



The Scientific Temper

JULY 2026 Vol. 17 (7): 6599-6617

ISSN: 0976-8653, E-ISSN: 2231-6396

DOI: 10.58414/SCIENTIFICTEMPER.2026.17.7.2498

<https://scientifictemper.com/>

RESEARCH ARTICLE

Multimodal Biometric Recognition Based on Face and Video Via Vision Attention Fusion Transformer

S. PREETHI¹, P. JOSEPH CHARLES²

Affiliated to Bharathidasan University, Tiruchirappalli- 620002.

Corresponding author: josephcharles_cs1@mail.sjctni.edu

ABSTRACT

Biometric recognition systems have become an indispensable piece of contemporary security and recognition administration frameworks. Multimodal biometric systems incorporate multiple biometric features for verification provide significant advantages over uni-modal biometric systems, to name a few being, improved security, enhanced accuracy and considerable amount of robustness. Across a sample of individual modalities, face recognition is extensively applied for its reliability and ease of use, voice recognition is eminent for its high accuracy and stability. This study highlights the significance of multimodal biometric authentication method using face images and voice signals called, Multi Task Cascaded Pre-emphasis and Vision Attention Fusion Transformer (MTCP-VAFT) to enhance the security of existing multimodal biometric recognition systems in indoor surveillance videos. It introduces a novel pre-processing model that simultaneously processes face images using Multi Task Cascaded Neural model and voice signals utilizing and Pre-emphasis Filter. This work also proposes a novel Vision Attention Fusion Transformer based multimodal biometric authentication to exploit their complementary advantages and mitigate attacks. Finally the method incorporates the two modalities via an attention fusion mechanism to highlight the multimodal biometric recognition system reliability under varying circumstances. Experimental results on the MSU-AVIS dataset demonstrate the efficiency of our method, showing notable improvements achieving overall multimodal biometric recognition accuracy by 97% and improving the peak signal to noise ratio by 48.25dB. These findings demonstrate that modality-aware fusion using MTCP-VAFT method can delivery secure and flexible biometric authentication suitable for deployment on high-security platforms.

Keywords: *Multimodal Biometric, Multi Task Cascaded Neural, Pre-emphasis Filter, Vision Transformer, Attention Mechanism, Fusion*

INTRODUCTION

Biometric authentication is progressively utilized as secure and suitable optimal to conventional identity verification. Yet, systems that depend on a single trait such as face or audio frequently fail in practical settings owing to spoofing, noise and variance across users. Existing multimodal methods address certain of these issues, however, depend on pre-trained models or depend on straightforward fusion mechanisms that do not adjust to data quality.

An innovative hybrid multimodal biometric system (MBS) incorporating visible (VIS) and infrared (IR) facial imagery was proposed (Mahum et al., 2025). Hybrid MBS applied weighted score-level fusion to ensure accurate and precise prediction. Moreover, to extract prominent features VGG16 was applied and dimensionality reduction was ensured by means of Principal Component Analysis (PCA). Finally Sequential Neural Networks (SNNs) was applied for Classification therefore enhancing robustness and accuracy for both verifications.

Despite improvements observed in precision and accuracy, however multimodality involving different biometrics was not analyzed. To address on this aspect, a Multi Task Cascaded Pre-emphasis and Vision Attention Fusion Transformer (MTCP-VAFT) is introduced that by focusing on the face image and voice signals results in improved precision and accuracy.

A user authentication method, Artificial Neural Network combined with Fuzzy Nonlinear Support Vector Machine (ANN + fuzzy-NL-SVM), was proposed to decode passwords from acoustic signals (Khan et al., 2023). This method used user speaking to enter their private password. Followed by which with the intent of tracking the code, an artificial neural network (ANN) was employed in addition to speech parameters like energy and Mel frequency cepstral coefficient. Next, by using frequency and phase properties the target's voice was compared with the identified password key, therefore safeguarding from speech forgeries. Finally a classifier employing nonlinear support vector machine was used that in turn not only improved true acceptance rate but also reduced false acceptance rate with improved accuracy. In spite of improved accuracy with minimal false acceptance, the method only supported voice signals. To focus on the aspect of multimodal biometric involving face images and voice signals as input, MTCP-VAFT is introduced with improved peak signal to noise ratio.

A cascaded transformer framework using multi-scale transformer feature extraction was proposed to address issues related to computational complexity (Li et al., 2025). By applying this technique not only minimized inference time but also improved accuracy to a greater extent. A detailed evaluation study focusing on voice-based recognition for nursing documentation was presented (Mahum et al., 2025). A small-scale transformer architecture with minimal error rates was designed (Soumya & Charles, 2026). Nevertheless, biometric systems remain highly susceptible to spoofing attacks that impersonate legitimate users. These vulnerabilities were addressed using a Vision Transformer-enhanced architecture with improved accuracy (Appati et al., 2025).

Advancements in audio synthesis techniques have led to the creation of highly realistic audio deepfakes, posing significant threats to public trust. An intelligent voice query method was designed to enhance detection efficiency (Gong & Li, 2025). A voice synthesis mechanism employing diffusion and entanglement models was proposed (Cheng et al.,

2024). By incorporating these techniques, accurate recognition was achieved. A review of machine learning and deep learning methods for deepfake detection was presented (Gong & Li, 2025). Voice-based attacks have become increasingly prevalent with the rapid digitalization of communication technologies. To address this challenge, a diffusion-aware attention mechanism combined with a dual-attention model was proposed, achieving a lower error rate in biometric recognition (Chen et al., 2024).

With the increasing dependence on digital technologies across various sectors, securing access to sensitive information has become a critical challenge. Conventional authentication mechanisms are increasingly inadequate because of their susceptibility to theft, spoofing, and human error. As attackers continue to develop more sophisticated attack strategies, the demand for secure and user-friendly authentication solutions has grown. A novel Swin Transformer-based framework was proposed to achieve precise speech-based biometric recognition (Mahum et al., 2025). A speech recognition technology based on a transformer end-to-end model was presented, resulting in improved overall recognition accuracy (Li et al., 2025). A multimodal fusion-based biometric authentication mechanism was designed using a context-driven scoring algorithm, resulting in a low equal error rate and improved authentication accuracy (Vatchala et al., 2025). Face-voice multimodal identity matching was investigated using cross-modal identity information to evaluate recognition accuracy (Smith et al., 2025). A Vision Transformer-based face recognition method was proposed to enhance recognition efficiency and robustness (Appati et al., 2025).

RELATED WORKS

Information security consists of protecting information's integrity, availability and confidentiality. Several materials and methods that can help assist with information security management. However to certain extent biometric systems have evolved with certain areas of data security. A dual shared-layer architecture based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) was proposed to improve feature extraction for biometric authentication (Zhang et al., 2024). However, the proposed approach did not incorporate multimodal fusion mechanisms, limiting its ability to exploit complementary information from multiple biometric modalities. A transformer-based end-to-end speech recognition model was proposed to

enhance recognition accuracy in real-time interactive environments (Li et al., 2025). Nevertheless, the study did not address multimodal biometric authentication by integrating complementary biometric modalities such as face and voice. A character-based transformer architecture was proposed for end-to-end automated speech recognition to improve recognition performance and reduce error rates (Loubser et al., 2024). A Swin Transformer-based framework was proposed to improve speech recognition performance for individuals with dysarthria by enhancing recognition accuracy and robustness (Mahum et al., 2025).

Human beings control the potentiality to connect voices with faces, even for unknown individuals. Over the recent years, deep neural networks by using computational mechanism mimics humans in connecting unheard voices with unseen face images.

A machine learning and metaheuristic-based framework was proposed to enhance voice-based authentication for secure mobile banking applications (Nosrati et al., 2024). An intelligent voice query framework was proposed to enhance the efficiency of industrial robot maintenance through speech-based interaction (Ren et al., 2024). The effectiveness of multimodal identity information for face and voice identity matching was investigated, revealing that multimodal information did not significantly improve identity matching accuracy (Smith et al., 2025). However, with the presence of both intrinsic and extrinsic variability hinder the performance improvement. Hybrid deep learning and customized CNN-based frameworks were proposed to improve the accuracy of handwriting-based learning disability detection (Soumya & Charles, 2026a, 2026b).

A voice-based structured nursing documentation framework integrating automatic speech recognition and large language models was developed to improve the accuracy and efficiency of clinical documentation (Su et al., 2026). A multi-stage face-voice association learning framework incorporating keynote speaker diarization was

proposed to improve cross-modal association between facial and vocal representations (Tao et al., 2024). A shared-layer multimodal biometric authentication framework was proposed to enhance authentication performance through joint face and voice feature learning (Vatchala et al., 2025). A deep learning-based multimodal image fusion model incorporating multiple filters was proposed to improve image fusion quality by effectively preserving complementary information from multiple modalities (Wang et al., 2025). An interpretable adaptive spectro-temporal diffusion transformer was proposed to improve audio deepfake detection by effectively learning spectro-temporal speech representations (Wani et al., 2025).

