Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.071996

ARTICLE

A Quantum-Enhanced Biometric Fusion Network for Cybersecurity Using Face and Voice Recognition

Abrar M. Alajlan[1,*] **and Abdul Razaque**[2]

[1]Self-Development Skills Department, Commonrst Year Deanship, King Saud University, Riyadh, 11362, Saudi Arabia
[2]Department of Cybersecurity, International Information Technology University, Almaty, 050000, Kazakhstan

*Corresponding Author: Abrar M. Alajlan. Email: aalajlan1@ksu.edu.sa

Received: 17 August 2025; Accepted: 19 September 2025; Published: 30 October 2025

ABSTRACT: Biometric authentication provides a reliable, user-specific approach for identity verification, significantly enhancing access control and security against unauthorized intrusions in cybersecurity. Unimodal biometric systems that rely on either face or voice recognition encounter several challenges, including inconsistent data quality, environmental noise, and susceptibility to spoofing attacks. To address these limitations, this research introduces a robust multi-modal biometric recognition framework, namely Quantum-Enhanced Biometric Fusion Network. The proposed model strengthens security and boosts recognition accuracy through the fusion of facial and voice features. Furthermore, the model employs advanced pre-processing techniques to generate high-quality facial images and voice recordings, enabling more efficient face and voice recognition. Augmentation techniques are deployed to enhance model performance by enriching the training dataset with diverse and representative samples. The local features are extracted using advanced neural methods, while the voice features are extracted using a Pyramid-1D Wavelet Convolutional Bidirectional Network, which effectively captures speech dynamics. The Quantum Residual Network encodes facial features into quantum states, enabling powerful quantum-enhanced representations. These normalized feature sets are fused using an early fusion strategy that preserves complementary spatial-temporal characteristics. The experimental validation is conducted using a biometric audio and video dataset, with comprehensive evaluations including ablation and statistical analyses. The experimental analyses ensure that the proposed model attains superior performance, outperforming existing biometric methods with an average accuracy of 98.99%. The proposed model improves recognition robustness, making it an efficient multimodal solution for cybersecurity applications.

KEYWORDS: Biometric system; multi-modal biometric system; machine learning; deep learning; residual network; quantum networks

1 Introduction

The Internet is the fundamental factor that connects computer systems, facilitates global connectivity and data exchange. Cybersecurity is crucial for securing data, businesses, and individuals from cyber threats across all sectors [1]. It involves protecting data, servers, programs, and network infrastructures from unauthorized access by enabling cyber protective measures and policies. At its core, cybersecurity ensures that sensitive systems and data are accessible only to authorized users [2]. This highlights the requirement of an authentication step to ensure digital trust and security. Reliable user authentication ensures compliance with regulations, maintains user privacy, and safeguards sensitive data. Authentication is the foundational step in cybersecurity, ensuring that only authorized users can access systems and resources [3]. Unauthorized access due to unreliable authentication leads to productivity impacts, reputational damage, and financial

losses. In server communication, reliable authentication ensures that the client can access the data only after proper verification [4]. Username and password are the essential authentication elements used to verify a user's identity and grant access to protected resources. In this context, Biometric authentication, including voice, face, and fingerprint recognition, delivers improved security based on unique biological traits that are far more difficult to replicate compared to conventional passwords or tokens [5]. Biometric technologies offer a secure means of authentication and verification by utilizing human characteristics such as fingerprints, facial recognition, and voice or speaker identification. Also, it is a strong verification model that streamlines the current multi-factor verification model and replaces passwords [6].

A unimodal biometric authentication system functions by relying on a single biological or behavioral characteristic of the user, identifying and verifying individuals based solely on one biometric trait [7]. Although unimodal biometric authentication systems offer several advantages for identity recognition, they still have issues in terms of privacy, accuracy, scalability, and security [8]. The integration of both voice and face modalities enhances the accuracy and robustness against spoofing and adversarial attacks. Multimodal biometric systems are promising solutions for real-world applications, improving accuracy and security using multiple biometric traits [9]. Recently, the integration of Deep Learning (DL) has improved advanced biometric systems by enabling the extraction and fusion of high-quality features from voice and face [10].

1.1 Motivation

In an era where cyber threats are growing increasingly sophisticated, traditional authentication methods are proving insufficient to ensure secure and foolproof identity verification. These approaches are often vulnerable to environmental distortions, spoofing attacks, and adversarial manipulations. While DL has advanced feature extraction, the limitations in feature richness and model generalization persist. Moreover, classical systems struggle to efficiently represent the complex spatial-temporal correlations. To address these gaps, it is necessary to provide a more resilient, intelligent, and secure identity recognition framework. The motivation of this research lies in bridging this gap by harnessing the power of quantum computing to improve the representational depth and robustness of facial recognition. In addition to this, this model simultaneously leverages temporal modeling techniques to extract fine-grained speech dynamics. By fusing these modalities in a unified system, the proposed work aims to deliver a novel biometric authentication framework that is accurate, secure, scalable, and adaptable to real-world conditions.

1.2 Novelty

The novelty of the research work is outlined as follows:

Hybrid Quantum-Classical Model: The proposed model incorporates a Quantum Residual Network (QResNet) to enhance representational capacity and learning depth beyond traditional Convolutional Neural Networks (CNNs). It leverages quantum-enhanced processing by encoding classical facial features into quantum state representations.

Pyramid-Attentive Temporal Voice Modeling: The proposed work uses a Pyramid-1D Wavelet Convolutional Bidirectional Network (P-1DWCBiNet) that integrates the pyramid wavelet convolutions and bidirectional Gated Recurrent Units (BiGRU) for voice-based identity recognition. Thereby, the model effectively captures the fine-grained speech patterns and long-range dependencies important for robust voice-based identity recognition.

Biometric Fusion Using Early Concatenation: The biometric fusion strategy fuses the normalized voice and face embeddings. This strategy preserves complementary modality-specific features, resulting in a unified and highly discriminative identity representation.

Improved Cybersecurity* via *Dual-Biometric Verification: The proposed model significantly enhances resistance to spoofing and impersonation attacks by verifying both facial and vocal biometrics.

1.3 Contribution

The major contributions are stated below.

Dual-Modality Biometric Model for Cybersecurity: To improve identity verification in cybersecurity, the proposed QBioFusionNet model integrates both voice and face features. The developed mode is more effective compared to the unimodal systems by reducing spoofing susceptibility and enhancing reliability.

Improved Temporal Voice Modeling: The P-1DWCBiNet model that combines the multi-scale wavelet convolution and BiGRU structure enables the effective extraction of hierarchical temporal voice features that are important for reliable speech-based identity recognition.

Enhanced Facial Feature Representation: The QResNet model converts the classical facial image features into quantum state representations, improving the representation power and learning ability of the model for enhanced recognition accuracy.

Unified Identity Embedding: A unified representation is generated by using an early fusion mechanism that fuses the voice and face feature vectors. This strategy assists in capturing the complementary nature of both modalities, enhancing the robustness of the identity verification process.

Superiority Performance: The proposed model's performance is evaluated using standard evaluation parameters, demonstrating that it gains high recognition performance compared to other biometric methods.

The other parts of the research work are provided as follows: related works are discussed in Section 2, the problem statement and objective function are provided in Section 3, a detailed explanation of the proposed methodology is presented in Section 4, the comprehensive experimental analysis is conducted in Section 5, and the conclusion and future works are provided in Section 6.

2 Related Work

Abdelfatah et al. [11] presented a three-factor biometric quantum identity authentication approach that integrates passwords, client image, and voice to enhance cybersecurity. This approach was divided into two important phases, namely client enrollment and authentication. The first phase ensured that the client's voice signal was encrypted by quantum encryption with low layers incorporating quantum circuits. In the second phase, quantum secure direct communication was integrated with quantum voice encryption to provide enhanced, double-layer security. The overall experimental results demonstrated that the model gained better performance rates.

Al-Abboodi et al. [12] developed a DL-based face recognition system within an IoT environment. Spatial-Temporal Interest Point (STIP) was deployed for the feature extraction, focusing on capturing features related to facial activities. A statistical method has selected the important features that contribute to facial recognition. Facial detection was performed using a CNN, and its parameters were tuned by Galactic Swarm Optimization (GSO). Experimental analyses ensured that the model attained high effectiveness on the face image dataset.

He et al. [13] presented an Audio-Visual Speech Recognition using Multimodal Generative Adversarial Network (AVSR-GAN). The AVSR-GAN model incorporates two-stream networks designed to process visual and audio stream information, where the generator and discriminator work together to improve

performance. To ensure privacy and security, this work incorporated an authentication model with federated learning approach (FL). Evaluations on Lip Reading Sentence 2 and 3 datasets confirmed the model's robustness.

Suneetha and Anitha [14] developed a speech-based emotion recognition model using an Improved and Faster Region-based CNN (IFR-CNN). The IFR-CNN approach stood out by learning and storing affective states while monitoring and recovering speech properties. For effective speech-based emotion recognition, spectral characteristics and temporal features were extracted. The IFR-CNN model's efficiency was better on the speech datasets.

Talaat [15] developed a lie detection model by using an explainable enhanced recurrent neural network (ERNN) with a fuzzy logic system. This work has employed a Long Short-Term Memory architecture (LSTM) trained on the audio recordings dataset. The ERNN model demonstrated high performance with reduced training time, highlighting its efficiency and effectiveness.

Gokulakrishnan et al. [16] presented an optimized facial recognition approach that detects criminal activities by applying DL. This work employed a Strawberry-CNN (SbCNN), which demonstrated effective face verification and criminal face detection with an optimization fitness function. The experimental results on the face expression database confirmed that the SbCNN model delivered efficient performance.

Alsafyani et al. [17] developed an effective model integrating DL and picture optimization with cryptography. A 5D conservative chaotic model with an optical chaotic Map (OChaMap) was deployed to enhance the security of the key. Encrypted facial images by applying Crypto General Adversarial Neural Network with OChaMap were employed to extract the involved items. The experimental analyses showed that the Cry_GANN_OChaMap model reached better effectiveness rates.

Lin et al. [18] developed an optimized face recognition access control system with Principal Component Analysis (PCA) to ensure effective facial recognition. Facial recognition was carried out by using an aggregating spatial embedding algorithm. Beta prior and full probability Bayesian method were integrated to optimize the PCA model. Moreover, the efficiency of the face recognition was improved by integrating the K-means Clustering Algorithm. The overall experimental outcomes demonstrated that the system attained greater performance while maintaining lower training time.

