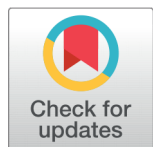


ORIGINAL ARTICLE



OPEN ACCESS

Received: 01/08/2026

Accepted: 13/08/2026

Published: 24/09/2026

Citation: Preethi S, Joseph Charles P (2026) Deep Learning with Additive Angular Spatio Temporal Vision Transformer for Multimodal Biometric Recognition Vision Attention Fusion Transformer. Indian Journal of Science and Technology 19(33): 2344-2357. <https://doi.org/10.17485/IJST/v19i33.1128>

* **Corresponding author.**

preethidiwakarmphil@hcctrichy.ac.in

Funding: None

Competing Interests: None

Copyright: © 2026 Preethi & Joseph Charles. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](https://www.indjst.org/))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Deep Learning with Additive Angular Spatio Temporal Vision Transformer for Multimodal Biometric Recognition Vision Attention Fusion Transformer

S Preethi^{1*}, P Joseph Charles²

¹ Research scholar, Department of computer science, Affiliated to Bharathidasan University, Tiruchirappalli-620002, Tamil Nadu, India

² Assistant Professor, Department of computer science, Affiliated to Bharathidasan University, Tiruchirappalli-620002, Tamil Nadu, India

Abstract

Objectives: To achieve audio-visual multimodal biometric recognition with higher accuracy. **Method:** Deep learning method that incorporates CNN and ViT into a unified framework, called, hybrid convolutional neural network with additive angular vision transformer (CNN-AAViT) introduced for audio-visual multimodal biometric recognition. The key contribution of the proposed method is the deep feature fusion layer that incorporates local spatial features and global context features in one shared feature space. This enables the implementation of face detection and audio recognition utilizing streamlined framework. **Findings:** results of CNN-AAViT improved in PSNR by 84dB, precision by 0.91, sensitivity by 0.90 and specificity by 0.80 and accuracy by 0.91 using face images. Proposed CNN-AAViT enhanced PSNR by 84dB, precision by 0.89, sensitivity by 0.89 and specificity by 0.87 and accuracy by 0.88 for voice signals when compared to existing methods techniques. **Novelty:** This work involves the integration of hybrid CNN-ViT architectures for both face and voice biometric modalities within a unified one-shot learning framework. Specifically, CNNs capture fine-grained spatial patterns such as facial features and micro-expressions, while ViTs model long-range dependencies in facial and spectrogram representations. The extracted multimodal features are projected into a shared joint embedding space, enabling direct face-voice association. Furthermore, the use of Additive Angular Margin Loss (AAM-Softmax) enhances inter-class separability and intra-class compactness using cosine similarity, thereby improving the reliability of multimodal matching, particularly when only limited samples are available.

Keywords: Convolutional Neural Network; Vision Transformer; Deep Learning; Spatial Features; Temporal Features; Additive Angular Margin

1 Introduction

Biometric recognition systems have become an essential component of modern security and identity management. Multimodal biometric recognition combines multiple traits, such as face, voice, iris, fingerprint, and ECG, to improve accuracy, security, and robustness over unimodal systems. However, existing approaches still face challenges related to noise, limited training samples, computational complexity, spatial–contextual feature extraction, and effective cross-modal representation.

A multimodal biometric recognition system using ECG signals was proposed in⁽¹⁾ to improve resistance to spoofing attacks; however, its robustness under noisy conditions was not sufficiently examined. The ANN+fuzzy-NL-SVM framework in⁽²⁾ improved authentication accuracy but provided limited emphasis on sensitivity and specificity. CNN- and RNN-based methods in⁽³⁾ enhanced biometric security, although accuracy and precision were not comprehensively evaluated. These studies demonstrate the need for recognition frameworks that jointly improve accuracy and robustness.

Vision Transformer-based multimodal confusion-state detection was investigated in⁽⁴⁾, but the method was limited in capturing long-range dependencies across multimodal information. A CNN-Transformer hybrid was introduced in⁽⁵⁾ to combine local and contextual feature extraction; however, it was not designed for face–voice association or one-shot authentication. Similarly, the CNN-Convolutional Vision Transformer framework in⁽⁶⁾ improved deepfake detection performance but focused on detecting manipulated content rather than cross-modal biometric matching. A pre-trained Swin Transformer was employed for iris recognition in⁽⁷⁾, while a Mamba Transformer was applied to audio-spectrum biometrics in⁽⁸⁾. Although these approaches demonstrate the effectiveness of transformers, they primarily address individual modalities rather than joint face–voice recognition.