Motivated by the above issues, in this work a novel method called, Multi Task Cascaded Pre-emphasis and Vision Attention Fusion Transformer (MTCP-VAFT) employing face images and voice signals is proposed. The detailed explanation of MTCP-VAFT method is provided in the following sub-sections.

MATERIALS AND METHODOLOGY

The paramount ground of multimodal biometric recognition is to improve security and accuracy in identity verification by fusing two or more recognizable biometric identifiers (e.g., face or voice). This model efficiently gets rid of the intrinsic vulnerabilities and false-match rates connection to depending on a single biometric trait. In our work multimodal biometric recognition with face images and voice signals are employed with the objective of providing accurate and frictionless identity verification. By incorporating two different physiological and behavioral markers create a more robust security solution. In this work multimodal biometric authentication method employing face images and voice signals noted as, Multi Task Cascaded Pre-emphasis and Vision Attention Fusion Transformer (MTCP-VAFT) to improve the security of multimodal biometric recognition systems is presented in this section. Figure 1 shows the flow diagram of MTCP-VAFT method.

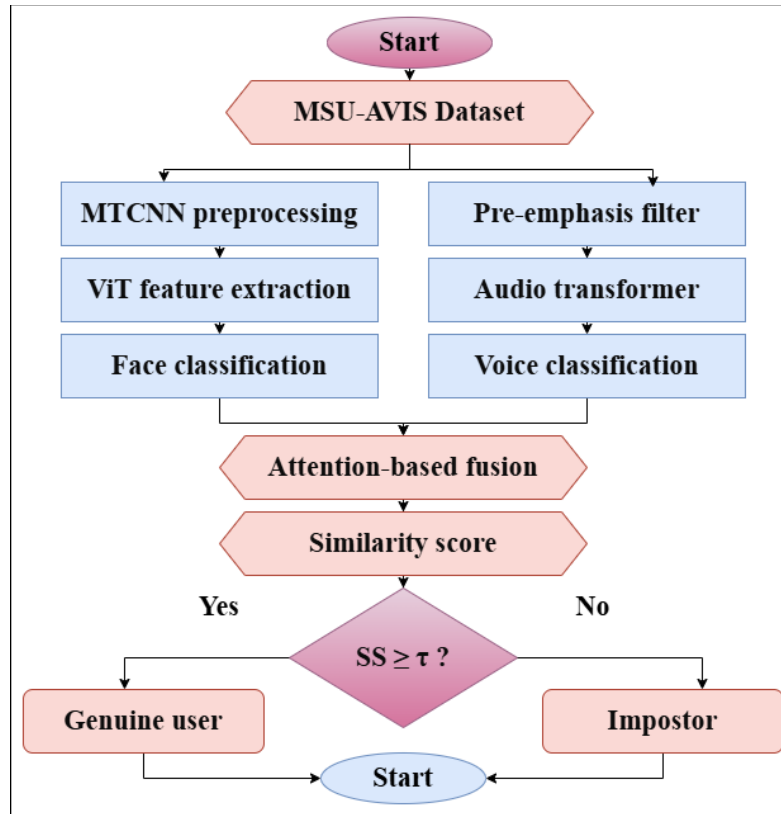


Figure 1 Flow diagram of MTCP-VAFT method

As shown in Figure 1, the proposed MTCP-VAFT method incorporates a Multi-Task Cascaded Convolutional Neural Network (MTCNN) and pre-emphasis filter-based preprocessing to optimize raw face and voice biometric data by removing noise and tampered artifacts, aligning the data, and enhancing discriminative features prior to the feature extraction process. This in turn prevents false biometric authentications. Followed by which to extract fine-grained local features and to capture hierarchical global associations within facial data and voice signals Vision Attention Fusion Transformer based multimodal biometric authentication is applied which is then forwarded to a classification head to perform efficient differentiation between genuine and impostor effectively. The elaborated description of the MTCP-

VAFT method is provided in the following subsections.

Multi Task Cascaded Neural and Pre-emphasis Filter-based Pre-processing

The purpose of pre-processing in biometric recognition is to clean and standardize raw sample prior to the feature extraction process. It isolates unique identifying characteristics and formats the data to make certain high accuracy and reliable biometric recognition. In our work Multi Task Cascaded Neural and Pre-emphasis Filter-based Pre-processing is applied for multimodal face and voice signals. Figure 2 shows the structure of Multi Task Cascaded Neural and Pre-emphasis Filter-based Pre-processing model.

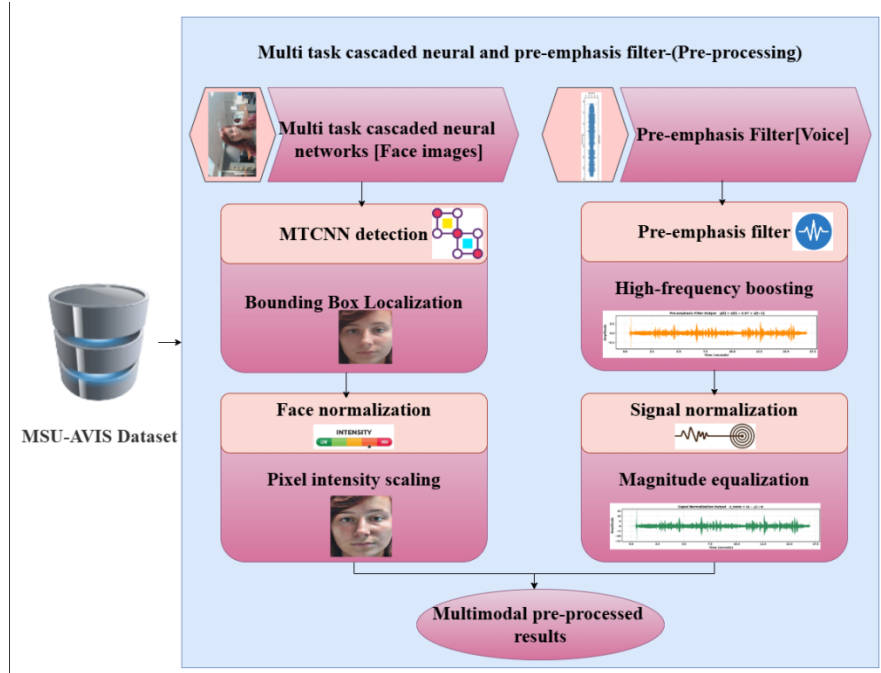


Figure 2 Structure of Multi Task Cascaded Neural and Pre-emphasis Filter-based Pre-processing

As illustrated in the figure 2, the sample face images and voice signals are initially obtained from MSU-AVIS Dataset. Here, to start with the face images are subjected to Multi Task Cascaded Neural Networks-based pre-processing. By applying this model standardizes face sample images so that the biometric recognition process accurately maps and compares facial features alone. Followed by which Pre-emphasis Filter-based Pre-processing is applied to the voice samples. By applying this filter discards unwanted background interference and ensures consistent format for the biometric recognition. In this manner by applying the pre-processing for sample face images and sample voice signals clean the noise before passing them to extraction phase.

Multi Task Cascaded Neural Networks-based pre-processing [face images]

The pre-processing step is initially done to make certain uniform and high-quality data for feature extraction. The two pre-processing steps performed in our work for face biometric recognition are face detection and normalization. Initially face regions in the MSU-AVIS Dataset are detected by means of Multi Task Cascaded Neural Networks (MTCNN). This face detection model not only localize face region by means of bounding boxes in face images but also optimizes the detection process by reducing the loss involved in bounding box identification

process. This is mathematically defined as given below.

$$L_{BB} = \sum_{i=1}^N IoU(GB_i, PB_i) \quad (1)$$

From the above equation (1) the localized bounding box ' L_{BB} ' of face sample images are obtained by applying intersection over union function ' IoU ' to ground truth bounding box ' GB_i ' and predicted bounding box ' PB ' respectively. Followed by which to fine tune pixel intensities and ensure consistence with the proposed method, normalization function is defined as given below.

$$PF = I_{Norm} = L_{BB} \left[\frac{FS(p,q) - \mu}{\sigma} \right] \quad (2)$$

From the above equation (2) ' $FS(p, q)$ ' represent the intensity of face sample pixel at position ' p, q ' with mean intensity of the sample pixel represented by ' μ ' and standard deviation of the sample pixel denoted by ' σ ' respectively. This normalized results ' I_{Norm} ' in turn makes certain that all sample input values are scaled, therefore enhancing the model convergence during training.