Ragab et al. [19] presented a Hunter-Prey Optimizer with DL-enabled Biometric Verification to enhance cybersecurity (HPODL-BVCS). The biometric images were gathered from the biometric dataset and then pre-processed utilizing Bilateral Filtering (BF). The critical features were extracted by applying the ShuffleNetv2.3 model. Here, a convolutional autoencoder (CAE) model offered an effective classification. The results analysis demonstrated that the model achieved superior accuracy, contributing to enhanced cybersecurity.

Aly et al. [20] developed a DL-based approach that enhances facial expression recognition in the online learning context. This facial expression recognition system has involved refinement of the ResNet Deep CNN, leading to improved precision in expression recognition. The model ensured effective classification using a convolutional block attention mechanism (CBAM) that attends to important spatial and channel features. This demonstrated that the model delivered higher performance and lower execution time.

For the human-computer interface, Alrowais et al. [21] developed a Modified Earthworm Optimization with DL Assisted Emotion Recognition (MEWODL-ER). The feature extraction step generated important feature vectors using the GoogleNet architecture. In the classification phase, the Quantum Autoencoder (QAE) model effectively classified the emotions associated with human–computer interface applications. The model delivered high recognition performance in experimental evaluations.

Gona et al. [22] developed a Transfer Learning Convolutional Neural Network (TL-CNN) for the biometric system. The gathered images from the different biometric datasets were improved, and the pattern-specific features were extracted using a Convolutional Residual Network (CRN). This work has demonstrated that the TL-CNN attained greater performance across multiple biometric datasets.

Sun et al. [23] evaluated the anti-spoofing performance of the Vector Tracking Loop (VTL) in complex environments, such as high-interference or signal degradation scenarios. Their findings demonstrated that the VTL method maintained robust tracking and spoof detection accuracy.

Radad et al. [24] developed a novel encoder-based CNN architecture that incorporated grayscale structural information, which enhanced the detection of spoofing attacks. This approach leveraged texture analysis and achieved high classification performance by distinguishing real and fake face inputs.

Ma et al. [25] proposed a face anti-spoofing algorithm based on pseudo-negative feature generation. This method generated artificial negative samples to enhance model generalization and reduce the risk of overfitting. From experimental computation, this model significantly improved detection performance on unseen spoofing patterns.

Limitations and Research Gaps

Although existing methods have achieved significant advancements in identity identification, they still encounter some limitations that affect their overall performance. The limitations need to be resolved to enable more accurate and secure voice and face recognition. The following points represent the key limitations of the existing methods.

Vulnerability to Spoofing and Impersonation: Existing unimodal biometric systems are vulnerable to spoofing attacks utilizing videos, photos, or synthetic voices. The proposed model utilizes complementary information from both spatial and quantum facial features and temporal voice dynamics. This dual-modality integration improves resistance to impersonation and spoofing attempts.

Limited Representation in Unimodal Systems: Existing unimodal systems that rely on either voice or text lack sufficient discriminatory power to ensure robust identity recognition. The proposed QbioFusionNet combined both modalities (voice and face) using a novel multimodal fusion mechanism. This demonstrates that the proposed system remains accurate and reliable for identity recognition.

Redundant Feature Extraction: The traditional biometric methods utilize handcrafted features or CNNs, ensuring that the high-level discriminative quantum information has not been captured. In the proposed system, classical images are encoded into quantum states using QResNet, providing an enriched feature representation. Meanwhile, P-1DWCBiNet effectively extracts the deep temporal features from voice data to improve the overall representation quality.

3 Problem Statement

In critical applications, such as financial transactions, surveillance, and secure access control, highly reliable user identity verification is essential. However, traditional unimodal biometric systems that utilize a single trait encounter challenges due to inherent limitations:

- Face recognition is vulnerable to changes in lighting, pose, and spoofing attacks utilizing photos or videos.
- Voice recognition has poor performance due to mimicry, throat infections, and noise.

Traditional biometric systems suffer from inter-class similarity and intra-class variability, resulting in High False Acceptance (HFA) and High False Rejection Rates (FRR). To address these limitations, this work develops a DL-based effective multimodal biometric authentication that combines the face and voice

modalities. The primary objective is to develop an effective system capable of learning discriminative, complementary, and invariant features from both modalities.

Objective Function

The objective of the proposed system is to minimize total authentication error by learning feature representations and using a fusion mechanism. The fusion mechanism increases the similarity between genuine user pairs and maximizes the dissimilarity for imposter pairs, thereby improving verification accuracy. Let's consider be the feature extractor for voice modality with parameters and as a feature extractor for face modality with parameters θ_f. $F = \phi\left(f_v, f_f; \theta_f vg\right)$ is the fusion function with fusion parameters θ_{fvg}. $D\left(F_1, F_2\right)$ specifies the distance function and $x \in \{0,1\}$ indicates the binary label. To efficiently train the proposed multimodal biometric authentication framework, the learning objective is formulated as the minimization of the contrastive loss function $L_{contrastive}$:

$$O\left(m \cdot n\right) \tag{1}$$

In the above equation, n indicates the margin that separates genuine and imposter pairs. This encourages small distances for genuine pairs and large distances.

4 Proposed Methodology

Fig. 1 illustrates the steps of the proposed model for identity recognition. In the data collection step, the VidTIMIT dataset is employed, offering raw video frames and audio recordings that serve as input for further processing. The raw video frames were pre-processed by performing face detection and cropping to isolate the facial region from the image background. Subsequently, the cropped face images underwent resizing, normalization, and gray-scale conversion. The local Binary Pattern (LBP) ensures that the local texture information is extracted effectively. For the audio recordings, voice activity detection is deployed to identify and segment speech regions. The noise reduction step removes the unwanted background noise from the audio recordings. The framing and windowing step divides the audio signal into short frames and minimizes edge effects. Log Compression and Mel-Frequency Cepstral Coefficients Extraction (MFCC) ensure effective audio extraction. To improve model robustness, augmentation steps are employed for both modalities, offering diverse variations of the original data. The proposed model performs feature extraction using P-1DWCBiNet for voice and QResNet for facial images. The extracted features were fused to create a unified and discriminative biometric signature, enhancing identity verification accuracy. The proposed model represents a cutting-edge approach that strengthens authentication reliability and resilience against spoofing and cyber threats.

4.1 Data Collection

Data collection is the fundamental step that captures the required data using various data sources to train the proposed system. In the proposed biometric system, the VidTIMIT audio-video dataset is deployed to gather the data. The details of the VidTIMIT audio-video are as follows:

VidTIMIT audio-video Dataset: The VidTIMIT dataset encompasses video and audio recordings from 43 individuals, each representing a set of short sentences (https://conradsanderson.id.au/vidtimit/, accessed on 18 September 2025). In this dataset, the digital video camera is employed for the audio and video recordings. This dataset is recorded in 3 sessions, including session 1, session 2, and session 3. Here, 10 sentences are assigned to each person. There are six sentences for Session 1, two sentences for Session 2, and the last two sentences for Session 3. The video data is presented as a sequentially numbered series of JPEG images, each with 512 × 384 pixels.

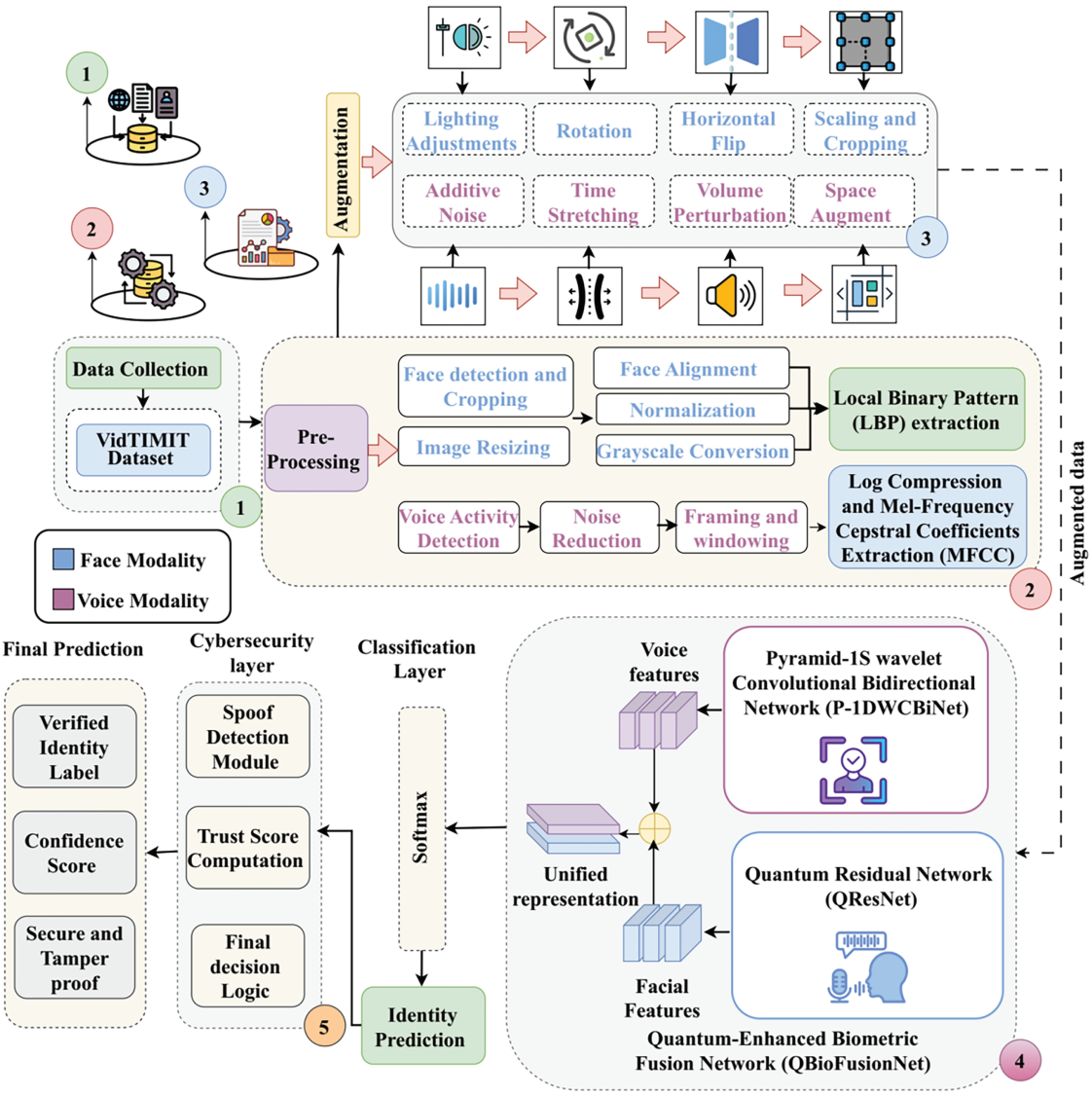


Figure 1: Proposed QbioFusionNet model

4.2 Pre-Processing

Pre-processing the collected data is a prominent step that significantly determines the overall efficiency and performance of the proposed system. In the proposed work, distinct pre-processing steps are employed for both audio and face modalities.

