the upsampled output from the preceding stage. Subsequently, each combined feature map is processed using (*3 X 3*) convolutional layers to produce the final set of feature maps (P2, P3, P4, P5). An additional level, (P6), is obtained by applying a max-pooling operation to (P5).

Region Proposal Network

The layer's convolution feature map before a 3*3 convolution layer processes data. After that, two parallel branches receive the result and use it to regress the bounding box coordinates and get the objectness score.

Various feature maps of varying sizes, which are used to search for objects, we utilize one anchor step and three anchor proportions for a feature hierarchy.

Mask Representation

The spatial details of an item are contained in a mask. Since it required pixel-to-pixel connectivity from the layer above, it couldn't condense output to a fully connected layer to enhance categorization and boundary box analysis. R-CNN predicts masks using a fully connected network. After receiving a RoI as input, this ConvNet produces the m*m mask representation. Additionally, we use 1*1 convolution to compress the channels to 256 and to make inferences from the input image, upscale this mask. It utilizes RoIAlign to provide our fully connected mask-predicting network. The region proposal network generates feature maps of different sizes, which are then converted into fixed-size feature maps using RoIAlign. Two distinct architectures were suggested by the Mask R-CNN research. The fully connected layer (FPN Network) receives the input of the mask-generating CNN in one form simply before, whereas it is passed after the application of RoIAlign in another (ResNet C4). The overall convolution network output of this mask implementing branch is K * (m*m), where m = 14 for ResNet-C4 and 28 for ResNet_FPN and K is the number of classes.

RoI Align

According to RoI pool, region suggestions are used by RoI align to produce fixed-size regions of interest.

Mask R-CNN uses Faster R-CNN's two-stage structure with an instance segmentation approach, as illustrated in Figure 2. The primary advancement lies in the integration of a third component referred to as the mask branch which functions alongside the region proposal and classification branches, allowing each identified item to have a segmentation mask created by the model. A binary segmentation mask is predicted for each region of interest by the mask branch, capturing exact pixel-level object boundaries, after the region proposal network (RPN) generates candidate regions. Consequently, Mask R-CNN is capable not only of detecting and classifying objects but also of generating detailed instance-specific segmentation masks. The mask branch employs a pixel-level alignment strategy commonly realized through spatially precise RoI feature extraction to maintain accurate correspondence between each proposed region and its predicted mask. As a result, Mask R-CNN produces, for every detected object, a high-resolution pixel-wise segmentation mask, a bounding box, and a corresponding class name.

3.5. Feature Extraction using Vision Transformer (ViT)

For computer vision and deep learning image analysis, the Vision Transformer was designed. It processes images by dividing them into patches and applying transformer-based attention mechanisms to extract meaningful features [28]. The Vision Transformer algorithm performs image recognition by:

Splitting images into patches

Encoding patches into embeddings

Learning global relationships using transformer attention

Producing predictions using the CLS token representation

The ViT constructs connections between regional and distant pixels using attention. The input image divides into sections to activate the attention mechanism. This method is essentially similar to applying a convolutional layer with a fixed kernel, resulting in a four-dimensional tensor structured by batch size, spatial dimensions (rows and columns), and feature depth. As a result, I ∈ $\mathbb{R}H\times W\times C$ is converted into PP ∈ $\mathbb{R}N\times P2\times C$, where C represents H and W represents image width, height, and channel number. In contrast, N is the quantity of patches determined as,

$$N = \frac{H\times W}{p^2} \qquad (1)$$

where P stands for patch size. All the patch and the image have dimensions of 6 x 6 and 78 × 78. The number of patches is computed from the image and patch size: ($H\times WP$2)=78×78(6)2=169. Following patch partition, the raw image (I) is linearly projected into a 64-dimensional 1D embedding vector from a 2D matrix, PP, $PPLinear\_Projection$:

$$PP_{Linear\_Projection} = [[\ I_1^1\ I_1^2\ I_1^{64}]\ [I_2^1\ I_2^2\ I_2^{64}]\ [I_{169}^1\ I_{169}^2\ I_{169}^{64}\ ] \qquad (2)$$

Due to the high computational cost associated with transformer architectures, the image patches are first encoded using positional embedding, allowing them to be processed efficiently in structured groups and scaled to larger image resolutions. The positional encoding, denoted as $EPOS$, is computed using a combination of sine and cosine functions at varying frequencies, thereby preserving spatial information within the sequence of embedded patches. For positional encoding, cosine functions are applied to patches at odd positions, while sine functions are used for those at even positions. In this formulation, pos denotes the position of a patch within the sequence, and i represents the embedding dimension. The positional information is thus encoded across different dimensions using sinusoidal functions of varying frequencies, where d indicates the maximum length of the patch sequence. Subsequently, each linearly projected patch representation is combined with its corresponding positional embedding, resulting in the final embedded patch representation.

$$E_{POS} = \{sin\left(\frac{pos}{1000^{\frac{2t}{d}}}\right), \qquad t\ is\ even\ cos\left(\frac{pos}{1000^{\frac{2t}{d}}}\right), \qquad t\ is\ odd \qquad (3)$$

$$EP = concatenate(PP_{Linear\ projection}, E_{POS}) \qquad (4)$$

The generated embedded patches are sent to the encoder after linear projection and positional encoding. A multi-layer perceptron (MLP), a multi-head self-attention (MSA) module, and a normalizing layer are among the six layers that comprise each of the eight identical blocks that comprise the encoder. Initial, the encoder block's input (represented by EP) combines with the output of the MSA module through a residual connection, facilitating effective feature propagation and stable training. Subsequently, the intermediate representation is processed through a normalization layer, followed by an MLP that incorporates layers using dropout for regularization that are completely connected. A skip (residual) connection links the original input to the attention output, reinforcing positional information by reintroducing the initial embedded patch to the subsequent layer. The attention mechanism is computed using three projection matrices key K, query Q, and value V which are derived through learned weight matrices W_Q, W_K, and W_V, respectively, as expressed in the following formulations:

$$Query,\ Q = EP.WQ \qquad (5)$$

$$Key,\ K = EP.WK \qquad (6)$$

$$Value,\ V = EP.WV \qquad (7)$$

where the weight matrices are and EP is the embedded patch W_Q, W_K, $W_V \in \mathbb{R}dmodel\times dk$. Equation (14) is used to determine the attention for the MHA layer's head, a single attention function that is executed several times in parallel. where the attention value is kept from being deleted by the dot-scaled product dk. The correlation between two visual patches may be scored using this attention value. The proposed framework depicts the standard four-headed multi-head attention (MHA) as follows:

$$Attention, (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (8)$$

$$MHA = Attention, (Q, K, V)_{\times 4} \qquad (9)$$

The subsequent layers use the multi-layer perceptron, it uses a dense layer with Gaussian error linear unit (GELU) activation to produce non-linearity in this process. The Gaussian distribution's cumulative distribution is denoted by ϕ. After removing the layer's output, which is Transformer, we flatten it using the following formulas:

$$GELU\ (x) = xP\ (X < x) = x\phi\ \ (x)$$

$$Transformer_{feature\_shape} = GELU\ (MHA)$$

$$Y_{Transformer} = flatten\ (Transformer_{featur\_shape})$$

Following that, the classifier for deep neural networks is given the $Y_{Transformer}$. Additionally, the ViT network's transformer feature extractor's optimum parameters. On the validation dataset, the model's performance was taken into consideration while adjusting the parameters, based on dataset size and complexity, these were projected.