Reviews of multimodal vision recognition⁽⁹⁾, audio classification⁽¹⁰⁾, and transformer-based identification⁽¹¹⁾ provide comprehensive insights into existing techniques but do not establish a unified solution for face–voice matching. A multi-scale kernel-based face network in⁽¹²⁾ extracted discriminative facial features but involved numerous hyperparameters, increasing the risk of overfitting. Convolutional patches integrated with Vision Transformers in⁽¹³⁾ improved classification accuracy, but the approach remained focused on visual data. Feature-level fusion of face and speech was investigated in⁽¹⁴⁾, although effective alignment of heterogeneous feature spaces remained challenging. Deep learning-based fusion of signature and facial information was presented in⁽¹⁵⁾, but it considered different biometric traits from the face–voice combination targeted in this study.

CNN-based biometric authentication in⁽¹⁶⁾ achieved a low false acceptance rate, but CNNs alone have limited capability for modeling long-range contextual relationships. The CNN-Swin Transformer framework with multi-head self-attention in⁽¹⁷⁾ improved spatial and contextual feature extraction; nevertheless, one-shot cross-modal face–voice matching was not addressed. A deep learning-based multimodal face–voice biometric system was presented in⁽¹⁸⁾, making it closely related to the proposed research. However, it does not fully exploit the complementary capabilities of CNN and Vision Transformer architectures or explicitly employ a shared embedding space with angular-margin optimization for one-shot matching. Other multimodal approaches combining fingerprint and finger-vein biometrics⁽¹⁹⁾ and angular-margin-based face recognition⁽²⁰⁾ demonstrate the effectiveness of multimodal fusion and discriminative loss functions, but they do not address joint face–voice embedding. Transformer-based biometric recognition with reduced computational cost was investigated in⁽²¹⁾, while multimodal recognition and fusion strategies were explored in⁽²²⁾ and⁽²³⁾. However, these approaches do not adequately address the combined requirements of efficient local–global feature extraction, cross-modal representation, and limited-sample matching.

Overall, existing studies have successfully explored individual aspects of biometric recognition, including spoofing resistance, CNN-based feature extraction, transformer-based contextual modeling, multimodal fusion, and angular-margin learning. Nevertheless, a significant gap remains in **unified one-shot face–voice recognition**, particularly in learning a discriminative common representation while simultaneously capturing local spatial and long-range contextual characteristics.

To address this gap, the proposed work develops a **hybrid CNN–Vision Transformer-based multimodal face–voice recognition framework**. CNNs extract fine-grained facial characteristics, including facial features and micro-expressions, while ViT-based processing of voice spectrograms captures long-range contextual dependencies in speech. The modality-specific representations are projected into a **shared joint embedding space** to enable direct face–voice identity association. In addition, **Additive Angular Margin Loss (AAM-Softmax)** enhances inter-class separability and intra-class compactness, while cosine similarity is used to determine the final matching score. The key novelty therefore lies in integrating **local CNN representation, transformer-based contextual modeling, shared cross-modal embedding, and angular-margin-based one-shot matching** within a unified face–voice biometric framework. This combination directly addresses the major limitations identified in existing studies and provides a more discriminative and robust solution for limited-sample multimodal biometric recognition.

1.1 Research gap

Research has been conducted on multimodal biometric recognition using machine learning, deep learning, CNNs, and ViTs. Existing studies have focused on improving recognition accuracy, security, and authentication performance; however, it failed to address the joint modelling of spatial, temporal, and contextual dependencies across heterogeneous biometric modalities. Conventional CNN-based approaches are effective in extracting local spatial features but have limited capability in capturing long-range dependencies, higher computational complexity. Furthermore, limited attention was given to hybrid CNN–ViT architectures for multimodal biometric recognition involving both audio and visual biometric information, for effectively modelling temporal relationships and cross-modal interactions. Therefore, there is a need for an effective hybrid deep learning framework. CNN and Vision Transformer mechanisms are employed to jointly capture local spatial features and long-range contextual dependencies for audio and visual modalities with improved recognition accuracy, precision, robustness, and security in multimodal biometric surveillance systems.

1.2 Novelty and Contributing remarks

The major novelty and contributions of the proposed work are as follows:

- Proposed hybrid CNN-AAViT is designed for audio-visual multimodal biometric recognition.
- Face and Voice Feature Vector Formalization and Pre-processing are designed to avoid differences in illumination and improve visibility in low light conditions. Discriminative regions are identified within facial frames and reliable representation of voice information.
- Hybrid CNN with Spatio Temporal ViT-based Feature Extraction is employed to capture fine-grained spatial and low-level local features from facial images and audio spectrograms using CNN, and Vision Transformer models temporal relationships and long-range contextual dependencies through self-attention, including the evolution of facial and audio patterns over sequential observations. This combined spatial and temporal representation is intended to improve the robustness of biometric feature extraction and enhance sensitivity and specificity.
- Additive Angular Margin Classification is utilized for multimodal biometric recognition that aids in maximizing inter-class distance and minimizing intra-class variations via geodesic distance therefore improving the overall recognition accuracy.
- Finally, the quantitative analysis results of performance evaluation represent that the packet delivery ratio in addition to end-to-end delay with energy consumption, throughput and network lifetime of the proposed CNN-AAViT method are better than existing methods.