4.2.1 Face Modality Pre-Processing

Face Detection and Cropping: In the pre-processing stage, the detection of the face in each video frame is the primary task, ensuring accurate localization of the facial regions. Let's consider the actual image frame as $I = \mathbb{R}^{H\times W\times 3}$, where the third dimension is related to the Red, Green, and Blue (RGB) channels, the image's height and width are indicated as H and W. This step generates a bounding box $B = (y, x, wt, hg)$, where represents the width of the detected face region, indicates the height of the detected face region, and specifies the top-left corner. Following this, the face image is cropped as, $I_{Crop} = I[x : x + hg, y : y + wt]$ and it considers only the face area for the following steps.

Face Alignment: Face alignment is the next step following face detection. It standardizes orientation while minimizing pose-related discrepancies. The facial features, such as the nose tip and eye centers, are detected. These features are indicated as a set of points $P = \left\{\left(y_j, x_j\right)\right\}_{j=1}^{m}$. In this step, the affine transformation matrix $M \in \mathbb{R}^{2\times3}$ is utilized to align these features to a canonical configuration $P' = \left\{\left(y'_j, x'_j\right)\right\}_{j=1}^{m}$. This matrix reduces the distance between P and P' and this transformation generates the aligned face image.

$$I_{aligned} = M\left(I_{crop}\right) \tag{2}$$

Image Resizing: Resizing the aligned face images was essential to ensure consistent input for the proposed model. The aligned face images are resized to 224 × 224.

$$I_{resized} = Resize\left(I_{aligned,}224, 224\right) \tag{3}$$

Normalization: The image normalization step is employed to reduce variations and pixel intensity. Therefore, each pixel value of the images is scaled to the range of $[0,1]$ utilizing the following equation:

$$I_{norm} = \frac{I_{resized}}{255} \tag{4}$$

Grayscale Conversion: Grayscale conversion is an important step in face detection, used for transforming the original images to grayscale, making it effective for the proposed model to identify facial features. The proposed system minimizes the dimensionality of input data and preserves important details that are essential for face recognition by removing color information. The grayscale images are utilized as a uniform input format for the following step, demonstrating that the proposed model concentrates on intensity-based patterns.

Local Binary Pattern Extraction: An LBP that captures local texture information from the grayscale images is employed to extract the features [23]. The face image is specified as $I\left(y, x\right)\mathbb{R}^{H\times W}$. A 3×3 neighborhood is assumed for every (y, x). The following equation provides the intensity of the center pixel $j_d = I\left(y, x\right)$ with its 8 neighboring pixels.

$$S\left(j_m - j_d\right) = \begin{cases} 1 & if\ j_m \geq j_d \\ 0 & if\ j_m < j_d \end{cases} \tag{5}$$

The following equation calculates the LBP for each pixel, generating an LBP code for the pixel.

$$LBP_{y,x} = \sum_{m=0}^{7} S\left(j_m - j_d\right) \cdot 2^m \tag{6}$$

Following the calculation of LBP, the LBP-coded image is acquired from the entire face image with its calculated LBP values. This involves dividing the LBP-coded image into small non-overlapping regions. Then, LBP values are calculated for each region. Consider $R_j\left(l\right)$ as the frequency of the LBP code in block. The final face feature vector is generated by integrating each histogram.

$$r_{face} = \left[R_1, R_2, \ldots, R_M\right] \in \mathbb{R}^c \tag{7}$$

In the above equation, r_{face} indicates the texture-based identity features of the face, c indicates the total number of histogram bins, and M specifies the number of blocks.

4.2.2 Voice Modality Pre-Processing

Voice Activity Detection: Raw audio recordings from the datasets cannot be provided to the proposed system as they contain unwanted silence and background noise, which capture all sounds present in the environment. Consider raw waveform be $y(\mathrm{e})$, e represents the time index. A Voice Activity Detection (VAD) algorithm outputs a binary mask $n(e) \in \{0,1\}$ that detects the speech regions. The following equation represents the filtered signal that retains only speech.

$$y_{vad} = (e) = y(e).n(e) \tag{8}$$

Noise Reduction: Background noise is removed to ensure the clarity of the voice recordings. The following equation represents the clean speech signal for the estimated noise signal $\widehat{m}(e)$.

Framing and Windowing: From the clean audio, overlapping frames with length L and stride S are multiplied by the Hamming window $w[m]$, reducing discontinuities at the edges.

$$y_w[m] = y[m].w[m] \tag{9}$$

Log Compression and Mel-Frequency Cepstral Coefficients Extraction: Logarithmic compression is used for Mel energies. The log-Mel energies are decorrelated, and cepstral coefficients are extracted by applying the Discrete Cosine Transform (DCT). This generates MFCCs, providing a matrix $D \in \mathbb{R}^{A \times c}$, where c indicates the number of coefficients for each frame and A represents the number of frames.

$$S_{\log}[n] = \log(S[n]) \tag{10}$$

$$d[m] = \sum_{n=0}^{N-1} S_{\log}[n] \cos\left[\frac{\pi m}{N}(n+0.5)\right] \tag{11}$$

4.3 Data Augmentation via Lighting and Rotation

The data augmentation step is initially performed on the original data to augment the size and diversity of the datasets using a variety of transformations. The augmentation steps enhance the robustness and generalization ability of the biometric authentication approach. In this work, suitable data augmentation steps are applied to both modalities (face and voice). For face images, augmentation steps, such as lighting, random rotations, flipping, scaling, and cropping are deployed. These allow the model to recognize the same individual under different visual conditions, simulating real-world variations, such as viewpoint shifts, head pose differences, and illumination changes.

Lighting Adjustment: This step includes the modification and contrast of an image. The following equation shows how each pixel's intensity is modified. In the equation below, $I(y,x)$ is the actual image, β denotes the brightness offset, and α indicates the contrast factor.

$$I_{aug}(y,x) = \alpha.I(y,x) + \beta \tag{12}$$

Rotation: The rotation process rotates the image around a center point. In the following equation, (y',x') denotes the rotated pixel coordinates, (y_d,x_d) represents the image center, and θ is the rotation angle.

$$\begin{bmatrix} y' \\ x' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} y - y_d \\ x - x_d \end{bmatrix} + \begin{bmatrix} y_d \\ x_d \end{bmatrix} \tag{13}$$

Horizontal Flip: This transformation mirrors the image with a vertical axis. The below equation computes the new horizontal position of all pixels, flipping the image horizontally.

$$I_{flip}(y,x) = I(W-1-y,x) \tag{14}$$

Scaling and Cropping: This step is employed to resize and crop an image. In Eq. (15), s is the scaling factor.

$$I_{scale} = Crop(Resize(J, sH, sW)) \tag{15}$$

The augmentation techniques are applied to the audio signals, which include noise to mimic background sounds, time stretching to vary speaking speed, and pitch shifting to alter vocal tone, volume perturbation, and spectrogram masking (SpecAugment). The utilization of these steps assists the system in maintaining variability in speaker characteristics and acoustic environments. The augmentation steps are detailed as follows:

Additive Noise: This step adds random noise to the actual audio signal. The following equation indicates the augmented audio, where $y(t)$ represents the original audio, $m(g)$ denotes the noise, and γ.

$$y_{aug}(g) = y(g) + \gamma.m(g) \tag{16}$$

Time Stretching: This technique changes the speed of the audio without changing the pitch. In the below equation, α is the time-scale factor:

$$y_{aug}(g) = y\left(\frac{1}{\alpha}\right) \tag{17}$$

Volume Perturbation: This step is employed to change the pitch of the audio with no speed alteration. From Eq. (18), δ is the amplitude gain factor.

$$y_{aug}(g) = \delta.y(g) \tag{18}$$

SpaceAugment: This technique utilizes masks for the spectrogram of the audio. The following equation represents the frequency mask and time mask on the spectrogram O.

$$f_0 : f_0 + f_w \tag{19}$$

$$e_0 = e_0 + e_m \tag{20}$$

In the above equations, $O \in \mathbb{R}^{F \times T}$ is the spectrogram, f_w, e_w denote the mask widths, f_0 and e_0 are the random start. With the utilization of these mechanisms, the proposed model learns the invariant and discriminative features from the two modalities. This reduces the likelihood of the overfitting issue and improves the accuracy and robustness in real-world scenarios.

4.4 Quantum-Enhanced Biometric Fusion Network

QBioFusionNet is a multimodal biometric recognition approach combining DL with quantum-enhanced modeling to ensure robust and effective human identification. It combines the information from both sources voice signals and facial images. Voice signals are processed via P-1DWCBiNet, and facial images are processed via QResNet. These two branches are designed according to their data modality and utilize temporal modeling and attention mechanisms for voice features, and Quantum state encoding and entangled

feature learning for facial features. Following independent feature extraction, a fusion module integrates the learned embeddings into a shared representation for identity classification. The QBioFusionNet improves the generalization and performance in real-world biometric applications.

4.4.1 Voice Recognition with Pyramid-Aware 1D Wide Convolutional Bidirectional Network

The P-1DWCBiNet model is the combination of the Pyramid Attention Mechanism (PAM) and a deep neural architecture, which improves speaker recognition performance utilizing voice signals. The deep neural architecture incorporates kernel convolutions, BiGRU, and a global average pooling (GAP) classifier. The PAM allows the proposed model to pay attention to salient temporal segments across multiple resolution scales [26]. The layer architecture of P-1DWCBiNet is illustrated in Fig. 2.