3.6. Classification using hybrid deep classifier as FaceNET model (FN) with Quantum Convolutional neural network (QCNN)

To increase the accuracy of facial recognition systems, a hybrid deep classifier that combines FaceNet and QCNN combines traditional deep learning with quantum computing methods. In difficult problems like the identification of same or similar faces in computer vision and quantum machine learning, this method is extremely useful. For sophisticated categorization, the hybrid model leverages the advantages of both the FaceNet and the Quantum Convolutional Neural Network.

3.6.1. Facenet

FaceNet employs deep convolutional networks to learn compact embedding representations directly, rather than relying on intermediate bottleneck layers used in earlier approaches. It follows a one-shot learning paradigm, enabling the model to generalize from a limited number of face samples. As a result, once trained, it can recognize new faces without requiring retraining. FaceNet learns facial representations in a Euclidean embedding space, where distances between feature vectors reflect the similarity between faces. This representation enables efficient face recognition and classification by comparing the distances among the learned embedding. Positive (same identity), negative (different identity), and anchor images compose each triplet in FaceNet's triplet-based learning technique with online triplet mining during training. The architecture produces discriminative face embedding by processing input batches via a deep convolutional neural network and L2 normalization.

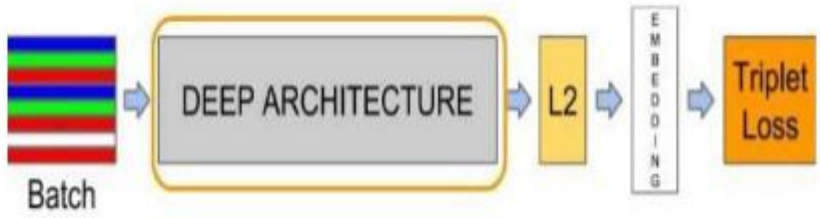


Fig. 3. The model structure of the FaceNet

FaceNet training uses a triplet loss function with an anchor, positive sample, and negative sample. The objective is to decrease the anchor's distance from the positive (same identity) and maximize its distance from the negative thereby improving the discriminative capability of the learned embedding.

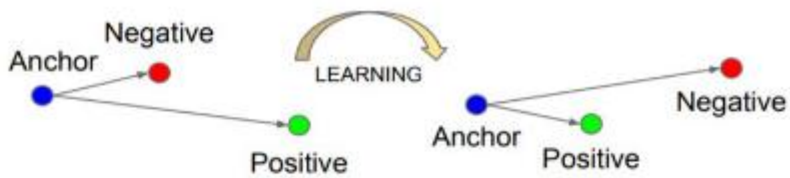


Fig. 4. The triplet loss training

A triplet-based loss function inspired by Large Margin Nearest Neighbour enables FaceNet learn compact 128-dimensional embedding. The training process relies on triplets composed of matched (positive) and unmatched (negative) face samples, with the loss function enforcing positive and negative pairings are separated by a margin. The input images are tightly cropped face thumbnails, requiring minimal preprocessing beyond basic scaling and translation, without the need for explicit 2D or 3D alignment.

3.6.2. Quantum Convolutional Neural Network (QCNN)

QCNNs create hybrid quantum classical learning systems by combining quantum computing with CNNs. It combines the hierarchical structure of classical CNNs with quantum operations such as measurements, entanglement, and prediction to enhance record and handle intricate data patterns.

In a quantum computing setting, data are represented using qubits, enabling the integration of convolutional neural network structures within a quantum framework. This section outlines the architectural design of the QCNN model, which extends the core components of classical CNNs namely convolution and pooling layers into quantum systems, as illustrated in Figure 5. The following is a description of the fundamental concept:

1) Applying many qubit gates between neighboring qubits helps the hidden state is identified by the convolution circuit.

2) By using 2-qubit gates like CNOT gates or monitoring qubit percentage, the pooling circuit reduces quantum system size.

3) Repeat the pooling and convolution circuits described in (1)–(2).

4) When the system size becomes sufficiently small, a fully connected quantum circuit is employed to produce the final classification output. The architecture commonly adopted for this purpose is the Multi-Scale Entanglement Renormalization Ansatz (MERA), which is designed for efficient simulation of quantum many-body states. In MERA, the quantum system's size grows exponentially with depth through the addition of qubits initialized in the $|0\rangle$ state. In contrast, a QCNN utilizes this structure in a reversed manner to progressively reduce the system size while extracting relevant features. The reversed application of MERA enables an exponential reduction in system size, making it well-suited for QCNN architectures. The QCNN framework proposed by Iris Cong further enhances this structure by incorporating QEC, leading to improved robustness and performance. Within the MERA representation, each class label is associated with a corresponding quantum state $|\psi\rangle$. Since QCNN operates as the inverse of Multi-Scale Entanglement Renormalization Ansatz (MERA), an input state $|\psi\rangle$ that is consistent with the MERA representation can be mapped to its corresponding class label as a well-defined solution. In contrast, if an input state $|\psi'\rangle$ lies outside the space generated by MERA, the QCNN may fail to produce a definitive classification. This

limitation can be mitigated by incorporating Quantum Error Correction (QEC), which introduces additional degrees of freedom and enhances the model's capacity to handle such states.