Pre-emphasis Filter-based Pre-processing [Voice]

Due to the presence of background noise in voice samples, pre-processing is considered as an essential step. These systems such as biometric recognition necessitate efficient feature extraction from voice files where the most of the spoken

component comprises of salient features connected to secured user authentication. This work utilized Pre-emphasis based filtering and signal normalization to reach its objective. Human speech holds substantially energy in low frequencies than in high frequencies. Pre-emphasis balances this out, therefore ensuring flattening of spectrum. In addition to boost phenome detection for biometric recognition high-frequency sounds contain crucial information that aid in differentiating between certain letters. With this objective of this to boost the strength of high-frequency audio samples while suppressing low-frequency audio samples. The proposed work employs pre-emphasis filtering. This is mathematically formulated as given below.

$$y(t) = VS(t) - \alpha * VS(t - 1) \quad (3)$$

From the above equation (3) the output of the voice sample ' $y(t)$ ' is obtained based on the input voice sample ' $VS(t)$ ', previous input voice sample ' $VS(t - 1)$ ' with respect to the filter coefficient ' α ' respectively. In addition to ensure that the voice signals are analogous in spite of any dissimilarity in magnitude, signal normalization is performed as given below.

$$PV = VS_{Norm_i} = \frac{VS_i - VS_i[\mu]}{VS_i[\sigma]} \quad (4)$$

From the above equation (4) the voice signal normalized result of the ' VS_{Norm_i} ' sample ' VS_i ' is arrived at by using the voice signals mean ' $VS_i[\mu]$ ' and standard deviation ' $VS_i[\sigma]$ ' respectively. The pseudo code representation of Multi Task Cascaded Neural and Pre-emphasis Filter-based Pre-processing is given below.

Algorithm 1 Multi Task Cascaded Neural and Pre-emphasis Filter-based Pre-processing

Input: Dataset ' DS ', face sample ' $FS = \{FS_1, FS_2, \dots, FS_M\}$ ', voice sample ' $VS = \{VS_1, VS_2, \dots, VS_N\}$ '
Output: high-fidelity pre-processed samples
Step 1: Initialize '$M = 50$', '$N = 50$' Step 2: Begin //Face Pre-processing Step 3: For each dataset 'DS' with face sample 'FS' //Multi Task Cascaded Neural Networks-based face region detection Step 4: Detect face region according to (1) //Normalization Step 5: Generate normalized face region according to (2) Step 6: Return pre-processed face image 'PF' Step 7: End for //Voice pre-processing Step 8: For each dataset 'DS' with voice sample 'VS' //Pre-emphasis based filtering Step 9: Perform filtering according to (3) //Signal Normalization Step 10: Generate normalized voice signals according to (4) Step 11: Return pre-processed voice signals 'PV' Step 12: End for Step 13: End

As given in the above algorithm with the face samples and video signals obtained from the MSR-AVIS dataset is subjected to pre-processing

separately. Initially Multi Task Cascaded Neural Network is applied to the face samples that by discarding background interference and refining

concentrate on the region of interest. This in turn generates high-resolution face images with improved peak signal to noise ratio. Followed by which by applying the Pre-emphasis Filter to the voice signals ensures that the voice authentication process depends on clear acoustic features instead of ambient audio noise. By incorporating these two functions separately not only geometrically aligns facial landmarks but also normalize spectral features therefore improving the overall PSNR.

Vision Attention Fusion Transformer-based multimodal biometric authentication

With the obtained pre-processed face images and voice signals are subjected to prominent feature extraction and classification. With the extracted face and voice features and classified results, attention fusion function is applied to ensure accurate and precise multimodal biometric authentication. Figure 3 shows the structure of Vision Attention Fusion Transformer based multimodal biometric authentication model.

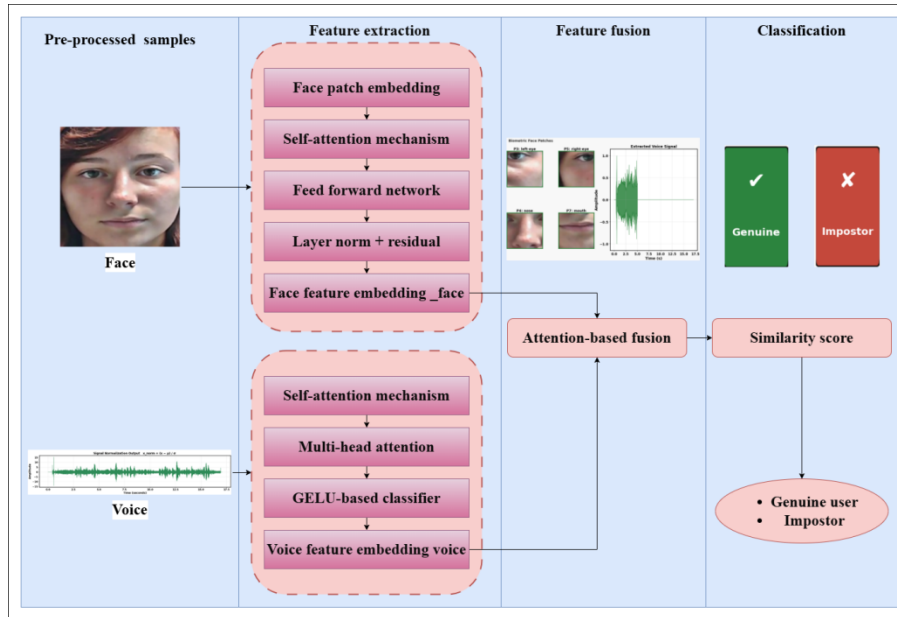


Figure 3 Structure of Vision Attention Fusion Transformer based multimodal biometric authentication model

As illustrated in the figure 3 the Vision Attention Fusion Transformer based multimodal biometric authentication model involving face and voice biometrics influences deep learning frameworks with the objective of processing and merging facial image features and vocal spectrograms. This model notably boosts security and multimodal biometric recognition accuracy by weighting the significance of each modality in an arbitrary fashion. Finally, the attention based fusion employed in our work can ensure secure and flexible multimodal biometric authentication pertinent for high security platforms.

The proposed feature extraction model employing Vision Transformer (ViT) applies a mechanism of self-attention to process pre-processed face images and voice signals. This is applied as sequences of patches for feature extraction. It is examined for its potential to capture global contextual relationships within facial images and video signals that is

paramount for understanding the subtle differences between genuine and impostor.

The essential of vision transformer is the patch embedding procedure when a pre-processed face image is initially transformed into a sequence of fixed size patches. Instead of providing pre-processed face image directly, the vision transformer employed in our work glances at an image $A \in \mathbb{R}^{H \times W \times C}$ as composed of non-overlapping patches $P * P$. These embeddings are then fed into transformer layers together with positional encodings and instead of using conventional convolution operations, the proposed model obtain spatial relationships of patches, forming a sequence as given in the below equation.

$$Seq_P = PF[flatten(A[i:i+P, j:j+P]) \forall (i, j) \in (0,0), \dots, (H-P, W-P)] \quad (5)$$

Also, by applying the above patch embedding results aid in obtaining the segmented regions of the

face (i.e. eyes, nose and mouth). From the above equation (5) results the embeddings are then projected into a lower dimensional space as given below.

$$Res = [Seq_p^1 TM; \dots; Seq_p^N TM] + Pos \quad (6)$$

From the above equation (6) ' TM ' denotes the trainable matrix with ' Pos ' denoting the positional embedding and finally ' N ' denoting the total number of patches. With this result the face spatial patches are mapped into a lower dimensional vector for further analysis.

Rather than considering all pixels same as the conventional convolution neural networks do, the self-attention mechanism of vision transformer employed in our work enables the method to evaluate the correlation between different facial features in an arbitrary fashion. This is performed by efficiently employing local and global information. This self-attention mechanism gives different weight to patches and pay focus to the most relevant patches when determining between genuine and impostor. For each patch embedding, three vectors are evaluated as given below.

$$Q_i = Res_i W_Q, K_i = Res_i W_K, V_i = Res_i W_V \quad (7)$$

From the above equation (7), ' W_Q ', ' W_K ' and ' W_V ' represents the trainable weight matrix for query, key and value respectively. Then the attention score for each vector is as given below.

$$Atten(Res_i, Res_j) = Softmax \left(\frac{Q_i K_j^T}{\sqrt{d}} \right) * V_j \quad (8)$$

From the above equation (8) the attention function employed ' $Atten$ ' normalizes the similarity score using ' d ' for facial feature dimensions. Followed by which the multi head self attention mechanism splits facial features in numerous subspaces with the intent of ascertaining distinct regions simultaneously. This is mathematically defined as given below.

$$MHSA(Res) = Concat \left(Atten(Res_i, Res_j), \dots, Atten(Res_M, Res_N) \right) \quad (9)$$

To increase representational capacity, non-linear variations in facial structures like angles or complex physiological features the proposed vision transformer is defined by adding a feed forward network (FFN). The FFN fine tune the features extracted from the self-attention mechanism. This in turn ensures high accuracy classification. The mathematical model is formulated as given below.

$$FFN(Res) = GELU(Res W_1 + B_1) W_2 + B_2 \quad (10)$$

From the above equation (10) the weight ' W_1 ' allows the FFN to isolate and analyze complex and subtle facial feature interactions (i.e. eye spacing) whereas using ' W_2 ' acts as a filter, retaining only the most significant extracted facial characteristics via bias ' B_1 ' and ' B_2 ' respectively.