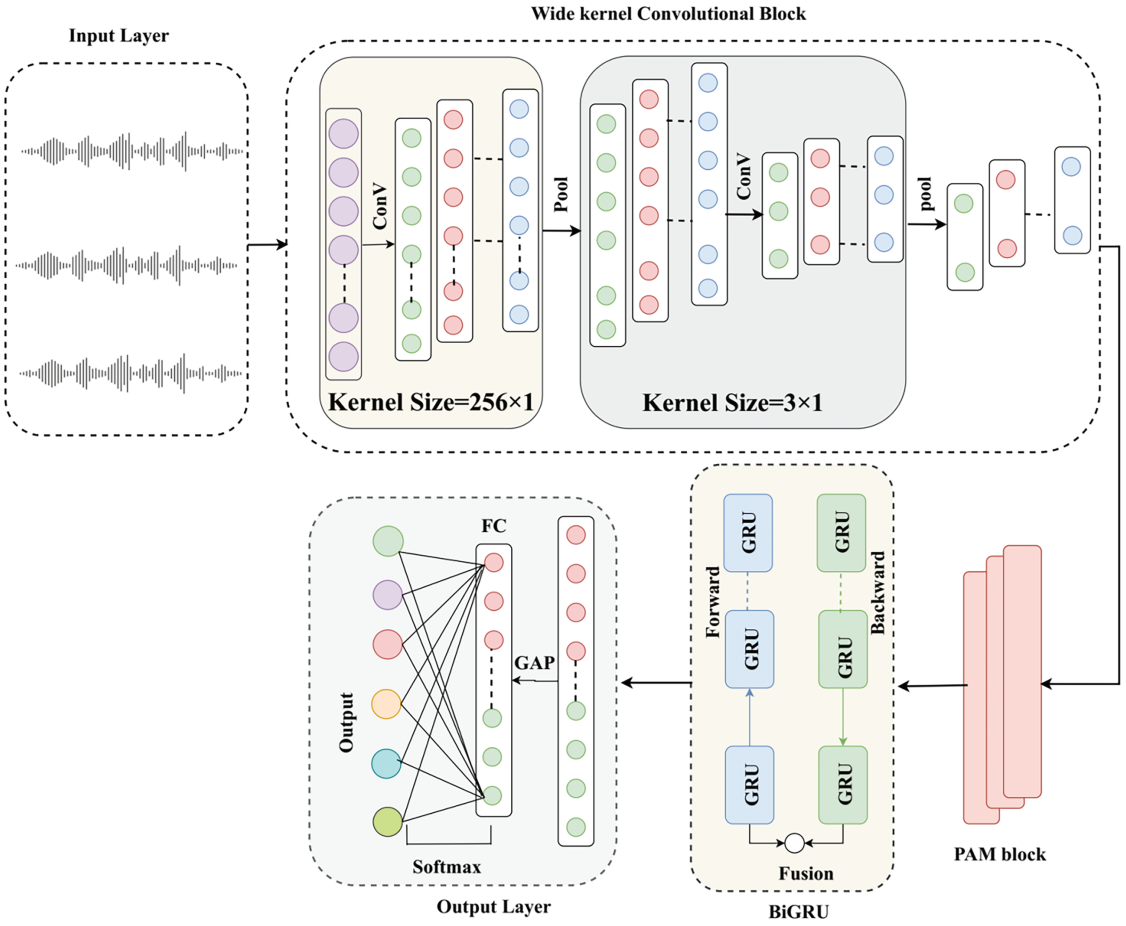


Figure 2: P-1DWCBiNet model

Input Representation

The preprocessed audio signals are transformed into a standardized 1D representation. Consider $y(e) \in \mathbb{R}^T$ as the time-domain speech waveform of length T. This signal is directly provided to the one-dimensional convolutional bidirectional gated recurrent unit (1D-WCBGRU) model without the MFCC feature extraction, leading to end-to-end learning.

Wide Kernel Convolutional Block

A wide 1D convolutional kernel is employed to convolve the input waveform for the local feature extraction from the speech. The large kernel assists in capturing low-frequency components and long-range temporal dependencies.

$$h_1(e) = \mathrm{Re}LU\left(WE_1 * y(e) + b_1\right), \quad WE_1 \in \mathbb{R}^{256} \tag{21}$$

From the above equation, b_1 is the bias term, WE_1 indicates the wide convolutional filter of size 256, and $*$ represents the 1D convolution operation. The addition of the max pooling layer minimizes the dimensionality and improves translational invariance.

$$h_1^{Pool} = \max Pool\left(h_1(e)\right) \tag{22}$$

High-resolution features are refined by using a second convolutional layer that has a smaller kernel size.

$$h_2(e) = \mathrm{Re}LU\left(WE_2 * h_1^{pool} + b_2\right), \quad WE_2 \in \mathbb{R}^2 \tag{23}$$

The layer normalization is utilized in order to stabilize and normalize intermediate activations.

$$\hat{h}_2(e) = LayerNorm\left(h_2(e)\right) \tag{24}$$

Pyramid Attention Mechanism

The PAM is introduced to improve the model's capability to attend to salient speech features across different temporal scales. It models the contextual information at multiple temporal resolutions, improving feature representation to enable advanced voice recognition. The output of the convolutional block is denoted as $F \in \mathbb{R}^{T\times d}$, where d represents the feature dimensionality and T denotes the temporal length of the feature sequence. PAM initially aggregates the local context using average pooling at multiple temporal scales $s \in \{1, 2, 4\}$. A down-sampled representation is acquired for each scale s.

$$F_s = AvgPool_s(\mathrm{F}) \tag{25}$$

The pooled features F_s are up-sampled back to the original length T, ensuring compatibility with the original temporal resolution.

$$\tilde{F}_s = Upsample_s\left(F_s\right) \tag{26}$$

After multi-scale context extraction, PAM utilizes a linear transformation to calculate the attention weights for every scale.

Bidirectional Gated Recurrent Units

BIGRU is deployed to model long-term dependencies in speech, especially in speaker traits that manifest over time. The forward and backward GRU operations with normalized convolutional output are presented in the following equation.

$$\vec{h}_e = GRU_{fwd}\left(\hat{h}_2(e)\right) \quad \vec{h}_e = GRU_{bwd}\left(\hat{h}_2(t)\right) \tag{27}$$

The following equation indicates the combined hidden representations:

$$h_e^{BiGRU} = \left[\vec{h}_e; \overleftarrow{h}_e\right] \tag{28}$$

Global Average Pooling

A GAP layer is employed to map the sequence of hidden states to a fixed-length vector. The addition of the GAP demonstrates that the model has the ability to accept variable-length inputs and generate consistent speaker embeddings.

$$f_{voice} = \frac{1}{T}\sum_{e=1}^{T} h_e^{BiGRU} \tag{29}$$

Output Layer and Classification

The output layer enables the fully connected layer (FC) to process a fixed-length vector $f_{voice} \in \mathbb{R}^c$ for effective speaker identity classification.

$$\hat{x} = Soft\max\left(W_{fc}.f_{voice} + b_{fc}\right) \tag{30}$$

In the above equation, $W_{fc} \in \mathbb{R}^{C\times c}$, C represents the number of speaker classes and $\hat{x} \in \mathbb{R}^C$ represents the predicted probability distribution over classes.

4.4.2 Quantum Residual Network for Face Recognition

The QResNet initially encodes the facial features that are extracted from the input images into quantum state representation. Fig. 3 illustrates the QResNet-Face model. Assume $I \in \mathbb{R}^{H\times W\times C}$, where the number of color channels is denoted as C. Local features $F \in \mathbb{R}^{S\times d}$ are extracted by Multilayer Backpropagation (MBP), where S represents the spatial feature locations and d denotes the dimension of the local feature vector, therefore $N = S \times d$, representing a total number of features.

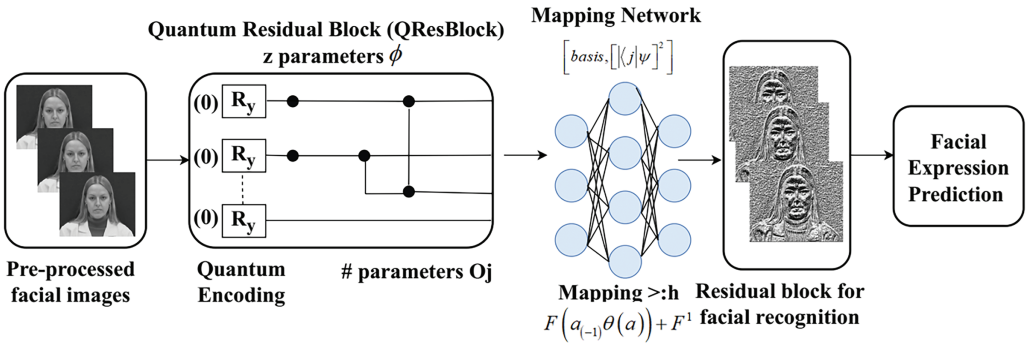


Figure 3: QResNet-Face model

A quantum registers of $N = \left[\log_2 N\right]$ qubits is constructed for encoding these features into a quantum state, generating a 2^M. The face features are encoded by the quantum state $|\psi\rangle \in \mathbb{C}^{2^M}$, such that the squared amplitudes $p_j = |\langle j|\ \psi\rangle|^2$ for basis states $|j\rangle$, $j = 0, 1, \ldots, 2^M - 1$. This represents the probabilities ranging between [0, 1] associated with feature parameters. These probabilities need to be mapped again to classical neural network parameters $\theta_j \in \mathbb{R}$. This is completed by a classical parameter mapping network $L_{\vec{\gamma}}$ with $\vec{\gamma}$. It utilizes $y_j = \left[bin\,(j), p_j\right]$ as a input vector and outputs the parameter $\theta_j = L_{\vec{\gamma}}\left(y_j\right)$, where the binary representation of the index j is represented as $bin(j)$. The parameterized rotation gates $Ry(\mu)$ acting on every qubit are employed to construct the quantum circuit. Eq. (31) defines a quantum rotation gate, specifically a rotation around the Y-axis of the Bloch sphere. This gate is applied to each qubit in the quantum circuit to encode classical information (such as facial features) into quantum states. The degree of rotation was

controlled by a trainable parameter, determining the extent of change in the quantum state. These rotations served as the fundamental components of the quantum encoding layer, enabling the model to convert classical input data into a quantum representation that could be further processed and entangled using additional quantum operations.

$$Ry(\mu) = \begin{bmatrix} \cos(\mu/2) & -\sin(\mu/2) \\ \sin(\mu/2) & \cos(\mu/2) \end{bmatrix} \tag{31}$$

From the above equation, μ is the trainable angle parameter. Controlled-NOT (CNOT) gates that are arranged in a linear chain introduce entanglement between qubits. This improves the expressiveness of the quantum state and maintains parameter scalability on the order of $O(poly\log(N))$. A full quantum residual block encompasses multiple layers, each layer parameterized by a vector $\vec{\phi}$. The residual learning framework is employed to update the classical features $F^{(a)}$:

$$F^{a} = \rho\left(F^{(a-1)}, \vec{\theta}^{(a)}\right) + F^{(a-1)} \tag{32}$$

In the above equation, ρ indicates the quantum-enhanced transformation function and $\vec{\theta}^{(a)}$ represents the classical parameter that is derived from the quantum measurement probabilities at the layer a. The residual connection assists gradient flow in the training process and resolves the vanishing gradient issues in Deep Networks. The training of the QResNet model covers the reduction of the cross-entropy loss for facial expression classification over classes.

$$L_{CE} = -\frac{1}{M_d}\sum_{m=1}^{M_d}\sum_{c=1}^{C} x_{m,C}\log\hat{x}_{m,\mathrm{C}} \tag{33}$$

From Eq. (33), M_d represents the number of training samples, $x_{m,C}$, and $\hat{x}_{m,C}$ denotes the true and predicted one-hot labels, respectively. The gradients for the quantum parameters $\vec{\phi}$ are acquired by the parameter-shift rule. Gradients for the classical parameters are computed via backpropagation.

$$\frac{\partial\langle O\rangle}{\partial\phi_k} = \frac{1}{2}\left[\langle O\rangle_{\phi_k+\frac{\pi}{2}} - \langle O\rangle_{\phi_k-\frac{\pi}{2}}\right] \tag{34}$$

From the above equation, represents the expectation of the measurement observable dependent on the parameter ϕ_k.