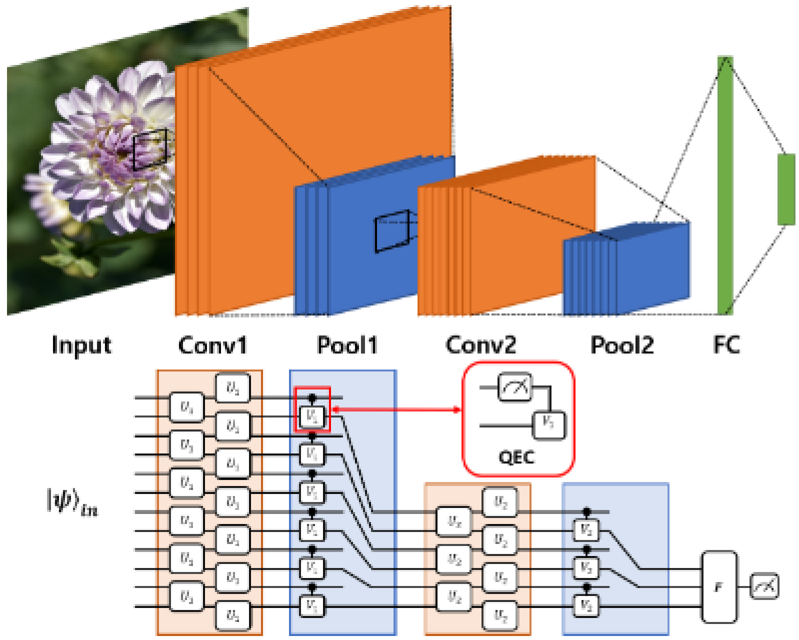


Fig. 5. Simple example of CNN and QCNN.

According to the new provided state in MERA, in the pooling layer, the result should be |0〉 , when the data supplied to QCNN is |ψ〉 . However, it is possible that the measured result will be |1〉 if the input data is |ψ′〉 , which MERA cannot produce. They are used to adjust the result if |1〉 is measured by applying an extra gate to the surrounding qubits. Through more predictable measurement outputs, the approach may provide improved performance. However, optimization of quantum circuits may experience barren plateaus, when gradients become small and learning slows down, making the training of QCNN models challenging. Increasing the QCNN's performance by the use of parameter initialization, which was introduced in this study.

Parameter initialization

Assuming that *itter(ψi)* indicates the number of iterations that classifier Ψi has spent in the network and Pa(Ψi) indicates the frequency of accurate prediction of Ψi. The weight ω(Ψi) of the classifier is expressed as follows:

$$P_a(\Psi_i) > P_a^{\pi} \text{ then } \omega(\Psi_i) = P_a(Psi_i) \qquad (9)$$

else
$$\omega(\Psi_i) = \frac{P_a(\Psi_i)}{\sqrt{itter(\Psi_i)}} \qquad (10)$$

where the average classifier accuracy in the hybrid model π is represented by P_a^{π}. The model Ψ's final prediction is found to be:

$$\Psi(x) = i \quad if$$

$$\sum_{t=1}^{T_{max}} \omega(\Psi_t) F_t^{(i)}(x) = \frac{max}{j\epsilon\{1,,2,......J\}} \sum_{t=1}^{T_{max}} \omega(\Psi_t) F_t^{(i)}(x) \qquad (11)$$

This hybrid approach assigns weights to individual classifiers based on their predictive accuracy and computational cost. Classifiers whose weights fall below a predefined threshold are excluded during the initial stage. By using accuracy as the primary optimization criterion, the framework ensures that the hybrid model achieves improved overall performance.

4. RESULTS AND DISCUSSION

Matlab is used to construct the suggested facial expression recognition system. The LFW dataset, accessible at the given Kaggle link, is used to assess the suggested technique. LFW is an increasingly common set of face images intended to aid with unconstrained facial recognition research. The dataset was designed and researchers at the University of Massachusetts Amherst maintain; comprehensive references are included in the acknowledgments section. The Viola-Jones object identification framework was used to identify and align faces in 13,233 images of 5,749 individuals that were gathered from online sources. Among these, 1,680 individuals have at least two distinct images. The original dataset includes four variations of LFW images, along with three types of aligned versions. Experimental results indicate that the proposed Improved CNN–based face recognition model achieves superior performance compared to Adaptive Boosting (AB) and Random Forest (RF) in terms of F-measure, recall, accuracy, and precision.

Accuracy

The weighed percentage of facial expressions is correctly identified by the measurement accuracy. It is represented as,

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (23)$$

Where,

TP-True Positive

TN-True negative

FP-False Positive

FN-False Negative

Precision

The ratio of precisely anticipated beneficial results to all anticipated observations is defined as precision.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (24)$$

Recall

Compare successfully anticipated beneficial results to all first-class observations to compute recall.

$$\text{Recall}=\frac{TP}{TP+FN} \quad (25)$$

F-Measure

The F-measure is determined using the weighed mean of accuracy and recall.

$$F=2\frac{P\times R}{P+R} \quad (26)$$

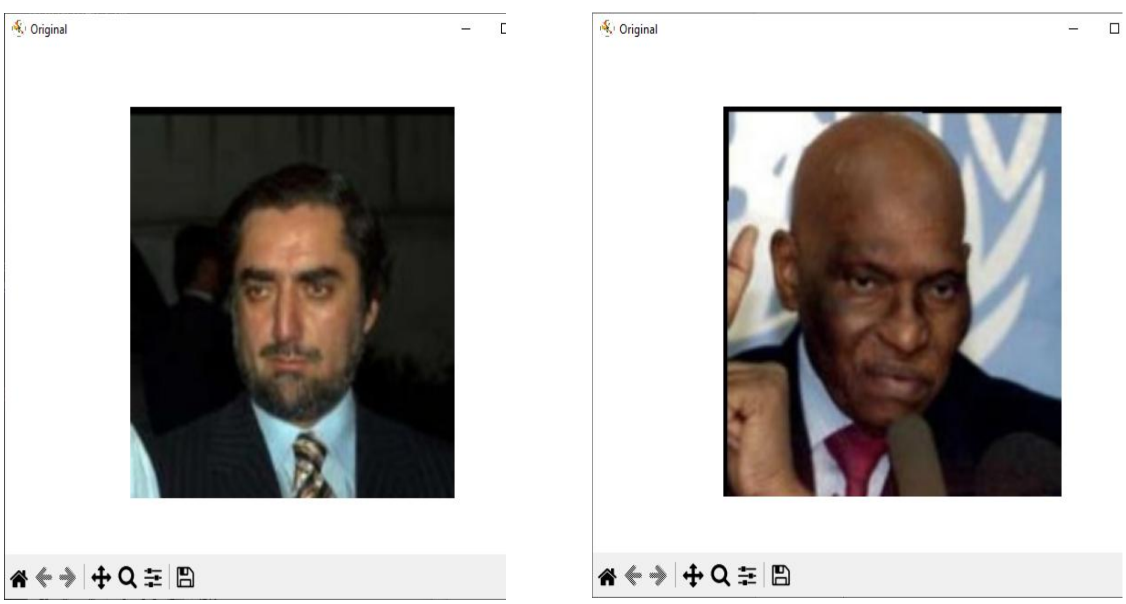
Dataset Selection:

The Fig 6 represents the sample images of the dataset.