Finally layer normalization is applied with the objective of making certain stable training and allows the method to preserve fine-grained facial features required for high accuracy biometric recognition. Also residual connection is used that by creating short cut straightforwardly flows backward during back propagation. This is mathematically defined as given below.

$$Res' = LN(Res + MHSA(Res)) \quad (11)$$

$$Res'' = LN(Res' + FFN(Res')) \quad (12)$$

From the above equations (11) and (12) by preventing the method from prevent the model from leaving indistinct features (i.e. skin texture) and retaining activation values within a harmonious scales, it makes certain high-dimensional face embeddings remain stable, resulting in more discriminative and reliable biometric recognition.

Finally, a classification token ' Res_{class} ' is introduced in the proposed biometric recognition. The classification token embeds the patch embedding sequence with the objective of learning a global representation of the entire face image. This classification token is then utilized for final classification, brining the proposed method to being able to correctly recognize genuine or impostor images. This is mathematically defined as given below.

$$Output = Softmax(Res_{class}, W_{class}) \quad (13)$$

The proposed method for extraction prominent features present in voice signals two core audio transformer modules, namely, self-attention and multi-head attention are used. First core module of audio transformer is the self-attention mechanism. The self-attention mechanism measures input representations by relating numerous positions of input sequences via, three inputs, namely, queries ' Q ', keys ' K ' and values ' V ' respectively. The result of a single query is evaluated as the aggregated sum values with each weight being ascertained via a query function that utilizes the related key. The self-attention mechanism based on dot product is defined as given below.

$$Attention(Q, K, V)[Q] = PV \left[softmax \left(\frac{Q K^T}{\sqrt{d}} \right) V[Q] \right] \quad (14)$$

From the above equation (14) segments of an audio signal are evaluated relative to all other segments with the objective of determining the most user-discriminative features. It assigns weight in an arbitrary fashion by allocation higher weight to unique phonetic or vocal traits. This in turn allows the method in isolating person's unique acoustic footprint. The second core module of audio transformer is the multi-head attention. This multi-head attention in voice signals process voice data parallel across distinct angles. Similar to self-attention mechanism, three linear projections are utilized for transforming the queries, keys, and values, respectively, into more discriminating representations. This is mathematically defined as given below.

$$MHSA(Q, K, V) = \text{Concat}(\text{Attention}[QW_i^Q, KW_i^K, VW_i^V]) \quad (15)$$

As given in the above equation (15) rather than using one set of attention weights, the multi head self attention mechanism splits the voice data stream into numerous parallel heads. Each head acts like an independent feature extractor concentrating on distinct voice features. Finally, Gaussian Error Linear Unit is applied as a classifier for voice extracted signals. The advantage of this function is its network's ability in capturing subtle acoustic nuances by weighting inputs in a stochastic function instead of applying a deterministic cut off. The Gaussian Error Linear Unit function employed for voice classification is defined as given below.

$$GELU(PV) = PV * \frac{1}{2} \left[1 + \text{ERF} \left(\frac{PV}{\sqrt{2}} \right) \right] \quad (16)$$

From the above equation (16) by introducing non-linearity Gaussian error function 'ERF' aids in capturing deep and intricate speech patterns accurately.

Attention-based Feature Fusion

The final fused embedding result is evaluated as the aggregated sum of the two modality-specific embeddings (i.e. face and voice signals). This is mathematically defined as given below.

$$Res_{fused} = \alpha \text{Output}(PF) + \beta \text{GELU}(PV) \quad (17)$$

With the above obtained results, the final step of the proposed multimodal biometric authentication system is responsible for converting fused embedding vector into a binary classification decision (i.e. genuine or imposter). Let ' Res_{fused} ' denote the output of the attention-based fusion layer. This fused results obtained distinguished information from both face images and voice signals. The similarity score result is generated as given below.

$$SS = \text{sigmoid} (W_{class} Res_{fused} + B_{class}) \quad (18)$$

From the above equation (18) ' W_{class} ' and ' B_{class} ' denotes the learnable training matrix results of the classification head for performing multimodal biometric recognition. The similarity score output ' SS ' denotes the methods confidence that the multimodal biometric input belongs to a genuine user. With the objective of transforming this probability distribution results into a decision making factor, the similarity score result ' SS ' is compared against a threshold ' τ '. Then, the output ' y ' is determined as given below.

$$y' = \begin{cases} 1, & \text{if } SS \geq \tau \text{ (genuine)} \\ 0, & \text{if } SS < \tau \text{ (impostor)} \end{cases} \quad (19)$$

The above obtained results ' y ' forms as the basis of multimodal biometric recognition system. By measuring the output based on the similarity score via decision logic in turn aids in improving overall accuracy and precision of multimodal biometric recognition in an extensive manner. The pseudo code representation of Vision Attention Fusion Transformer-based multimodal biometric authentication is given below.

Algorithm 2 Vision Attention Fusion Transformer-based multimodal biometric authentication

Input: Dataset ' DS ', face sample ' $FS = \{FS_1, FS_2, \dots, FS_M\}$ ', voice sample ' $VS = \{VS_1, VS_2, \dots, VS_N\}$ '
Output: Precise and accuracy multimodal biometric recognition
Step 1: Initialize '$M = 50$', '$N = 50$', pre-processed face sample image 'PF', pre-processed voice signals 'PV' Step 2: Begin Step 3: For each dataset 'DS' with pre-processed face sample image 'PF' //Face image extraction Step 4: Transform pre-processed face image into sequence of fixed size patches according to (5) Step 5: Project embeddings into a lower dimensional space according to (6)

```

Step 6: Generate three vectors separately for each patch embedding according to (7)
Step 7: Evaluate the attention score for each vector according to (8)
Step 8: Generate multi head attention results for determining different regions of facial features according to (9)
Step 9: Obtain non-linear variations in face features using FFN according to (10)
Step 10: Generate normalized face extracted results according to (11) and (12)
//Classification
Step 11: Generate face classified results according to (13)
Step 12: End for
Step 13: For each dataset 'DS' with pre-processed voice signals 'PV'
//Voice signal extraction
Step 14: Generate most discriminative voice features according to (14)
Step 15: Split voice data stream into numerous parallel heads according to (15)
//Classification
Step 16: Generate voice classified results according to (16)
Step 17: End for
//Feature fusion
Step 18: For each dataset 'DS' with pre-processed face sample image 'PF' and pre-processed voice signals 'PV'
Step 19: Generate the final fused embedding result according to (17)
Step 20: Evaluate the similarity score results according to (18)
Step 21: If 'y' = 1'
Step 22: Then multimodal biometric is genuine
Step 23: End if
Step 24: If 'y' = 0'
Step 25: Then multimodal biometric is impostor
Step 26: End if
Step 27: End for
Step 28: End

```

As given in the above algorithm the entire process of multimodal biometric authentication with pre-processed face images and pre-processed voice signals are split into three parts. They are feature extraction, classification and feature fusion. By incorporating facial and vocal biometrics utilizing sophisticated attention mechanisms process distinct inputs in a simultaneous fashion and relate them in shared embedding space. By processing both patches of a face and spectrograms of a voice via a shared Vision Transformer architecture, the proposed method learns deep cross-modal correlations. This in turn notably boosts authentication accuracy. In addition the crucial evolution lies in the multi-head attention layers. Rather than depending on hard coded rules, the

proposed method learns which pieces of the face associate to a specific phonetic trait in the voice. This collective characterization weights the significance of both modalities in an arbitrary fashion. The classified and fused results are then weighed using similarity score that in turn results in improved biometric recognition accuracy and precision.

CASE ANALYSIS AND INFERENCES

In this section detailed case analysis of multimodal biometric with face images and voice signals as input are detailed. To perform simulation MSU-AVIS Dataset was used. Though 2260 samples are present in the dataset, two samples are used for case analysis whereas 2260 samples were used for

performance evaluation. Figure 4 shows the face image and voice signals used for simulation.