4.4.3 Fusion Strategy in QBioFusionNet

The QBioFusionNet combines the biometric features extracted from voice and facial images using a sophisticated fusion strategy. The voice features $\vec{v} \in \mathbb{R}^{d_v}$ are generated by the P-1DWCBiNet branch, and facial features $\vec{f} \in \mathbb{R}^{d_f}$ are produced by QResNet. A unified representation for classification is generated by normalizing these two features, providing $\widehat{\vec{v}}$ and $\widehat{\vec{f}}$. The concatenation-based early fusion is used for the feature fusion:

$$\vec{w} = \left[\widehat{\vec{v}}; \widehat{\vec{f}}\right] \in \mathbb{R}^{d_v+d_f} \tag{35}$$

This process stacks the two normalized vectors into a single vector and ensures that the unique characteristics of the two modalities are preserved. The concatenated feature is passed through an FC layer for mapping it into a common latent space suitable for learning. In the equation below, $\sigma(.)$ represents the element-wise activation function, $\vec{b}_f \in \mathbb{R}^d$, and $\vec{w}'$ is a compressed and fused representation that encodes both voice and facial information.

$$\vec{w}' = \sigma\left(Z_f \vec{w} + \vec{b}_f\right) \in \mathbb{R}^{d_w} \tag{36}$$

In the above equation, $\sigma(.)$ is the element-wise activation, $\vec{b}_f \in \mathbb{R}^d$ denotes the bias vector, and the $Z_f \in \mathbb{R}^{d_w \times (d_v \times d_f)}$ is the weight matrix. The output $\vec{w}'$ is a fused representation encoding two modalities (voice and face). Based on this vector, the softmax classifier outputs identity predictions.

$$\hat{x} = soft\max\left(Z_c \vec{w}' + \vec{b}_c\right) \tag{37}$$

This fusion strategy enables the network to utilize complementary strengths, i.e., temporal audio dynamics from voice and spatial-quantum features from facial images. This offers a robust and discriminative representation for biometric recognition tasks. Fig. 4 illustrates the structural representation of the QBioFusionNet approach.

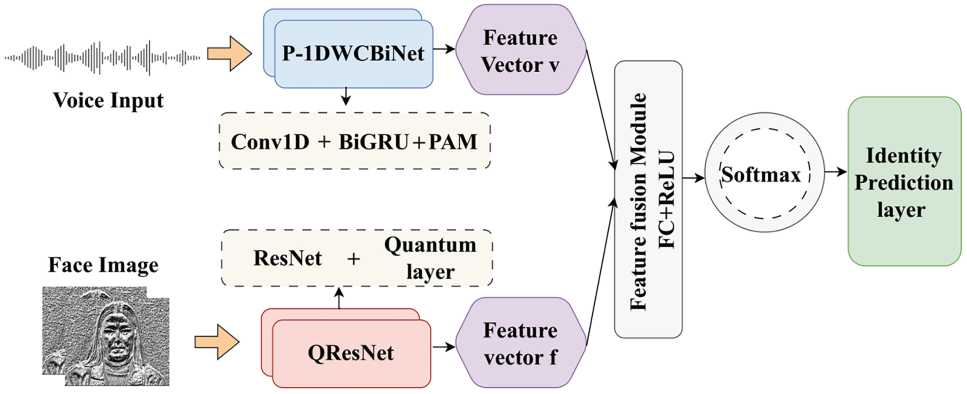


Figure 4: Structural representation of the QBioFusionNet

4.5 Cybersecurity-Aware Decision Module in QBioFusionNet

After the early fusion of normalized voice v and f, the combined representation is defined as follows:

$$w = \left[\hat{v}; \hat{f}\right] \tag{38}$$

This fused vector is fed to the FC layer to output a latent representation $h \in \mathbb{R}^k$.

$$h = \phi\left(Z_w + b\right) \tag{39}$$

From the above equation, $\phi(.)$ indicates the non-linear activation function (ReLU), $Z \in \mathbb{R}^{k \times d}$ is the weight matrix, and $b \in \mathbb{R}^k$.

Speed Detection Module

The binary cross-entropy loss is employed to train the parallel binary classification for the detection of adversarial or spoofed inputs.

$$S = \sigma\left(Z_S, h + b_S\right) \tag{40}$$

From the above equation, $\sigma(.)$ denotes the sigmoid function and $Z_S \in \mathbb{R}^{1\times k}$ indicates the spoof probability score. The following equation indicates the spoof detection loss L_{Spoof}:

$$L_{Spoof} = -\left[x_S \log(\mathrm{S}) + (1 - x_S)\log(1 - S)\right] \tag{41}$$

From the above equation, $x_S \in \{0,1\}$ specifies the ground truth label.

Trust Score Computation (Anomaly Detection)

Anomaly Detection is combined by applying a reconstruction-based scoring function. Let's assume the autoencoder $U(.)$ learn to reconstruct genuine feature representations. The following equation denotes the reconstruction loss.

$$L_R = \|w - U(w)\|_2^2 \tag{42}$$

This loss is normalized to compute the trust score $T \in [0,1]$:

$$T = \exp\left(-\alpha.L_R\right) \tag{43}$$

From the above equation, α denotes the tunable sensitivity parameter. A greater trust score represents input consistency with genuine biometric patterns.

Final Detection Logic

Consider $p \in \mathbb{R}^C$ is the identity prediction vector from the softmax output $Z_c h + b_c$ with the number of classes C. The following equation represents the final authentication decision $F \in \{0,1\}$:

$$F = \begin{cases} 1 & if\ \max(\mathrm{p}) > \mathrm{T}_p \wedge S < T_S \wedge T > T_T \\ 0 & Otherwise \end{cases} \tag{44}$$

From the above equation, T_T indicates the minimum trust score, T_S denotes the spoof detection threshold (0.5), and T_P represents the confidence threshold.

5 Experimental Analysis

This section provides a comprehensive evaluation of the proposed model's performance, employing standard metrics to assess its effectiveness in advanced voice and face recognition.

5.1 Experimental Setup

The experimental setup used a machine configured with an Intel Core i7 CPU, 32 GB of RAM, and an NVIDIA RTX 3090 GPU. Model development and execution are conducted using Python 3.10 and the TensorFlow library on an Ubuntu 20.04 platform.

5.2 Parameter Settings

The parameter setting phase involved configuring the model parameters to optimize overall performance. Table 1 presents the parameters along with their corresponding values.

Table 1: Parameter setting

Methods	*Parameters*	*Values*
P-1DWCBiNet	LayerNorm	64
	Conv1d_1	Activation function = ReLU, Kernel size = 256, Filters = 32, stride = 1,
	MaxPool1d_1	Kernel size = 2
	Conv1d_2	Activation function = ReLU, Filters = 64, stride = 1, kernel size = 2,
	MaxPool1s_2	Kernel size = 2
	Pyramid Attention	Scales s = {1, 2, 4}
	BiGRU	Input size = 64, hidden size = 32, bidirectional = true
	GAP	Output size = 1
	FC	In feature = 64, out-features = 6, activation function = softmax
QResNet	Fusion Layer	Concatenation + Dense + ReLU
	Classifier	Dense
	Optimizer	Adam
	Epochs	100
	Batch Size	64

5.3 Evaluation Parameters

Evaluation metrics are quantifiable measures that are employed to estimate the model's effectiveness. The different evaluation metrics are deployed, which are defined as follows:

F1-score: The F-score is an important metric used for classification tasks, combining precision and recall to accurately identify positive cases.

Recall: Using the recall metric, the actual positive classes in a dataset are measured.

Precision: Precision metric is employed to measure how many of the instances are correctly predicted as positive by the proposed model.

Accuracy: Accuracy reflects the overall correctness of the proposed model's classification.

Specificity: True negatives of the model are measured by using specificity.

Equal Error Rate: Equal Error Rate (EER) plays an important role in evaluating the trade-off between security.

Mean Average Precision: Mean Average Precision (mAP) is calculated based on the average precision across each class.

Matthews Correlation Coefficient: The Matthews Correlation Coefficient (MCC) is measured based on the four important elements.

5.4 Performance Analysis

The proposed model confirms its capability for voice and face recognition using several evaluation measures. Table 2 provides the visual representation of the pre-processed face images. The raw input images are gathered from the dataset and undergoes the pre-processing step. In the face detection and cropping stage, the facial images are detected, and the background areas are eliminated. The next step (face alignment) aligns the key landmarks of each face image. The resizing step provides the resized images, optimizing model compatibility. Ultimately, the normalization step ensures that uniform input data is generated for further steps. Table 3 shows the grayscale conversion and LBP transformation with a histogram employed for feature

extraction. The grayscale image is the simplified input representation by eliminating color information. The LBP-converted image highlights that relationships between a pixel and its neighboring pixels are encoded. LBP histogram is the discrimination feature vector utilized for face recognition.