Fig 6. Collection of input image from the dataset

Input Images:

Fig 7 shows the input images of all the sample images of the dataset.



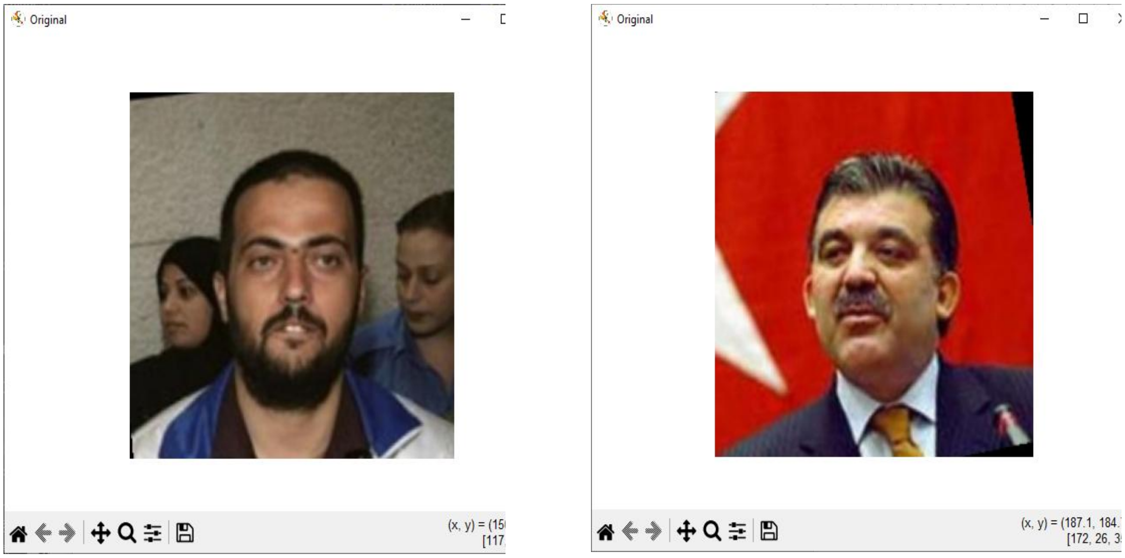
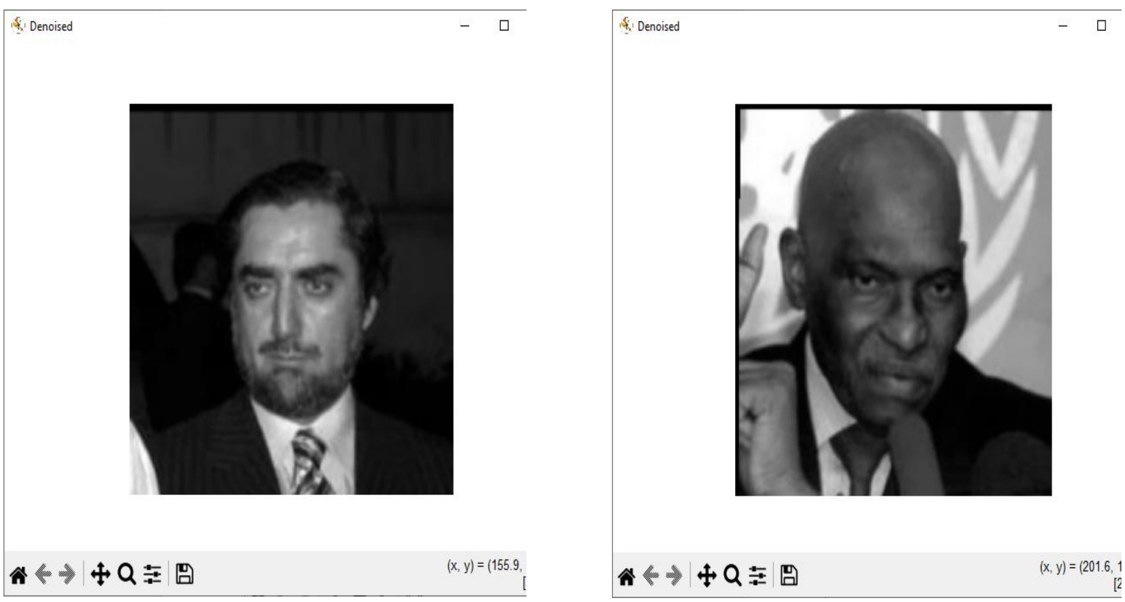


Fig 7 Input Images for all the sample dataset

Denoising Images:

Fig 8 represents the denoising images of all the sample datasets.