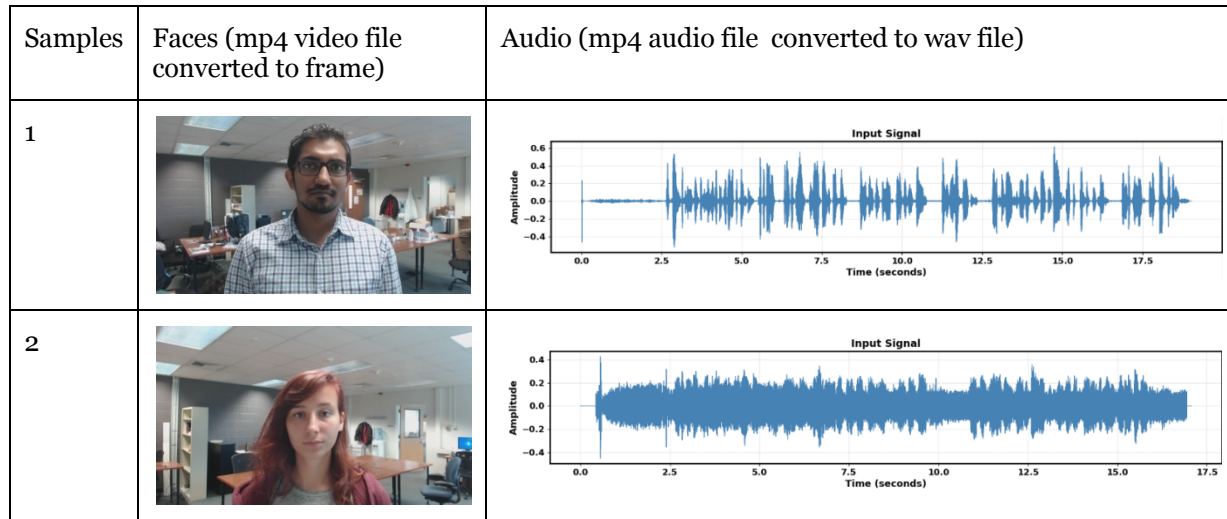


Figure 4 Face image and video samples

With the above two samples the entire process of multimodal biometric recognition is performed by applying the Multi Task Cascaded Pre-emphasis and Vision Attention Fusion Transformer (MTCP-

VAFT) method. Figure 5 shows the pre-processed results for the face image by applying multi task cascaded neural networks for not only detecting the face regions but also to perform normalization.

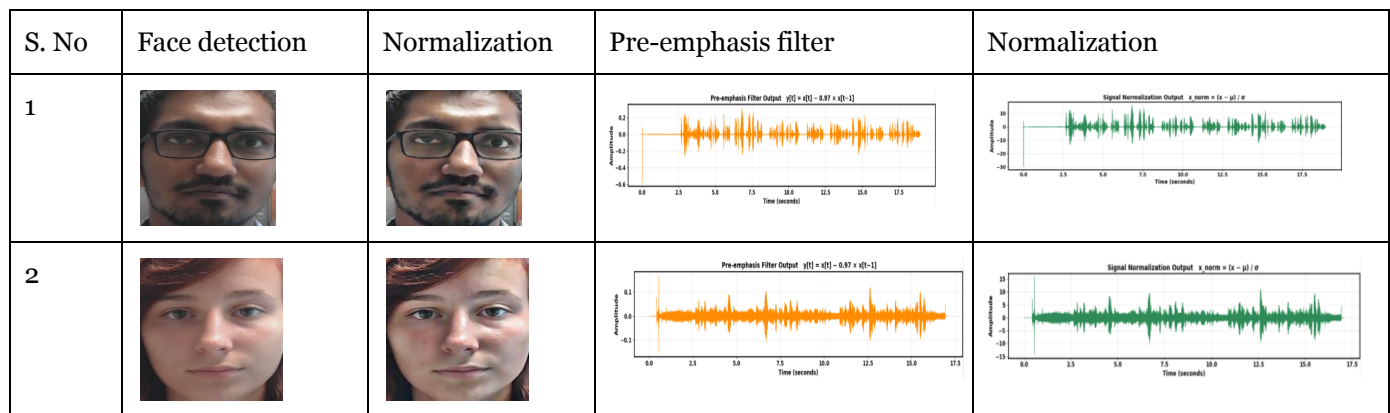


Figure 5 Face and voice based pre-processing

As provided in the above results pre-processing was applied separately to face images and voice signals. First to detect face and perform normalization function multi task cascaded neural network was applied. By applying this ensured precise detection of face regions. Moreover, by applying the pre-

emphasis filter for voice signals in turn aided in improving peak signal to noise ratio. Followed by which vision transformer was applied to extract facial features. Figure 6 shows the face extracted results.

Sample	Patch Results	Embedded	Facial feature extracted results	Voice feature extracted results

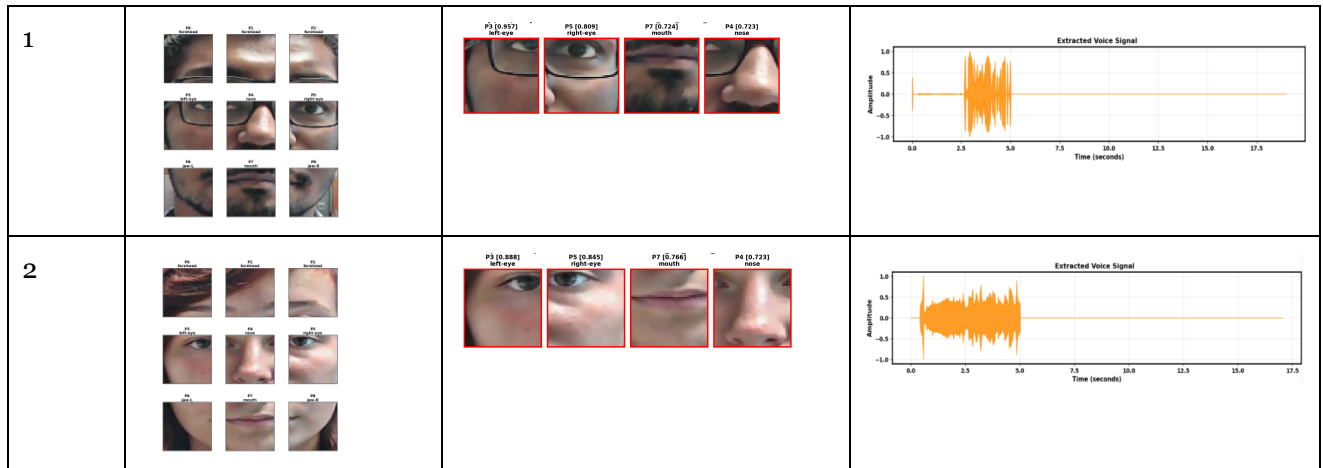


Figure 6 Facial feature extracted results

As provided in the above results by applying the self-attention and multi head attention mechanism of vision transformer in turn aids in extracting the prominent feature extracted results. Finally, the classified along with the multimodal biometric recognition results are provided in figure 7.

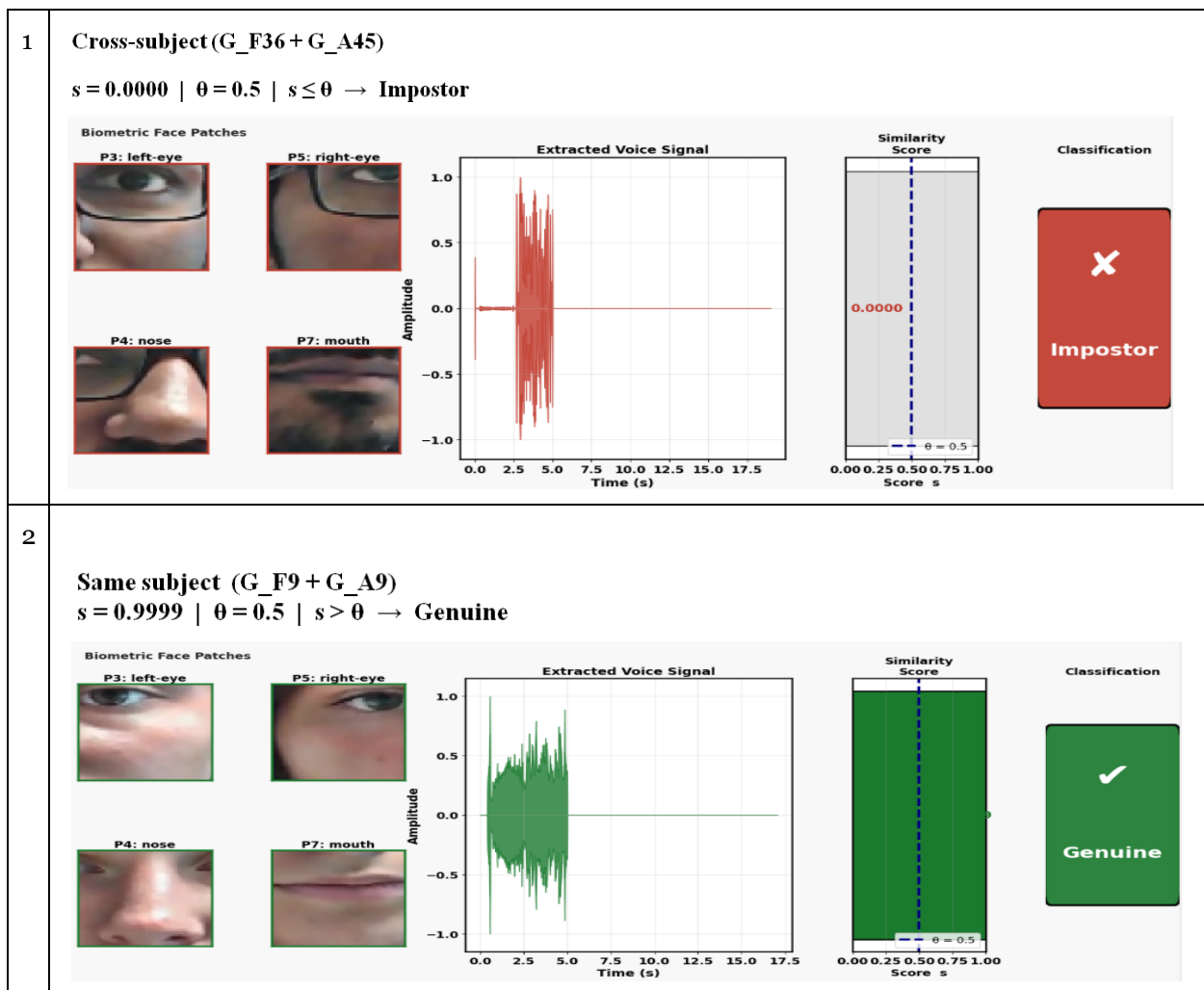


Figure 7 Multimodal biometric recognition results

As provided in the above results similarity score along with attention feature fusion was applied to measure and validate the multimodal biometric recognition results. An accurate and precise result was generated. The quantitative analysis of multimodal biometric recognition results are provided in the following sub-sections.