Table 2: Visual representation of the pre-processed face images

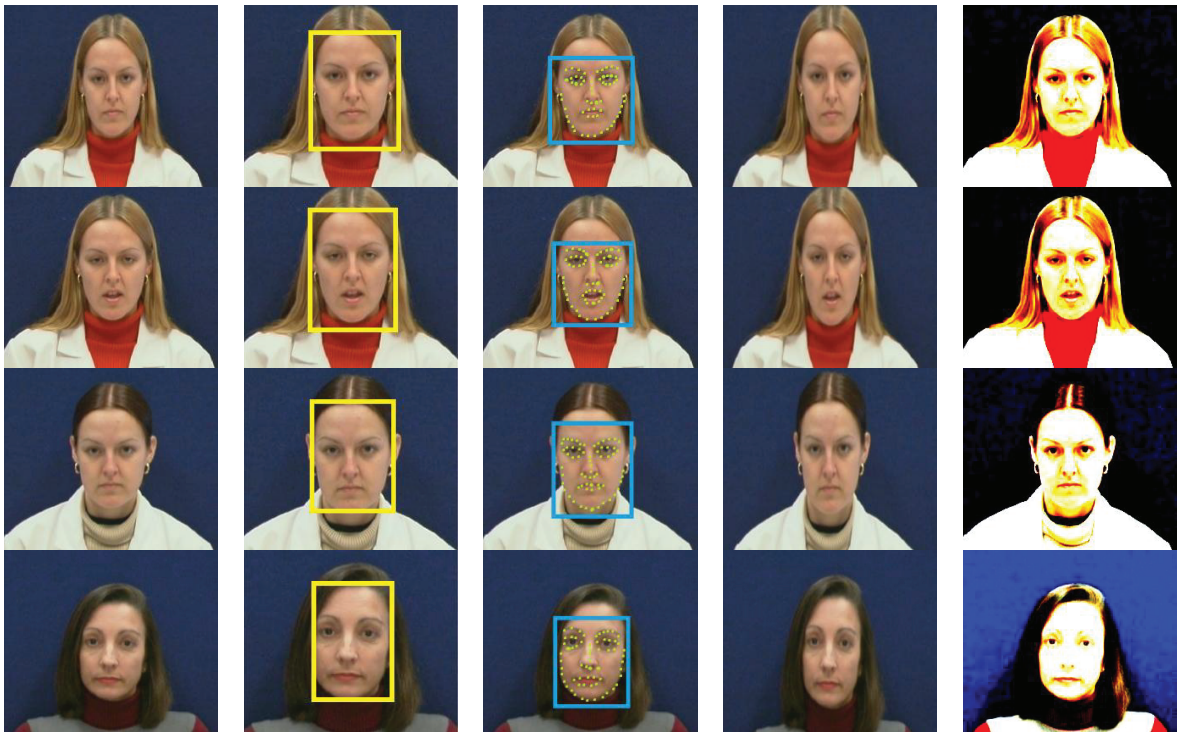
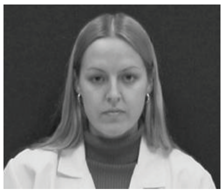
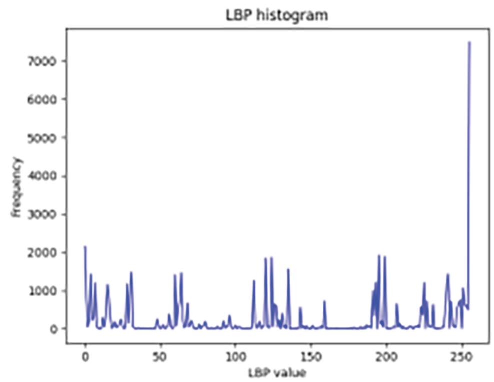
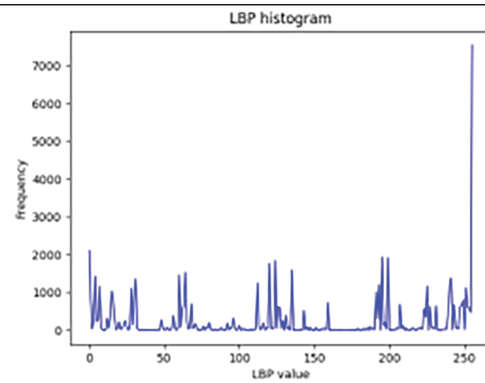
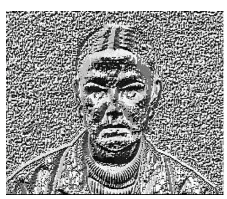
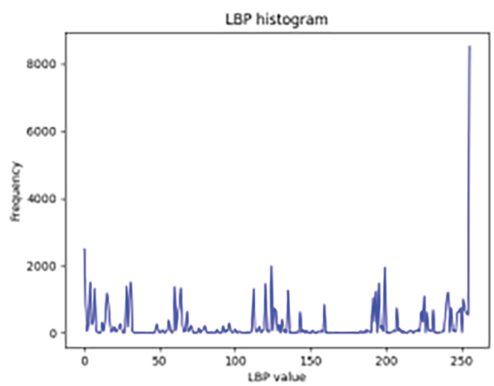
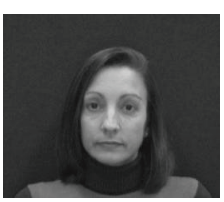
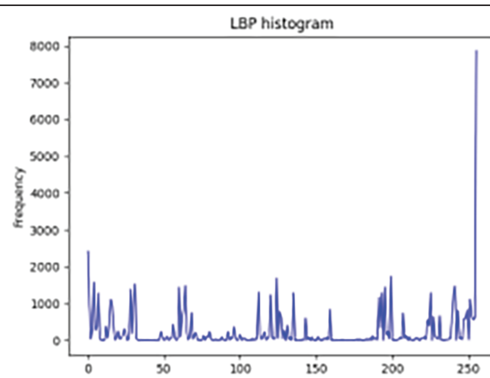
Input images	Pre-processed images			
	Face detection and cropping	Face alignment	Resizing	Normalization
				

Table 3: Visual representation of faces after applying grayscale conversion and LBP

<table>
<tr><th>Grayscale image</th><th>LBP converted image</th><th>LBP histogram</th></tr>
<tr><td>  </td><td>  </td><td> 
</td></tr>
<tr><td>  </td><td>  </td><td> 
</td></tr>
<tr><td>  </td><td>  </td><td> 
</td></tr>
<tr><td>  </td><td>  </td><td> 
</td></tr>
</table>

The Area Under the Receiver Operating Characteristic curve (AUC-ROC) and the Area Under the Precision-Recall curve (AUC-PR) analyses between the voice, face, and fusion models are illustrated in Fig. 5a,b. As shown in the graphs, the fusion strategy (QBioFusionNet) attains the highest AUC rate of 0.988 compared to the voice and face models. This ensures its superior ability to distinguish between different classes. The fusion model provides the best AUC-PR value of 0.975, which is higher than the face and voice models. This confirms the model's robustness in handling imbalanced data and attaining superior precision and recall.

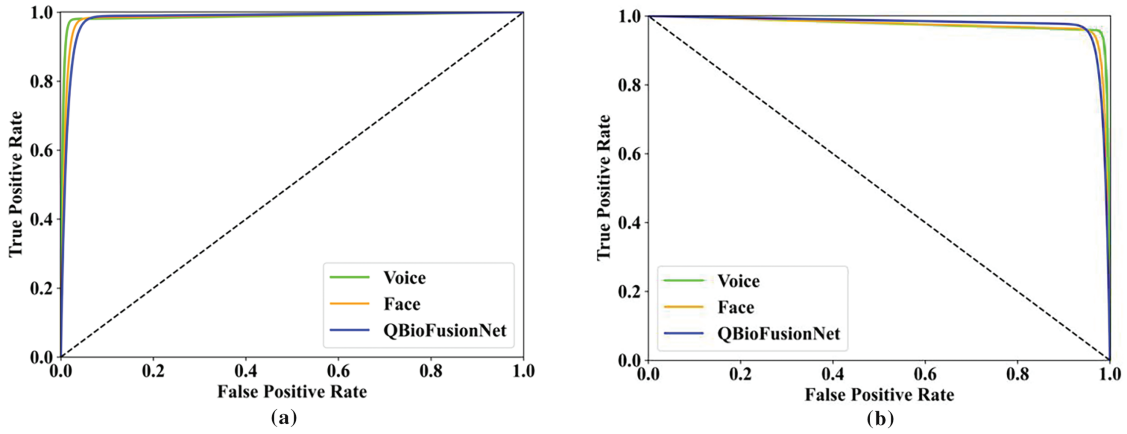


Figure 5: Performance analysis based on (**a**) AUC-ROC, (**b**) AUC-PR

The comparative performance across each modality and fusion method is represented in Table 4. As shown in the table, the performance of the voice (P-1DWCBiNet) and face (QResNet) is lower than the QBioFusionNet. It demonstrates higher recall, accuracy, precision, specificity, and F1-score, highlighting its strong capability for face and voice recognition. Its EER is lower than the voice and face models, indicating reliability in verification tasks. The AUC-ROC rate of the proposed method ensures high performance than the voice and face modalities.

Table 4: Performance analysis for voice, face, and fusion methods

Measures	Voice only (P-1DWCBiNet)	Face only (QResNet)	QBioFusionNet
Accuracy	97.90%	97.2%	98.99%
Precision	96.89%	96.17%	97.54%
Recall	95.67%	96.78%	97.32%
Specificity	95.10%	95.12%	96.78%
F1-score	96.28%	96.47%	97.43%
EER	5.5%	6.0	1.0%
AUC-ROC	0.978	0.972	0.988

5.5 Comparative Analysis

The comparative analysis ensures that the proposed model is a superior method compared to the existing methods that attain poor performance. Table 5 presents the evaluation of the proposed approach across various performance parameters. The proposed model delivers superior performance in F1-score, recall, specificity, accuracy, and precision than the AVSR-GAN [13], IFR-CNN [14], ERNN [15], SbCNN [16], HPODL-BVCS [19], and MEWODL-ER [21]. The proposed approach attains a higher F1-score and accuracy.

The higher F1-score specifies that it maintains a balanced trade-off between precision and recall, which is useful for biometric systems. The higher accuracy of the proposed model demonstrates robust recognition performance. The enhanced specificity confirms that the proposed model accurately identifies individuals, while the superior recall ensures it effectively captures all true positive identities.

Table 5: Performance comparison of the proposed method with related approaches

Methods	Precision	Specificity	F1-Score	Accuracy	Recall
Proposed	97.54%	96.78%	97.43%	98.99%	97.32%
AVSR-GAN	92.13%	90.43%	91.84%	93.76%	91.56%
IFR-CNN	93.81%	91.26%	93.36%	94.61%	92.91%
ERNN	93.60%	91.10%	93.12%	93.45%	92.65%
SbCNN	94.34%	92.53%	93.87%	95.60%	93.41%
HPODL-BVCS	93.40%	93.40%	93.09%	94.34%	92.79%
MEWODL-ER	94.92%	92.39%	94.03%	95.11%	93.15%

Comparative analysis of EER, mAP, and MCC are illustrated in Fig. 6. Existing methods deliver lower performance in EER, mAP, and MCC than the proposed model. It ensures a sufficient MCC of 0.97%, indicating that the developed approach is effective in identifying positive and negative instances. The superior mAP attained by the proposed model is 0.976, specifying that it effectively recognizes the correct speaker and verifies correct face images. The proposed model exhibits a lower EER compared to existing methods, indicating that it is more reliable and less prone to errors.