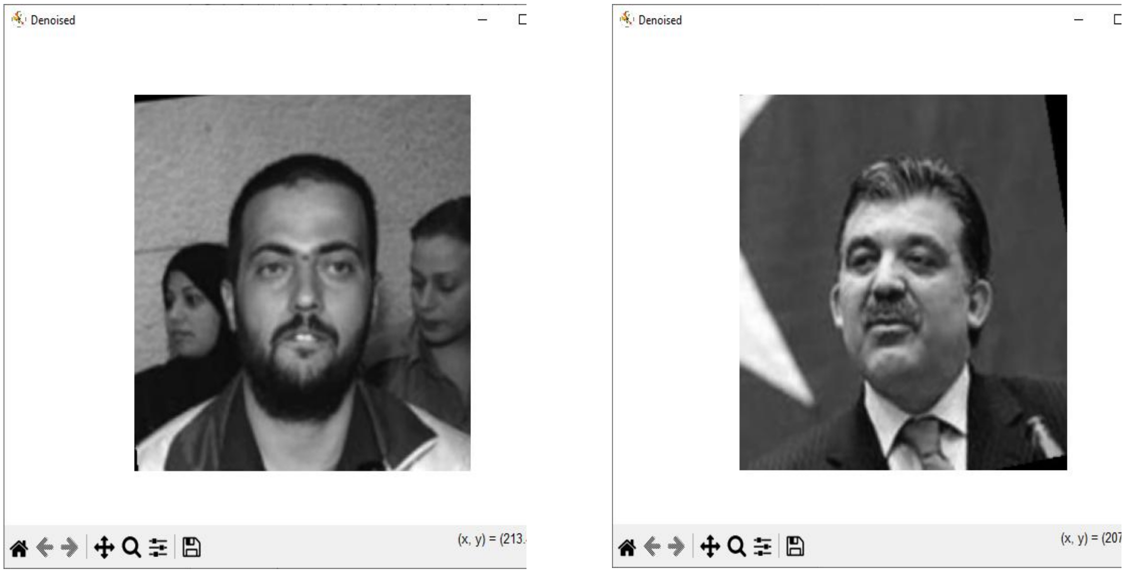
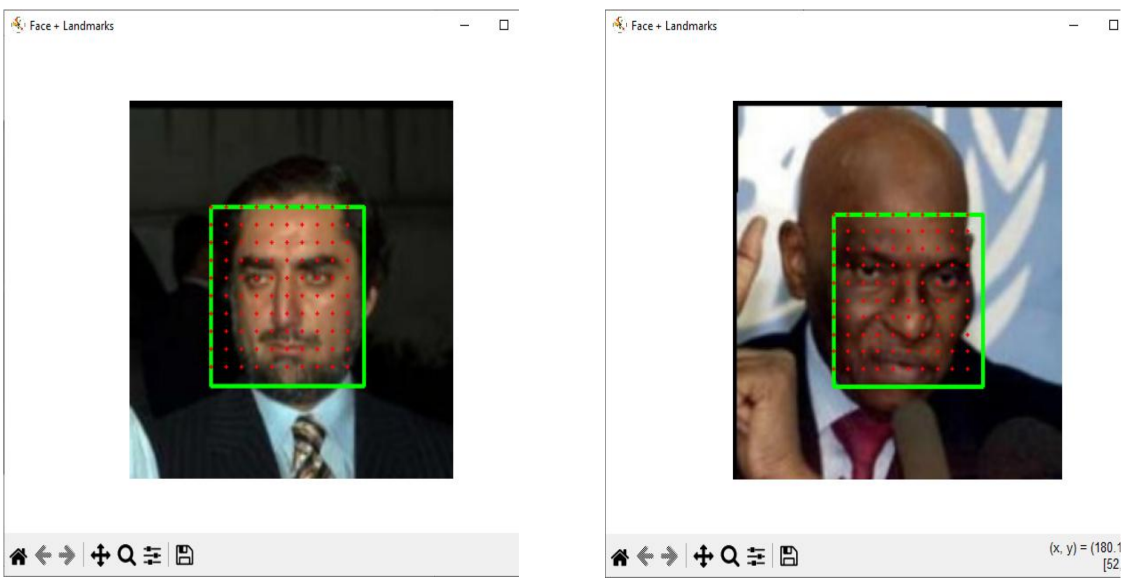


Fig 8 Denoising images for all the sample image dataset

Face + Landmark:

Figure 9 illustrated the face + landmarks of all the sample dataset.



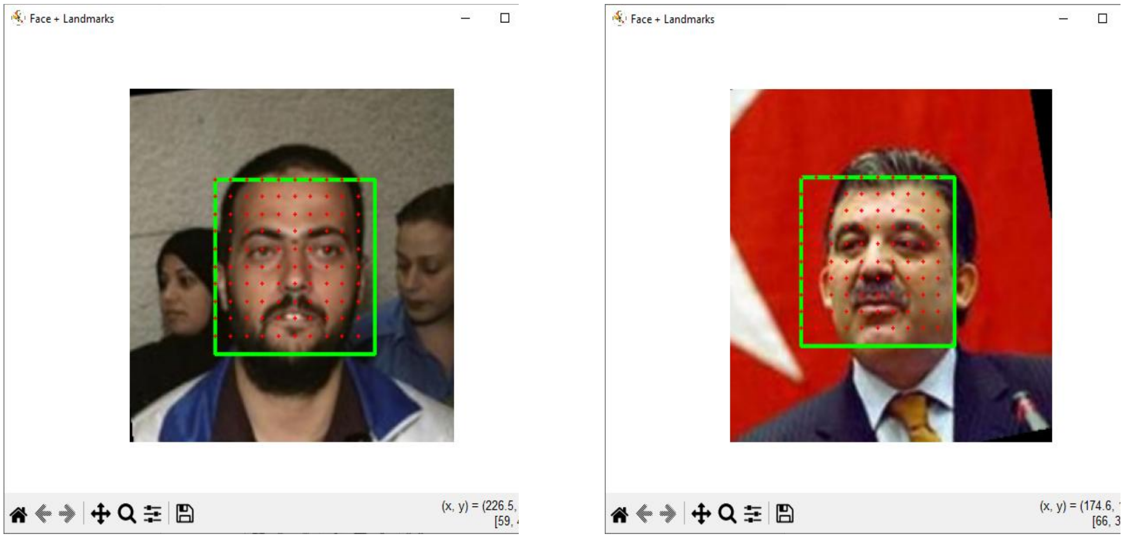
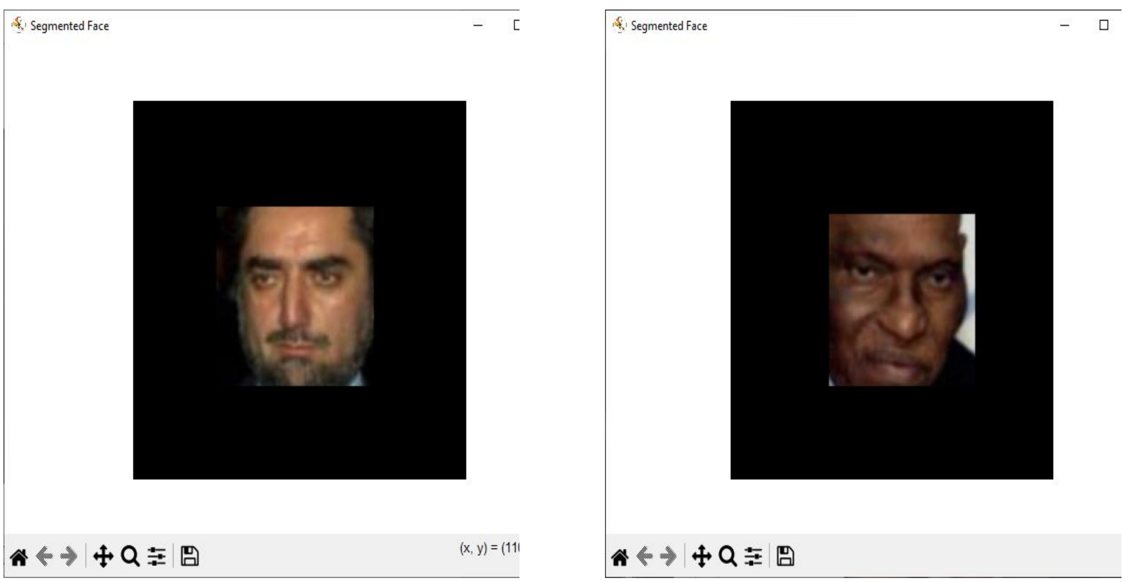


Fig 9 Face + Landmarks for all the sample image dataset

Segmented Images:

Fig. 10 shows the segmented images of all the sample datasets.