EXPERIMENTAL SETUP

In this section, performance measure of the proposed face images and video signals as multimodal biometric recognition method called, Multi Task Cascaded Pre-emphasis and Vision Attention Fusion Transformer (MTCP-VAFT) is designed. A detailed comparison was conducted with two existing methods, including the universal transformer for biometric matching proposed by Abdrakhmanova et al. (2024) and the fusion-enhanced hybrid multimodal biometric system integrating visible and infrared facial recognition proposed by Abdul-Al et al. (2026). The proposed MTCP-VAFT framework was implemented and validated using the Python high-level general-purpose programming language on the MSU-AVIS dataset. The experiments were conducted on an Intel Core i5- 6200U CPU @ 2.30GHz 4 cores with 4 Gigabytes of DDR4 RAM for the corresponding 100 samples of face data and voice signals. The software stack includes Python 3.11, along with PyTorch 2.0 for training and model development. To perform data manipulation and analysis Pandas 2.0.3 and NumPy 1.26.4 are used. Simulations are analyzed using five performance parameters, peak signal to noise ratio (PSNR), precision, recall, accuracy and F1-score. A fair comparison was conducted using identical samples from the MSU-

AVIS dataset across all three methods: the proposed MTCP-VAFT framework, the universal transformer for biometric matching proposed by Abdrakhmanova et al. (2024), and the fusion-enhanced hybrid multimodal biometric system proposed by Abdul-Al et al. (2026). All three methods were evaluated for multimodal biometric recognition using face images and voice signals as input modalities.

Performance analysis of Peak Signal-to-Noise Ratio

In multimodal biometric recognition, Peak Signal-to-Noise Ratio (PSNR) evaluates the image reconstruction accuracy. It is primarily utilized in measuring the recognition quality. It evaluates the ratio between maximum possible pixel value of an image (or spectrogram) and the power of corrupting noise. This is mathematically defined as given below.

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (20)$$

$$MSE = \frac{1}{m} \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^n [I(i, j) - K(i, j)]^2 \quad (21)$$

From the above equations (20) and (21) the PSNR value 'PSNR' is measured based on the maximum pixel value of image (or signals) 'MAX_I²' and the averaged squared difference between original image (or voice signals) 'I(i, j)' and the reconstructed image (or reconstructed voice signals) 'K(i, j)'. It is measured in terms of decibels (dB). Table 1 presents the PSNR results for face images and voice signals obtained using the proposed MTCP-VAFT framework and the methods proposed by Abdrakhmanova et al. (2024) and Abdul-Al et al. (2026).

Table 1 Performance analysis of PSNR for face image and voice signals

Samples	PSNR (dB) for face image			PSNR (dB) for voice signals		
	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM
226	48.25	42	36.35	51.35	47.55	44
452	51.35	46.35	41.25	56.75	48.35	43.25
678	55.25	40.15	35.35	61.35	53.25	48.55
904	58.45	43.25	38.55	65.25	57	52
1130	63.1	58.55	43.15	68.55	60.25	55.35
1356	68.35	63.25	58.35	72.35	64.35	69
1582	72.45	67.55	62.15	78.45	70.25	65.25
1808	75	70.25	65.85	82.55	74.55	69.15
2034	81.35	76.35	71.35	88.35	80.25	75.35
2260	85	80	75.25	91.45	83.55	78



Figure 8 PSNR versus samples for face images and voice signals

Figure 8 illustrates the graphical representations of PSNR for face images and voice signals by applying the proposed MTCP-VAFT and two existing methods, Hybrid MBS and ANN+fuzzy-NL-SVM. From the above figure with the samples ranging between 226 and 2260 involved in the simulation process, ten iterations were performed to analyze the PSNR for all the three methods. The simulation results indicate that the proposed MTCP-VAFT framework achieved higher PSNR values than the methods proposed by Abdrakhmanova et al. (2024) and Abdul-Al et al. (2026) for both face images and voice signals. Furthermore, the proposed MTCP-VAFT framework consistently attained superior PSNR values for voice signals compared with the existing methods. The reason behind the PSNR improvement using MTCP-VAFT method was due to the application of Multi Task Cascaded Neural and Pre-emphasis Filter-based Pre-processing. By applying this pre-processing model noise present in both the facial images and voice signals were eliminated via normalization. The combined preprocessing mechanism enhances high-frequency details while suppressing background noise, thereby improving the structural fidelity of face images and voice signals during multimodal biometric recognition. Consequently, the proposed MTCP-VAFT framework achieved a 12% improvement in PSNR over the method proposed by

Abdrakhmanova et al. (2024) and a 21% improvement over the method proposed by Abdul-Al et al. (2026) for face images. Similarly, for voice signals, the proposed framework improved the PSNR by 11% and 17% compared with the methods proposed by Abdrakhmanova et al. (2024) and Abdul-Al et al. (2026), respectively.

Performance analysis of Precision, Recall, Accuracy

Second in this section the performance analysis of precision, recall and accuracy is measured. This is mathematically defined as given below.

$$Pre = \frac{TP}{TP+FP} \quad (22)$$

$$Rec = \frac{TP}{TP+FN} \quad (23)$$

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (24)$$

From the above equations (22), (23) and (24), precision 'Pre', recall 'Rec' and accuracy 'Acc' is measured based on the true positive rate 'TP', false positive rate 'FP', true negative rate 'TN' and false negative rate 'FN' respectively. Table 2 and Table 3 given below lists the results of precision, recall and accuracy for the corresponding face image and voice signals using MTCP-VAFT, Hybrid MBS and ANN+fuzzy-NL-SVM.

Table 2 Performance analysis of precision, recall and accuracy for face image

Samples	Precision			Recall			Accuracy		
	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM
226	0.98	0.91	0.9	0.97	0.95	0.94	0.96	0.88	0.86
452	0.95	0.87	0.8	0.95	0.88	0.82	0.93	0.8	0.75
678	0.93	0.85	0.78	0.93	0.86	0.6	0.9	0.77	0.72
904	0.9	0.82	0.75	0.9	0.83	0.76	0.88	0.75	0.7
1130	0.88	0.8	0.73	0.88	0.81	0.74	0.85	0.72	0.67
1356	0.85	0.77	0.7	0.85	0.78	0.72	0.81	0.68	0.63
1582	0.83	0.75	0.67	0.82	0.75	0.67	0.78	0.65	0.6
1808	0.87	0.79	0.72	0.86	0.79	0.73	0.82	0.69	0.64

2034	0.9	0.82	0.74	0.88	0.81	0.74	0.85	0.72	0.67
2260	0.93	0.85	0.78	0.93	0.86	0.8	0.88	0.75	0.7

Table 3 Performance analysis of precision, recall and accuracy for voice signals

Samples	Precision			Recall			Accuracy		
	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM
226	0.98	0.92	0.9	0.98	0.95	0.94	0.97	0.89	0.86
452	0.95	0.87	0.77	0.96	0.88	0.8	0.95	0.85	0.8
678	0.94	0.86	0.76	0.95	0.87	0.79	0.93	0.83	0.78
904	0.92	0.84	0.74	0.93	0.85	0.77	0.91	0.81	0.76
1130	0.9	0.82	0.72	0.91	0.83	0.75	0.9	0.8	0.75
1356	0.88	0.8	0.7	0.89	0.81	0.73	0.88	0.78	0.73
1582	0.85	0.77	0.67	0.86	0.78	0.7	0.85	0.75	0.7
1808	0.87	0.79	0.69	0.9	0.82	0.74	0.89	0.79	0.74
2034	0.9	0.82	0.72	0.93	0.85	0.77	0.92	0.82	0.77
2260	0.93	0.85	0.75	0.95	0.87	0.79	0.94	0.84	0.79