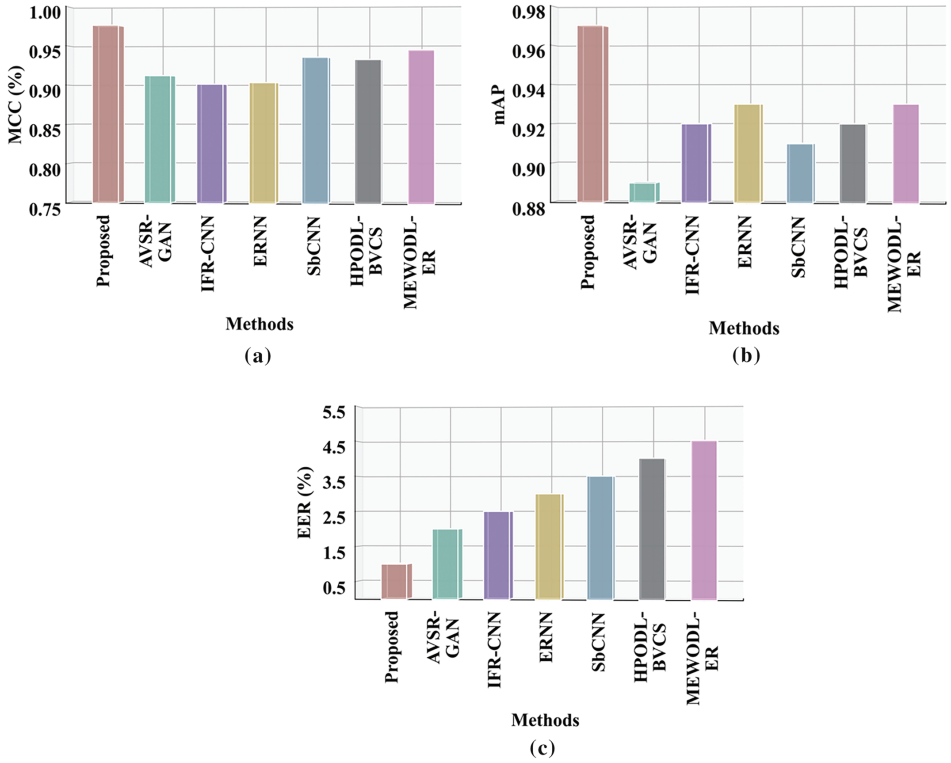


Figure 6: Comparative Estimation in terms of three standard indicators (**a**) MCC, (**b**) mAP, (**c**) EER

Table 6 presents the security analysis using several security measures. As shown in table, the proposed model demonstrates efficient performance compared to the existing methods. This highlights that the proposed model is more reliable and improves the security and trustworthiness of the biometric systems.

Table 6: Security analysis

Methods	Spoof attack detection (%)	False acceptance rate (FAR) (%)	False rejection rate (FRR) (%)	Data privacy leakage rate (%)	Model robustness score (%)	Authentication time (ms)
Proposed	98.7	1.2	1.5	0.8	97.3	120
AVSR-GAN	93.8	3.5	3.7	2.6	91.7	140
IFR-CNN	92.2	2.6	3.5	3.5	90.4	136
ERNN	91.7	4.1	5.7	4.7	91.4	150
SbCNN	90.4	5.7	2.6	3.2	93.8	135
HPODL-BVCS	91.4	6.1	7.3	5.7	91.8	145
MEWODL-ER	93.7	3.7	3.8	6.1	92.4	153

5.6 Ablation Study

Through an ablation study, this subsection illustrates that the proposed model performs more effectively when each individual component contributing to face and voice recognition is included. Table 7 shows the ablation study based on the standard evaluation parameters. The voice and face models attain poor performance across all parameters when compared to the proposed QBioFusionNet. As shown in the table, the fusion model significantly impacts performance when removing the quantum block. The proposed QBioFusionNet model gains superior performance, indicating that each component improves the model's performance for secure biometric authentication.

Table 7: Ablation study analysis

Model variant	Precision	Specificity	F1-Score	EER	Accuracy	Recall
Voice only (P-1DWCBiNet)	94.65%	92.53%	94.03%	3.5%	95.50%	93.42%
Face only (QResNet)	95.51%	93.21%	95.12%	4.0%	97.70%	94.73%
Fusion without quantum block	93.40%	93.40%	93.09%	4.5%	94.34%	92.79%
Proposed QBioFusionNet	97.54%	96.78%	97.43%	1.0%	98.99%	97.32%

5.7 Sensitivity Analysis

The robustness and adaptability of the developed system across six different real-world scenarios are evaluated in Table 8. Under Case 1, the proposed model ensures superior performance for effective voice and face recognition. Although cases 2 and 6 demonstrate stable detection and acceptable to reliable predictions, case 2 exhibits slightly lower performance. Case 3 shows less stable predictions, and case 4 demonstrates strong adaptability. However, case 5 faces challenges that impact the performance and are highly sensitive to facial obstructions. Overall, these outcomes ensure the strength of the proposed model under typical and slightly adverse conditions.

Table 8: Sensitivity analysis

Evaluation case	Overall identification performance	Correct match rate	Detection capability	Consistency of predictions	Rejection vs. Acceptance balance
Case 1 (Proposed model)	Excellent	Very high	High effective	Strong and consistent	Extremely balanced
Case 2 (moderate distortion)	Good	Moderately high	Stable	Acceptable	Slightly degraded
Case 3 (Increased noise)	Fair	Above average	Weakened	Less stable	Noticeable imbalance
Case 4 (Low light condition)	Very good	High	Robust	Quite stable	Well-controlled
Case 5 (Partial occlusion)	Moderate	Moderate	Affected	Inconsistent	Higher deviation
Case 6 (Language variation in voice)	Good	Relatively high	Effective	Reliable	Balanced

Table 9 compares classical biometric models with the quantum-enhanced QBioFusionNet, highlighting modest performance gains in accuracy and EER due to quantum integration.

Table 9: Comparative analysis based on classical versus quantum-enhanced components in QbioFusionNet

Aspect	Classical (Face + Voice)	Quantum-Enhanced (QBioFusionNet)	Analysis/Observation
Voice AUC-ROC	0.978	–	High performance achieved with classical audio processing.
Face AUC-ROC	0.972	–	The classical face model performs slightly lower than the voice.
Fusion AUC-ROC	–	0.988	Quantum-enhanced fusion improves AUC by ~1%.
Accuracy (Table 4)	97.90% (Voice)/97.2% (Face)	98.99%	Fusion provides ~1–2% gain, but it is marginal.
EER (Equal Error Rate)	5.5% (Voice)/6.0% (Face)	1.0%	Notable drop in EER is valuable for security applications.
Computational Overhead	Low	High (quantum simulation)	Quantum component increases complexity.
Hardware Dependency	Standard (CPU/GPU)	Quantum simulator	Limits real-world deployment feasibility.
Theoretical Value	Well-established	Experimental	QResNet serves as a forward-looking research exploration.

5.8 Complexity Analysis

Table 10 presents the computational complexity of each major component within the proposed QBio-FusionNet model and expands all notations used in the analysis.

- The QResNet was implemented using a quantum simulator to explore the feasibility of quantum-enhanced facial feature encoding.
- The model maps classical facial descriptors to quantum states using parameterized rotation gates and simulates entanglement through controlled quantum operations such as CNOT gates.
- While the model contributes to improved performance metrics, such as an AUC-ROC of 0.988 and an EER of 1.0% these results were obtained under simulated conditions.

Table 10: Component-wise complexity analysis

Component	Complexity	Remarks
Wide 1D Conv layers	$O(K \cdot T \cdot F)$ K: kernel size of the convolution T: number of time steps (input sequence length) F: number of filters used in the convolution layers	Large kernels increase computation
BiGRU	$O(n \cdot h^2)$ n:sequence length (number of frames) h: size of the hidden state in the BiGRU	Sequential operations are less parallelizable.
Pyramid attention	$O(s \cdot T)$ s: number of scales in the pyramid attention module T: temporal length of the input feature map	Multiple downsampling/upsampling layers
Quantum state encoding	$O(q \cdot d)$ q: number of qubits used in the quantum circuit d: dimensionality of local facial feature vectors	Per feature vector
Quantum simulation	$O(2q)$	Significant overhead on classical hardware
Fusion + FC	$O(m \cdot n)$ m: length of voice feature vector n: length of face feature vector	Relatively lightweight

5.9 Statistical Analysis

The Friedman test is conducted to statistically validate the performance differences between the existing methods based on standard evaluation parameters. Friedman test analysis is shown in Table 11. This comparison highlights the performance of the proposed biometric system relative to other biometric methods.

The table shows that the performance differences between the developed approach and existing methods are statistically significant. The proposed model consistently outperforms the alternatives, enhancing the robustness and reliability of biometric recognition tasks.

Table 11: Friedman test analysis

Methods	Accuracy	Precision	Recall	Specificity	F1-Score
Proposed	1	1	1	1	1
AVSR-GAN	6	7	7	7	7
IFR-CNN	4	4	4	5	4
ERNN	7	5	6	6	5
SbCNN	2	3	2	3	3
HPODL-BVCS	5	5	5	2	6
MEWODL-ER	3	2	3	4	2

In terms of the estimation indicators, mean rank, and different methods, the value of the chi-square value $\left(X_{\gamma}^{2}\right)$ is computed. The Friedman results for the significance threshold $\sigma = 0.05$ are presented in Table 12. The critical value is 12.592 for the degree of freedom 6(7-1). If the $P_{\gamma} < 0.05$, rejection of the null hypothesis is observed.

Table 12: Statistical findings of the Friedman results with threshold $\sigma = 0.05$

Methods	Chi-square $\left(X_{\gamma}^{2}\right)$	P_{γ} value
QBioFusionNet vs. AVSR-GAN	14.955	0.0043
QBioFusionNet vs. IFR-CNN	15.987	0.0052
QBioFusionNet vs. ERNN	15.456	0.0061
QBioFusionNet vs. SbCNN	17.653	0.0071
QBioFusionNet vs. HPODL-BVCS	14.255	0.0082
QBioFusionNet vs. MEWODL-ER	20.891	0.0053

5.10 Development Considerations

1. Model Latency

- The proposed QBioFusionNet architecture includes computationally intensive components such as wide 1D convolutional layers, BiGRU networks, and a simulated quantum residual network.
- In latency-sensitive applications such as real-time authentication or surveillance, this could impact system responsiveness and user experience.
- Benchmarking latency under realistic conditions will be essential for assessing real-world suitability.