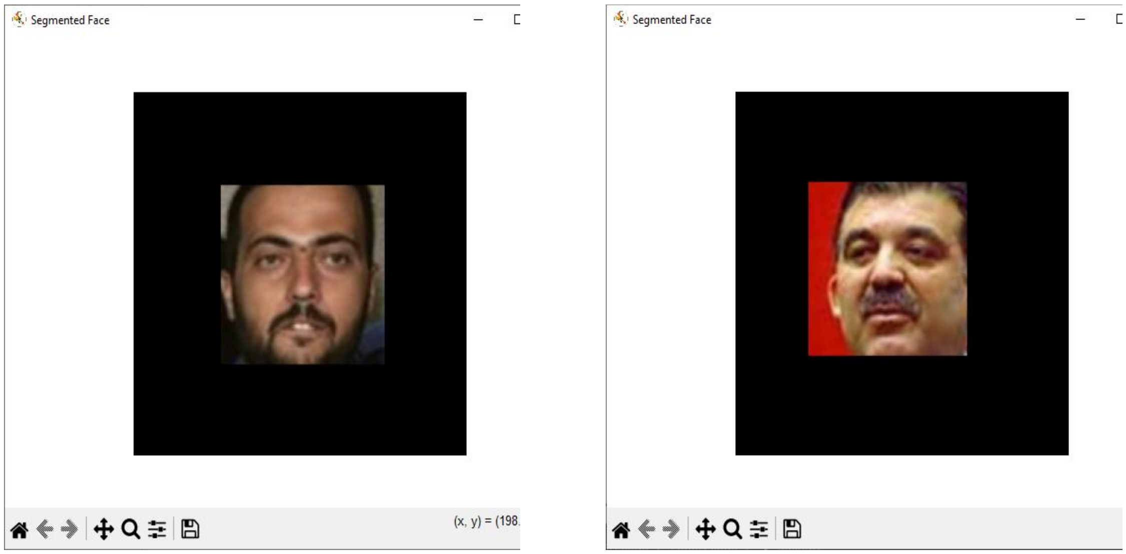


Fig 10 Segmented Image for all the sample image dataset

Identical Faces:

The identical faces are represented in Fig. 11 for all the sample datasets.

Fig 11. Identical Faces

Table 1. performance results between the proposed and existing methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
AB	85.12	81	84	83
RF	93	90.09	92.2	91.12
ICNN	94.54	95.7	93.3	94.49
FN-QCNN	97.78	98.1	97.45	97.7

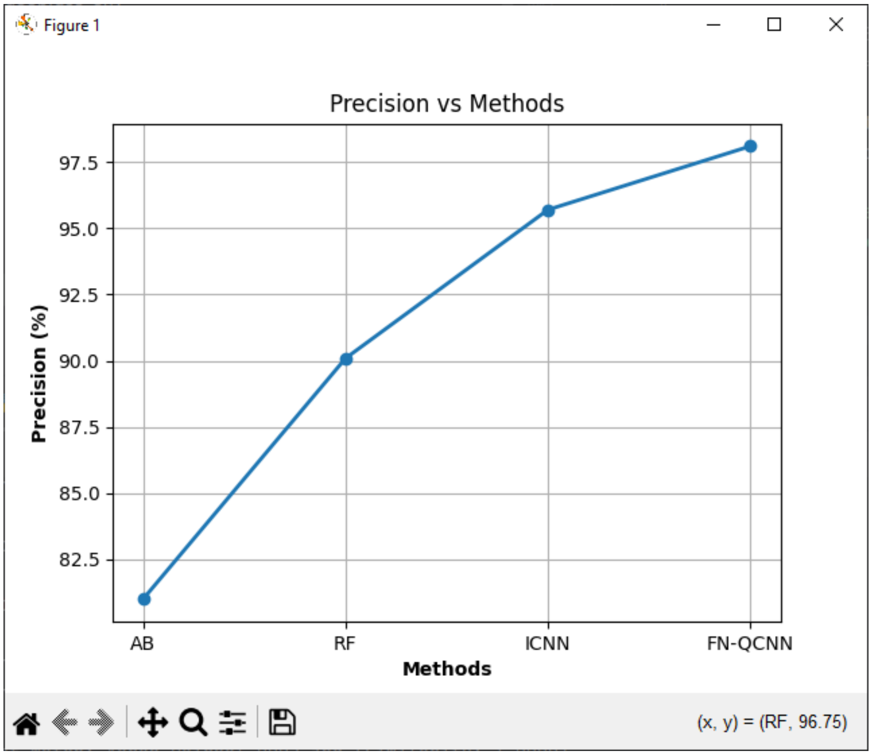


Fig.12. Precision comparison of the proposed FN-QCNN

The accuracy of the suggested method and existing facial recognition techniques is contrasted in Figure 12. The results demonstrate that the FN-QCNN model consistently achieves higher performance across the evaluated datasets compared to the selected machine learning algorithms. These findings align with the previously observed error rates and can be attributed to the non-redundant rule sets generated by the proposed classifier. From the results it concludes that the proposed FN-QCNN technique has high precision results compare to the existing classification techniques.