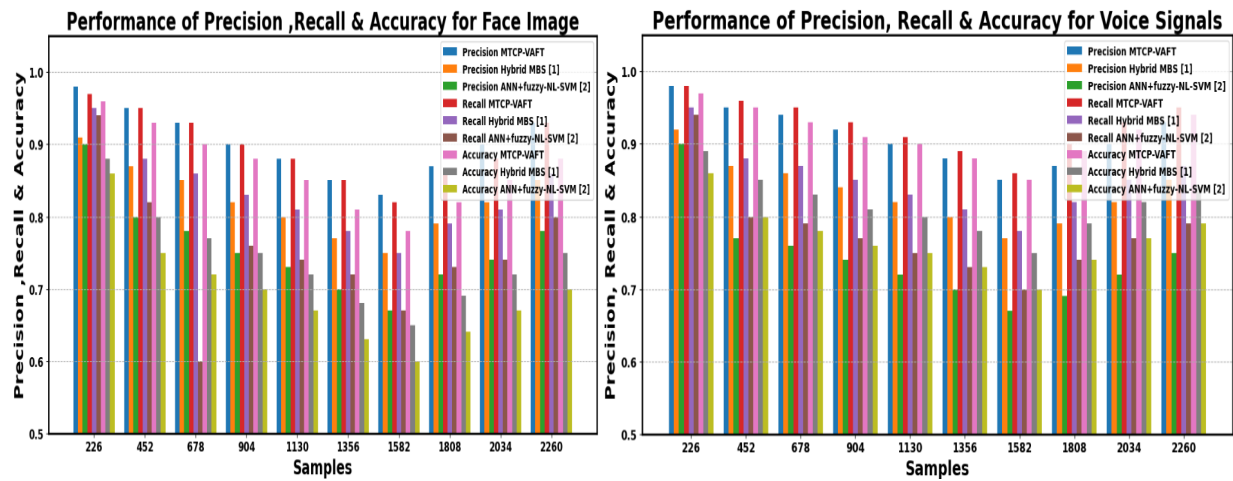
**Figure 9 Precision, recall and accuracy versus samples for face images and voice signals**

Figure 9 illustrates the graphical representations for precision, recall and accuracy using face images and voice signals as input. From the above two figures increasing the sample sizes ranging between 226 and 2260 did not have influence in the precision, recall and accuracy results. This is evident from the sample simulation. With simulations performed for the first seven iterations the results were less whereas for the rest of three iterations the results were high. This was also evident using both the samples, therefore corroborating the objective. The reason behind the improvement in precision, recall and accuracy was due to the application of Vision Attention Fusion Transformer based multimodal biometric authentication model. First, by concentrating on the most distinguishing regions of face images and voice signals as biometric trait via of Vision Attention Fusion Transformer it avoids depending on noisy or occluded data. The proposed MTCP-VAFT framework minimizes the false acceptance rate, thereby improving the precision of multimodal biometric recognition. Specifically, the

precision was improved by 9% and 16% over the methods proposed by Abdrakhmanova et al. (2024) and Abdul-Al et al. (2026), respectively, for face images, and by 9% and 19%, respectively, for voice signals.

Furthermore, instead of discarding an entire biometric sample when either the face or voice modality is compromised, the proposed cross-attention mechanism dynamically assigns greater importance to the unaffected modality. This adaptive weighting reduces the false rejection rate and consequently improves the recall of the proposed MTCP-VAFT framework by 7% and 16% over the methods proposed by Abdrakhmanova et al. (2024) and Abdul-Al et al. (2026), respectively, for face images, and by 8% and 16%, respectively, for voice signals.

In addition, rather than treating face images and voice signals as independent data streams, the proposed attention mechanism learns contextual dependencies between facial and speech

representations. This cross-modal alignment enables the framework to exploit complementary information from both modalities. Moreover, the self-attention and multi-head attention mechanisms suppress the influence of degraded biometric inputs by adaptively emphasizing the more reliable modality. For example, when noise degrades the voice signal, the Vision Transformer dynamically assigns greater attention to facial features to preserve recognition performance. Consequently, the proposed MTCP-VAFT framework improved the overall recognition accuracy by 15% and 20% over the methods proposed by Abdrakhmanova et al. (2024) and Abdul-Al et al. (2026), respectively, for face images, and by 11% and 16%, respectively, for voice signals.

Performance analysis of F1-score

Finally in this section to evaluate the efficiency of the classification involved in multimodal biometric recognition F1-score is measured. The F1-score result is mathematically defined as given below.

$$F1 - score = 2 * \frac{Pre * Rec}{Pre + Rec} \quad (25)$$

From the above equation (25) the results of 'F1 - score' is derived from the precision 'Pre' and recall 'Rec' rate. Finally, table 4 given below lists the results of F1-score for face images and voice signals using MTCP-VAFT, Hybrid MBS and ANN+fuzzy-NL-SVM.

Table 4 Performance analysis of F1-score for face images and voice signals

Samples	F1-score for face image			F1-score for voice signals		
	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM	MTCP-VAFT	Hybrid MBS	ANN+fuzzy-NL-SVM
226	0.97497436	0.92956989	0.91956522	0.98	0.934759358	0.91956522
452	0.95	0.87497143	0.80987654	0.9549738	0.874971429	0.78471338
678	0.93	0.85497076	0.67826087	0.9449735	0.864971098	0.77470968
904	0.9	0.8249697	0.75496689	0.924973	0.844970414	0.75470199
1130	0.88	0.80496894	0.73496599	0.9049724	0.824969697	0.73469388
1356	0.85	0.77496774	0.70985915	0.8849718	0.804968944	0.71468531
1582	0.8249697	0.75	0.67	0.8549708	0.774967742	0.68467153
1808	0.8649711	0.79	0.72496552	0.8847458	0.804720497	0.71412587
2034	0.88988764	0.81496933	0.74	0.9147541	0.834730539	0.74416107
2260	0.93	0.85497076	0.78987342	0.9398936	0.859883721	0.76948052

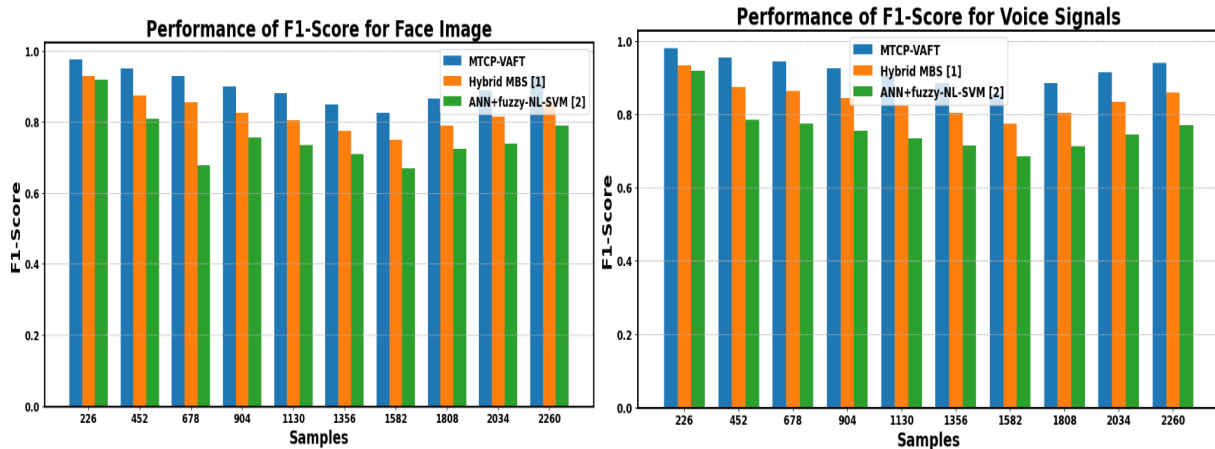


Figure 10 F1-score versus samples for face images and voice signals

Figure 10 presents the F1-score results obtained using the proposed MTCP-VAFT framework and the methods proposed by Abdrakhmanova et al. (2024) and Abdul-Al et al. (2026) for face images and voice signals. Ten iterations were conducted using sample sizes ranging from 226 to 2260 for all three methods. The results demonstrate that the proposed MTCP-VAFT framework consistently

achieved higher F1-scores than the existing methods.

The improved F1-score can be attributed to the use of the Gaussian Error Linear Unit (GELU) classifier for voice signal classification. Unlike conventional activation functions with hard thresholding, GELU weights the input according to a Gaussian

probability distribution, enabling smoother feature representation and improved robustness to noise in both face images and voice signals. In addition, the proposed attention-based feature fusion mechanism dynamically assigns higher weights to the most discriminative biometric modality, thereby enhancing multimodal feature representation. Consequently, the proposed MTCP-VAFT framework improved the F1-score by 8% and 26% over the methods proposed by Abdрахmanova et al. (2024) and Abdul-Al et al. (2026), respectively, for face images, and by 8% and 10%, respectively, for voice signals.