2. Hardware Compatibility

- The QResNet is implemented via simulation on classical hardware, which introduces significant overhead.
- Current quantum hardware is still limited in qubit count, gate fidelity, and accessibility.
- As a result, the model cannot yet be deployed as-is on quantum machines.
- Moreover, the combined model size and memory demands may exceed the capabilities of edge devices, posing challenges for embedded or mobile applications.

3. Deployment Challenges

- Beyond computational complexity, integrating the full QBioFusionNet pipeline in operational systems would require seamless coordination between audio and video streams.
- These engineering and system integration challenges remain unaddressed in the current scope of this research.

4. User Acceptability

- User-facing biometric systems must balance accuracy with usability.
- While QBioFusionNet shows strong recognition performance, no assessment was conducted regarding user response time or tolerance to partial occlusion.
- Acceptability studies, including response latency thresholds and demographic fairness, will be necessary before large-scale deployment.

6 Conclusion and Future Directions

This work presents an effective multi-model system that facilitates voice and face recognition, namely QbioFusionNet. The data collection is the main phase of the face recognition systems to acquire the required data (face images and voice recordings) using the VidTIMIT dataset. The collected data undergo the pre-processing steps, where suitable pre-processing steps are applied to each data point, which includes face and audio recordings. Advanced feature extraction strategies are deployed on the pre-processed voice and face modalities. Augmentation steps generate face and audio samples by creating variations in existing face images and voice recordings. Then, voice and face recognition are performed using robust methods, such as P-1DWCBiNet and QResNet. The P-1DWCBiNet method is used to process the audio recordings to extract the facial features, and the QResNet model enables the encoding of the facial features, capturing complex spatial patterns using quantum state representations. The proposed fusion mechanism integrates these multi-model features, improving the accuracy performance. The detailed experimental results are performed using different evaluation parameters, such as precision, accuracy, F1-score, recall, specificity, mAP, EER, and MCC. The comparative analysis demonstrates that the QBioFusionNet strategy gains greater efficiency compared to other biometric methods. Moreover, sensitivity analysis is conducted to evaluate the robustness of the QBioFusionNet mechanism across varying input conditions. This demonstrates its stability and reliability in real-world biometric recognition scenarios.

Despite the proposed model providing significant advancements for identity recognition, it has the following limitations that need to be solved in future works:

- **Quantum Simulator Constraints:** Instead of using real quantum hardware, the face model (QResNet) is evaluated utilizing a simulated backend.
- **Limited Demographic Diversity in Dataset:** The VidTIMIT dataset lacks explicit diversity in age, ethnicity, and accent, which may limit the model's generalizability. Despite achieving 98.99% accuracy, the small performance gap between unimodal and fusion models suggests possible overfitting to the training distribution, raising concerns about real-world demographic performance.
- **Limited Resistance to Advanced Spoofing Techniques:** While QBioFusionNet incorporates a spoof detection module and achieves a low EER of 1.0%, the framework does not explicitly address increasingly sophisticated attack vectors such as deepfake facial videos or AI-synthesized voice clones. These advanced spoofing methods could pose challenges to the system in high-security environments where zero tolerance to identity fraud is essential.

Future Works

- **Integration with Real Quantum Hardware:** The proposed model will be implemented and tested on real quantum computing hardware to explore the true quantum advantages and estimate the real-world feasibility under hardware constraints.
- **Defense against Deepfake and Synthetic Attacks:** Extend the current spoof detection module to incorporate deepfake detection algorithms, adversarial training strategies, and liveness analysis, making the system resistant to advanced generative attacks targeting voice and facial biometrics.
- **Broader Dataset Evaluation:** Future work will evaluate the model on larger, demographically diverse datasets to assess fairness and robustness across varied populations and reduce potential bias in biometric recognition.

Acknowledgement: The authors declare that no additional individuals contributed to this work outside of the listed authors. No external volunteers or patients were involved in this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Abrar M. Alajlan conceived the research idea, developed the theoretical framework, and performed the computational analyses. The analytical methods were verified by Abdul Razaque. All authors contributed to the interpretation of the results, participated in the critical revision of the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available at https://conradsanderson.id.au/vidtimit/ (accessed on 18 September 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Li J, Yi Q, Lim MK, Yi S, Zhu P, Huang X. MBBFAuth: multimodal behavioral biometrics fusion for continuous authentication on non-portable devices. IEEE Trans Inf Forensics Secur. 2024;19(11):10000–15. doi:10.1109/TIFS.2024.3480363.
2. Helmy M. Proposed cancelable biometrics system based on hybrid optical crypto-steganography audio framework for cyber security applications. Multimed Tools Appl. 2025;84(17):17979–8004. doi:10.1007/s11042-024-18232-w.
3. Reddy Rachapalli D, Dondeti V, Kalluri HK. Multimodal cancellable biometric template protection and person verification in transformed domain. IEEE Access. 2024;12(17):173557–82. doi:10.1109/access.2024.3501368.
4. Prasad PS, Lakshmi AS, Kautish S, Singh SP, Shrivastava RK, Almazyad AS, et al. Robust facial biometric authentication system using pupillary light reflex for liveness detection of facial images. Comput Model Eng Sci. 2024;139(1):725–39. doi:10.32604/cmes.2023.030640.
5. Krishna Prakasha K, Sumalatha U. Privacy-preserving techniques in biometric systems: approaches and challenges. IEEE Access. 2025;13(2):32584–616. doi:10.1109/access.2025.3541649.
6. Vekariya V, Joshi M, Dikshit S, bargavi Manju SK. Multi-biometric fusion for enhanced human authentication in information security. Meas Sens. 2024;31(2):100973. doi:10.1016/j.measen.2023.100973.
7. Sharma S, Saini A, Chaudhury S. Multimodal biometric user authentication using improved decentralized fuzzy vault scheme based on Blockchain network. J Inf Secur Appl. 2024;82(2):103740. doi:10.1016/j.jisa.2024.103740.
8. Mansour A, Eddermoug N, Sadik M, Sabir E, Azmi M, Jebbar M. A lightweight seamless unimodal biometric authentication system. Procedia Comput Sci. 2024;231(3):190–7. doi:10.1016/j.procs.2023.12.192.
9. Halbouni A, Gunawan TS, Habaebi MH, Halbouni M, Kartiwi M, Ahmad R. Machine learning and deep learning approaches for CyberSecurity: a review. IEEE Access. 2022;10:19572–85. doi:10.1109/access.2022.3151248.

10. Sumalatha U, Prakasha KK, Prabhu S, Nayak VC. A comprehensive review of unimodal and multimodal fingerprint biometric authentication systems: fusion, attacks, and template protection. IEEE Access. 2024;12(9):64300–34. doi:10.1109/access.2024.3395417.
11. Abdelfatah RI. Robust biometric identity authentication scheme using quantum voice encryption and quantum secure direct communications for cybersecurity. J King Saud Univ Comput Inf Sci. 2024;36(5):102062. doi:10.1016/j.jksuci.2024.102062.
12. Al-Abboodi RH, Al-Ani AA. A novel technique for facial recognition based on the GSO-CNN deep learning algorithm. J Electr Comput Eng. 2024;2024(1):3443028. doi:10.1155/2024/3443028.
13. He Y, Seng KP, Ang LM. Generative adversarial networks (GANs) for audio-visual speech recognition in artificial intelligence IoT. Information. 2023;14(10):575. doi:10.3390/info14100575.
14. Suneetha C, Anitha R. Speech based emotion recognition by using a faster region-based convolutional neural network. Multimed Tools Appl. 2025;84(8):5205–37. doi:10.1007/s11042-024-19004-2.
15. Talaat FM. Explainable enhanced recurrent neural network for lie detection using voice stress analysis. Multimed Tools Appl. 2024;83(11):32277–99. doi:10.1007/s11042-023-16769-w.
16. Gokulakrishnan S, Chakrabarti P, Hung BT, Shankar SS. An optimized facial recognition model for identifying criminal activities using deep learning strategy. Int J Inf Technol. 2023;15(7):3907–21. doi:10.1007/s41870-023-01420-6.
17. Alsafyani M, Alhomayani F, Alsuwat H, Alsuwat E. Face image encryption based on feature with optimization using secure crypto general adversarial neural network and optical chaotic map. Sensors. 2023;23(3):1415. doi:10.3390/s23031415.
18. Lin N, Ding Y, Tan Y. Optimization design and application of library face recognition access control system based on improved PCA. PLoS One. 2025;20(1):e0313415. doi:10.1371/journal.pone.0313415.
19. Ragab M, Alghamdi BM, Alakhtar R, Alsobhi H, Maghrabi LA, Alghamdi G, et al. Enhancing cybersecurity in higher education institutions using optimal deep learning-based biometric verification. Alex Eng J. 2025;117(2):340–51. doi:10.1016/j.aej.2025.01.012.
20. Aly M, Ghallab A, Fathi IS. Enhancing facial expression recognition system in online learning context using efficient deep learning model. IEEE Access. 2023;11:121419–33. doi:10.1109/access.2023.3325407.
21. Alrowais F, Negm N, Khalid M, Almalki N, Marzouk R, Mohamed A, et al. Modified earthworm optimization with deep learning assisted emotion recognition for human computer interface. IEEE Access. 2023;11:35089–96. doi:10.1109/access.2023.3264260.
22. Gona A, Subramoniam M, Swarnalatha R. Transfer learning convolutional neural network with modified Lion optimization for multimodal biometric system. Comput Electr Eng. 2023;108(2):108664. doi:10.1016/j.compeleceng.2023.108664.
23. Sun W, Sun F, Gao H, Zhang L, Xiao K, Xiao P. Anti-spoofing performance assessment of the vector tracking loop in challenging environments. IEEE Trans Instrum Meas. 2025;74(3):8510312. doi:10.1109/TIM.2025.3583376.
24. Radad M, Enab AE, Elagooz SS, El-Fishawy NA, El-Rashidy MA. Face anti-spoofing detection based on novel encoder convolutional neural network and texture's grayscale structural information. Int J Comput Intell Syst. 2025;18(1):175. doi:10.1007/s44196-025-00757-z.
25. Ma Y, Lyu C, Li L, Wei Y, Xu Y. Algorithm of face anti-spoofing based on pseudo-negative features generation. Front Neurosci. 2024;18:1362286. doi:10.3389/fnins.2024.1362286.
26. Ma X, Guo J, Sansom A, McGuire M, Kalaani A, Chen Q, et al. Spatial pyramid attention for deep convolutional neural networks. IEEE Trans Multimed. 2021;23:3048–58. doi:10.1109/TMM.2021.3068576.