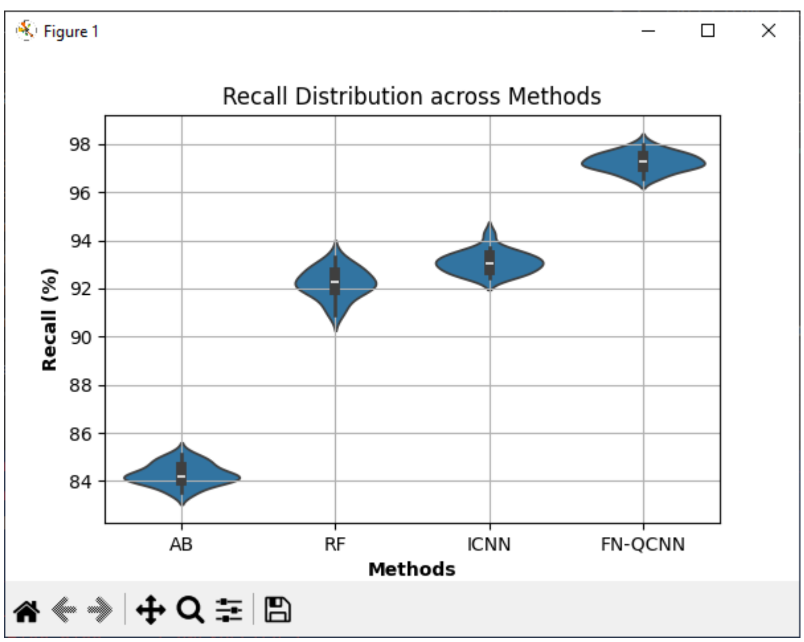


Fig.13. Comparison of the suggested FN-QCNN recall

The recall comparison between the suggested method and existing facial recognition techniques is shown in Figure 13. For every subject, a large number of images are used to train the FN-QCNN model, resulting in an expanded dataset that contributes to improved overall accuracy. By analysing the data, it is possible to determine that the light circumstances have an impact on the identification process. When there is minimal light, the recognition system often performs false recognition. This could be fixed by increasing the number of training images that were taken in low light while creating the face classifier. And also the parameter optimization using improved whale optimization increases the recall ratio.

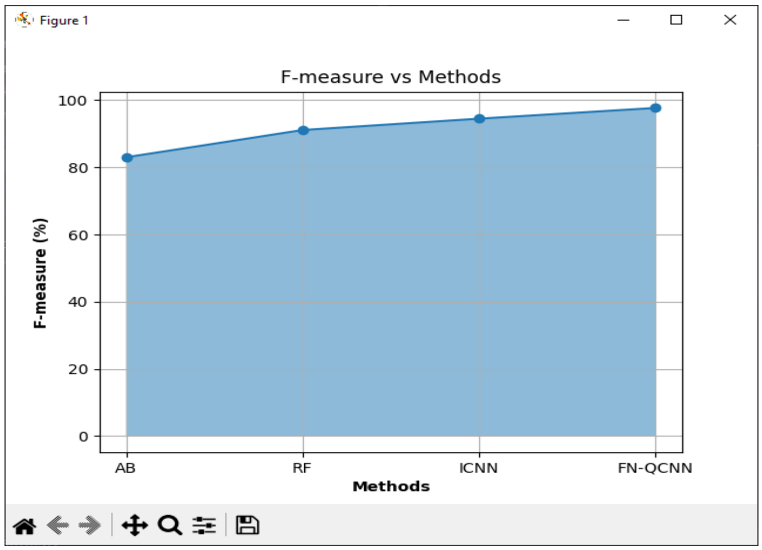


Fig.14. F-measure comparison of the proposed FN-QCNN

The F-measure comparison between the suggested method and existing facial recognition techniques is shown in Figure 14. The steps of categorization and feature extraction, according to the experiment data, effectively capture the most relevant variations in the LFW dataset, leading to a strong correlation with the target classes. The comparative analysis indicates that the proposed model achieves the highest accuracy among the evaluated machine learning approaches on the dataset. Additionally, the LFW dataset yields the best F-measure values, and the findings show that, in terms of F-measure, the FN-QCNN model regularly performs more effectively than existing methods.

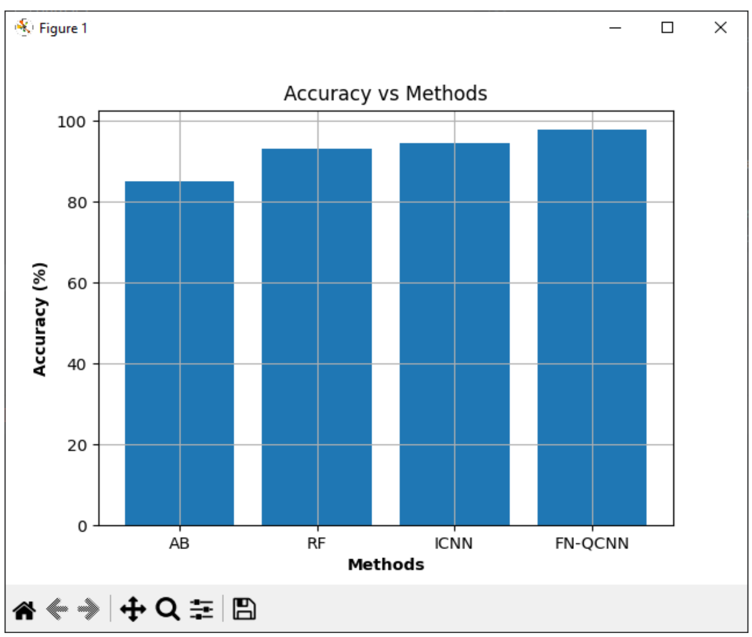


Fig.15. Comparison of the suggested FN-QCNN model's accuracy

The accuracy comparison between the suggested method and existing facial recognition techniques is shown in Figure 15. An effective deep learning model is characterized by its ability to make correct predictions while generalizing well to unseen data. Model performance is commonly evaluated using accuracy, which can be further interpreted through metrics include accuracy and sensitivities. From the results it concludes that the proposed FN-QCNN technique has high accuracy results compare to the existing classification techniques. It is observed that the suggested FN-QCNN technique provides the best accuracy when compared with other present techniques. Additionally, the total success rate of the facial recognition system increases by this preprocessing technique. The proposed FN-QCNN yields improved accuracy results of 97.78% where as ICNN technique yields is 94.54% RF technique yields is 93% and AB technique provides 85.12%.

5. CONCLUSION

In this research work, for identical face identification, a hybrid deep learning architecture that combines FaceNet and Quantum Convolutional Neural Networks was suggested. The suggested method's primary objective was to enhance the accuracy and reliability of face recognition systems when dealing with highly similar or identical facial images. The methodology first utilizes intelligent feature extraction model for face recognition that leverages two-dimensional (2D) facial images sourced from diverse origins to construct three-dimensional (3D) face meshes using 468 MediaPipe landmarks. The preprocessing pipeline incorporates Restormer to enhance image quality, followed by deep learning-based segmentation via Mask R-CNN architectures to effectively separate the foreground (face) from the background, thereby reducing noise and improving detection precision. Feature extraction are extracted by using the Vision Transformer (ViT). By mapping facial images into a structured embedding space, FaceNet enables efficient comparison between faces and highlights subtle differences something the human eye can find difficult to detect. Following feature extraction, the extracted embedding is processed by a Quantum Convolutional Neural Network after being encoded into quantum states. The QCNN model uses quantum operations including quantum convolution, pooling, and measurement to perform in sophisticated feature learning and classification. The ability of the system to identify complex patterns and correlations in face data is improved by this combination of deep learning and quantum computing technologies. The hybrid FaceNet–QCNN model demonstrates improved performance for identical face recognition by integrating the advantages of traditional deep learning and quantum machine learning. The suggested method presents a potential solution for challenging biometric identification tasks where traditional recognition systems may struggle to differentiate between highly similar faces..