CONCLUSION

This paper presented a novel multimodal biometric authentication method using face images and voice signals called, Multi Task Cascaded Pre-emphasis and Vision Attention Fusion Transformer (MTCP-VAFT) to facilitate security. The proposed MTCP-VAFT method performs the tasks of pre-processing, feature extraction, classification and feature fusion for secure multimodal biometric recognition. With the face images and voice signals as input, first pre-processing is applied separately using Multi Task Cascaded Neural Networks-based face region detection and Pre-emphasis based filtering for voice signals. Followed by which prominent face and voice features are extracted using Vision Attention Fusion Transformer based multimodal biometric authentication model. Finally, attention-based feature fusion along with similarity score was measured to ensure authentication process in a precise and accurate manner. The algorithm is outlined to optimize five performance metrics, improving PSNR, precision, recall, accuracy and F1-score. The proposed MTCP-VAFT method is assessed on various numbers of simulation runs. The performance of the proposed MTCP-VAFT method is compared to the face and voice biometric methods such as Hybrid MBS and ANN+fuzzy-NL-SVM using different performance metrics. Experimental results prove that the proposed MTCP-VAFT method efficiently performs multimodal biometric recognition in a significant manner.

ACKNOWLEDGEMENT

The authors thank DST-FIST, Government of India, for funding towards Infrastructure facilities at St. Joseph's College (Autonomous), Tiruchirappalli-620 002.

Source of support: None

Conflict of interest: None

REFERENCES

- Abdрахmanova, M., Yermekova, A., Barko, Y., Ryspayev, V., Jumadildayev, M., & Varol, H. A. (2024). One model to rule them all: A universal transformer for biometric matching. *IEEE Access*, 12, 102637–102651. <https://doi.org/10.1109/ACCESS.2024.3426640>
- Abdul-Al, M., Kyeremeh, G. K., Qahwaji, R., Ali, N. T., & Abd-Alhameed, R. A. (2026). Fusion-enhanced hybrid multimodal biometric system: Integrating visible and infrared facial recognition for robust authentication. *IEEE Access*, 14.
- Alharbi, B., & Alshanbari, H. S. (2023). Face-voice based multimodal biometric authentication system via FaceNet and GMM. *PeerJ Computer Science*, 9, e1468. <https://doi.org/10.7717/peerj-cs.1468>
- Appati, J. K., Tsifokor, E., Amissah, D. K., & Adjepon-Yamoah, D. E. (2025). Vision Transformer-enhanced multi-descriptor approach for robust age-invariant face recognition. *Applied AI Letters*, 6(3), e70000. <https://doi.org/10.1002/ail2.70000>
- Byun, D.-M., Kim, S.-B., & Lee, S.-W. (2025). Hierarchical diffusion model for zero-shot singing voice synthesis with MIDI priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33, 2326–2336. <https://doi.org/10.1109/TASLPRO.2025.3577324>
- Chen, W., Sun, Y., Xu, K., & Dou, Y. (2024). Contrastive learning-based chaining-cluster for multilingual voice-face association. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)* (pp. 11348–11354). Association for Computing Machinery. <https://doi.org/10.1145/3664647.3688981>
- Chen, W., Zhu, B., Xu, K., Dou, Y., & Feng, D. (2024). VoiceStyle: Voice-based face generation via cross-modal prototype contrastive learning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(9), Article 279. <https://doi.org/10.1145/3671002>
- Cheng, H., Guo, Y., Wang, T., Li, Q., Chang, X., & Nie, L. (2024). Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(11), Article 354. <https://doi.org/10.1145/3698247>
- Dalila, C., Omar Badis, E. A., Saddek, B., & Naït-Ali, A. (2020). Feature level fusion of face and voice biometrics systems using artificial neural network for personal recognition. *Informatica*, 44(1), 85–96. <https://doi.org/10.31449/inf.v44i1.2596>

- Gong, L. Y., & Li, X. J. (2025). Deepfake voice detection: An approach using end-to-end transformer with acoustic feature fusion by cross-attention. *Electronics*, 14(10), Article 2040. <https://doi.org/10.3390/electronics14102040>
- Kefalas, T., Fotiadou, E., Georgopoulos, M., Ma, P., Petridis, S., Stafylakis, T., & Pantic, M. (2024). KAN-AV dataset for audio-visual face and speech analysis in the wild. *Image and Vision Computing*, 140, 104851. <https://doi.org/10.1016/j.imavis.2023.104851>
- Keresh, A., & Shamo, P. (2024). Liveness detection in computer vision: Transformer-based self-supervised learning for face anti-spoofing. *IEEE Access*, 12, 185673–185685. <https://doi.org/10.1109/ACCESS.2024.3513795>
- Khan, A. A., Jahangir, R., Alroobaea, R., Alyahyan, S. Y., Almulhi, A. H., Alsafyani, M., & Wechtaisong, C. (2023). An efficient text-independent speaker identification using feature fusion and transformer model. *Computers, Materials & Continua*, 75(2), 4085–4100. <https://doi.org/10.32604/cmc.2023.036797>
- Li, P., & Ren, J. (2026). CCFormer: A cascaded transformer framework for precise temporal audio-visual deepfake localization. *Alexandria Engineering Journal*, 134.
- Li, P., Yang, C., & Mao, L. (2025). The analysis of transformer end-to-end model in real-time interactive scene based on speech recognition technology. *Scientific Reports*, 15(1), Article 17950. <https://doi.org/10.1038/s41598-025-02904-0>
- Li, X., Wen, Y., Yang, M., Wang, J., Singh, R., & Raj, B. (2023). Rethinking voice-face correlation: A geometry view. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)* (pp. 2458–2467). Association for Computing Machinery. <https://doi.org/10.1145/3581783.3611779>
- Li, X., Xu, C., & Yagi, Y. (2026). OU-CA: A face-voice multimodal benchmark for continuous user authentication. *IEEE Transactions on Biometrics, Behavior, and Identity Science*. Advance online publication. <https://doi.org/10.1109/TBIOM.2026.3690283>
- Loubser, A., De Villiers, P., & De Freitas, A. (2024). End-to-end automated speech recognition using a character based small scale transformer architecture. *Expert Systems with Applications*, 252, 124119. <https://doi.org/10.1016/j.eswa.2024.124119>
- Mahum, R., Ganiyu, I., Hidri, L., El-Sherbeeney, A. M., & Hassan, H. (2025). A novel Swin transformer based framework for speech recognition for dysarthria. *Scientific Reports*, 15, Article 20070. <https://doi.org/10.1038/s41598-025-02042-7>
- Nosrati, L., Bidgoli, A. M., & Haj Seyyed Javadi, H. (2024). Machine learning and metaheuristic algorithms for voice based authentication: A mobile banking case study. *International Journal of Computational Intelligence Systems*, 17.
- Ren, Z., Yu, Z., Zhang, W., & Lei, Q. (2024). Streamlining industrial robot maintenance: An intelligent voice query approach for enhanced efficiency. *IEEE Access*, 12, 121864–121881. <https://doi.org/10.1109/ACCESS.2024.3452269>
- Smith, H. M. J., Ritchie, K. L., Baguley, T. S., & Lavan, N. (2025). Face and voice identity matching accuracy is not improved by multimodal identity information. *British Journal of Psychology*, 116(2), 367–385. <https://doi.org/10.1111/bjop.12757>
- Soumya, K., & Charles, P. J. (2026). A hybrid hierarchical deep learning framework with multi-level feature fusion for handwriting-based learning disability detection. *Indian Journal of Science and Technology*, 19(16), 1029–1038. <https://doi.org/10.17485/IJST/v19i16.432>
- Soumya, K., & Charles, P. J. (2026). A customized CNN-based framework for learning disability detection using handwriting image classification. *The Scientific Temper*, 17(4), 6015–6024. <https://doi.org/10.58414/SCIENTIFICTEMPE R.2026.17.4.11>
- Su, M.-H., Wang, W.-C., Hsu, Y.-M., Hou, S.-Y., Su-Jung, & Chang, S.-S. (2026). Voice-based structured nursing documentation using automatic speech recognition and large language models: Development and evaluation study. *JMIR Nursing*, 9.
- Tao, R., Shi, Z., Jiang, Y., Truong, D.-T., Chng, E.-S., Alioto, M., & Li, H. (2024). Multi-stage face-voice association learning with keynote speaker diarization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)* (pp. 11342–11347). Association for Computing Machinery. <https://doi.org/10.1145/3664647.3688980>
- Vatchala, S., Yogesh, C., Govindarajan, Y., Raja, M. K., Amirtha Ganesan, V. P., Arul, A. V., & Ramesh, D. (2025). Multi-modal biometric authentication: Leveraging shared layer architectures for enhanced security. *IEEE Access*, 13, 28029–28041. <https://doi.org/10.1109/ACCESS.2025.3534223>
- Wang, Y., Li, Z., Wang, J., Yang, L., Dong, B., Zhang, H., & Liu, J. (2025). MFF: A deep learning model

for multi-modal image fusion based on multiple filters. *IEEE Access*, 13, 38076–38090. <https://doi.org/10.1109/ACCESS.2025.3540007>

Wani, T. M., Qadri, S. A. A., Ashraf, A., & Amerini, I. (2025). ASTDT: An interpretable adaptive

spectro-temporal diffusion transformer for audio deepfake detection. *EURASIP Journal on Information Security*, 2025, Article 32. <https://doi.org/10.1186/s13635-025-00217-3>