References:

1. Faridi, M. S., Zia, M. A., Javed, Z., Mumtaz, I., & Ali, S. (2021). A comparative analysis using different machine learning: an efficient approach for measuring accuracy of face recognition. International Journal of Machine Learning and Computing, 11(2), 115-120.
2. Sharma, S., Bhatt, M., & Sharma, P. (2020, June). Face recognition system using machine learning algorithm. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1162-1168). IEEE.
3. Komlavi, A. A., Chaibou, K., & Naroua, H. (2024). Comparative study of machine learning algorithms for face recognition. Revue Africaine de Recherche En Informatique et Mathématiques Appliquées, 40.
4. Damale, R. C., & Pathak, B. V. (2018, June). Face recognition based attendance system using machine learning algorithms. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 414-419). IEEE.
5. Singhal, P., Srivastava, P. K., Tiwari, A. K., & Shukla, R. K. (2021, September). A Survey: Approaches to facial detection and recognition with machine learning techniques. In Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021 (pp. 103-125). Singapore: Springer Singapore.
6. Amirgaliyev, B., Mussabek, M., Rakhimzhanova, T., & Zhumadillayeva, A. (2025). A review of machine learning and deep learning methods for person detection, tracking and identification, and face recognition with applications. Sensors, 25(5), 1410.
7. Teoh, K. H., Ismail, R. C., Naziri, S. Z. M., Hussin, R., Isa, M. N. M., & Basir, M. S. S. M. (2021, February). Face recognition and identification using deep learning approach. In Journal of Physics: Conference Series (Vol. 1755, No. 1, p. 012006). IOP Publishing.
8. Manna, S., Ghildiyal, S., & Bhimani, K. (2020, June). Face recognition from video using deep learning. In 2020 5th International conference on communication and electronics systems (ICCES) (pp. 1101-1106). IEEE.
9. Khan, S., Ahmed, E., Javed, M. H., Shah, S. A., & Ali, S. U. (2019, July). Transfer learning of a neural network using deep learning to perform face recognition. In 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-5). IEEE.
10. Singhal, N., Ganganwar, V., Yadav, M., Chauhan, A., Jakhar, M., & Sharma, K. (2021). Comparative study of machine learning and deep learning algorithm for face recognition. Jordanian Journal of Computers and Information Technology, 7(3).

11. Setiowati, S., Franita, E. L., & Ardiyanto, I. (2017, October). A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods. In 2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 1-6). IEEE.
12. Pandey, I. R., Raj, M., Sah, K. K., Mathew, T., & Padmini, M. S. (2019). Face recognition using machine learning. Int. Res. J. Eng. Technol, 6, 3772-3776.
13. Nachet, R., & Stambouli, T. B. (2022, December). Improved face recognition rate using convolutional neural networks. In 2022 2nd International Conference on New Technologies of Information and Communication (NTIC) (pp. 1-5). IEEE.
14. Gupta, P., Saxena, N., Sharma, M., & Tripathi, J. (2018). Deep neural network for human face recognition. International Journal of Engineering and Manufacturing (IJEM), 8(1), 63-71.
15. Hussain, S. A., & Salim Abdallah Al Balushi, A. (2020, January). A real time face emotion classification and recognition using deep learning model. In Journal of physics: Conference series (Vol. 1432, No. 1, p. 012087). IOP Publishing.
16. Benradi, H., Chater, A., & Lasfar, A. (2023). A hybrid approach for face recognition using a convolutional neural network combined with feature extraction techniques. IAES International Journal of Artificial Intelligence, 12(2), 627-640.
17. Ghasemzadeh, A., & Demirel, H. (2018). 3D discrete wavelet transform-based feature extraction for hyperspectral face recognition. Iet Biometrics, 7(1), 49-55.
18. Vishwakarma, V. P., & Dalal, S. (2020). Generalized DCT and DWT hybridization based robust feature extraction for face recognition. Journal of Information and Optimization Sciences, 41(1), 61-72.
19. Bendjillali, R. I., Beladgham, M., & Merit, K. (2019, March). Face recognition based on DWT feature for CNN. In Proceedings of the 9th international conference on information systems and technologies (pp. 1-5).
20. Almabdy, S., & Elrefaei, L. (2019). Deep convolutional neural network-based approaches for face recognition. Applied Sciences, 9(20), 4397.
21. Udawant, P., Pratap, R., Gupta, S., Upadhyay, V., Sabale, K., & Thakkar, H. K. (2024, January). A systematic approach to face recognition using convolutional neural network. In 2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) (pp. 1-6). IEEE.
22. Tao, K., He, Y., & Chen, C. (2019, November). Design of face recognition system based on convolutional neural network. In 2019 Chinese Automation Congress (CAC) (pp. 5403-5406). IEEE.
23. Verma, Kesari, Bikesh Kumar Singh, and A. S. Thoke. "An enhancement in adaptive median filter for edge preservation." Procedia Computer Science 48 (2015): 29-36.
24. Ibrahim, Haidi, Nicholas Sia Pik Kong, and Theam Foo Ng. "Simple adaptive median filter for the removal of impulse noise from highly corrupted images." IEEE Transactions on Consumer Electronics 54.4 (2008): 1920-1927
25. Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M. H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5728-5739).
26. M. Zubair. (Jan. 6, 2022). Face Detection With Mediapipe| Towards Data Science. [Online]. Available: https://towardsdatascience.com/write-a-fewlines- of-code-and-detect-faces-draw-landmarks-from-complex-imagesmediapipe-932f07566d11
27. Dollár, K. H. G. G. P., & Girshick, R. (2017, October). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).
28. Dutta, P., Sathi, K. A., Hossain, M. A., & Dewan, M. A. A. (2023). Conv-ViT: a convolution and vision transformer-based hybrid feature extraction method for retinal disease detection. Journal of Imaging, 9(7), 140.
29. William, I., Rachmawanto, E. H., Santoso, H. A., & Sari, C. A. (2019, October). Face recognition using facenet (survey, performance test, and comparison). In 2019 fourth international conference on informatics and computing (ICIC) (pp. 1-6). IEEE.
30. Oh, S., Choi, J., & Kim, J. (2020, October). A tutorial on quantum convolutional neural networks (QCNN). In 2020 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 236-239). IEEE..