2 Methodology

Multimodal biometric recognition objective is to enhance accuracy and fraud resistance by verifying a person's identity using two or more unique physiological features. Fusion data from numerous biometric inputs not only minimizes false matches but also reduces false rejections. This in turn makes it ideal for high security applications like banking, finance and secure facility access. A sophisticated deep learning method with hybrid convolutional neural network with additive angular vision transformer (CNN-AAViT) for audio-visual multimodal biometric recognition is proposed. The CNN-AAViT method for audio-visual multimodal biometric recognition is designed to incorporate the local, fine-grained spatial feature extraction of CNN with global, temporal and additive angle of ViT. This method generates highly accurate similarity embeddings by harnessing both the local feature extraction of CNNs and the global attention of ViTs. Figure 1 shows the architecture of CNN-AAViT for audio-visual multimodal biometric recognition.

2.1 Dataset Acquisition

The MSU-AVIS dataset utilized in our work for multimodal biometric recognition comprises of both audio and visual data. Dataset is gathered from https://iprobe.cse.msu.edu/dataset_detail.php?id=7&?title=MSU_AVIS_Dataset. The samples were collected from subjects while walking in a semi-constrained indoor environment. The subjects involved in the simulation mimics a real world surveillance scenario. The audio and visual data were obtained from 50 subjects. Among 50 subjects 16 were female and the rest of 34 subjects were male. The audio visual data were collected from both the camera and microphone. The data collected included 12 videos, each of 5 seconds duration. The video frames of each subjects' were obtained at a resolution of 1920 X 1080. In addition the audio were recorded at a sampling rate of 48kHz. The face images exhibit variations due to occlusions, indoor-illumination and accessories. The audio samples exhibit variations due to the indoor reverberations and background noise, etc.

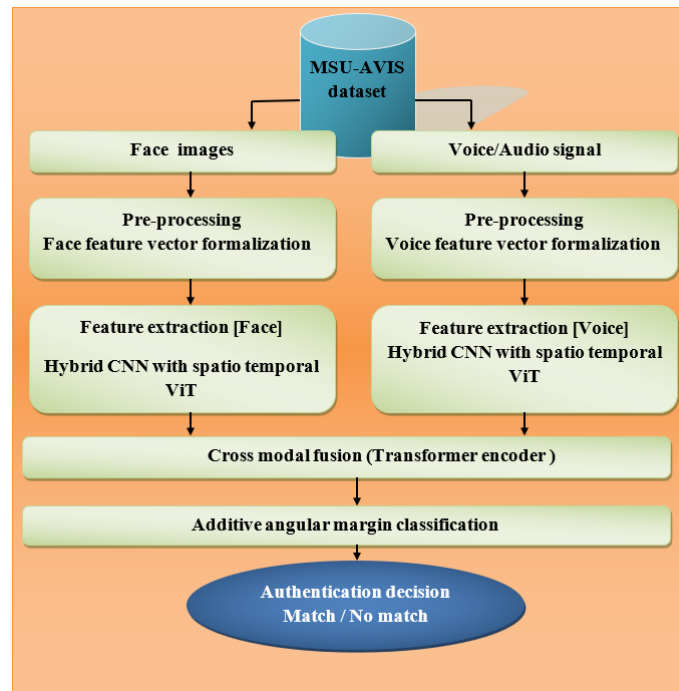


Fig 1. Architecture of CNN-AAViT for audio-visual multimodal biometric recognition

2.2 Face and Voice Feature Vector Formalization and Pre-processing

The MSU-AVIS dataset video stream comprises multiple image frames independently captured while walking around in a semi-constrained indoor environment mimicking a real world surveillance scenario. Face and voice pre-processing is performed to obtain consistent and discriminative feature representations from the two biometric modalities. For facial data, face detection, alignment, resizing, and pixel normalization are applied to reduce variations caused by pose and illumination, while CNN and ViT components capture local facial details and global dependencies. Similarly, voice signals are transformed into normalized spectrogram representations to capture relevant time-frequency patterns, with CNN extracting local spectro-temporal features and ViT learning their global relationships. The resulting fixed-length face and voice feature vectors provide reliable representations for subsequent multimodal feature fusion and biometric recognition.

2.2.1. Face feature vector formalization and Pre-processing

Let $S_f \subset \mathbb{R}^{m \times n \times 3}$ represent the visual data domain of RGB images with $m \times n \times 3$ size. Each visual data sample image S_f is pre-processed with the intent of detecting the bounding box and key points (i.e., two eyes, nose and two mouth corners) of the face. Followed by which affine geometric transformation and normalization is applied for face alignment to ensure consistency across different visual data samples. The visual data image is then resized and each pixel value is normalized in the range between '0' and '1'. The emerging transitional facial image, defined as $TS_f \subset \mathbb{R}^{m \times n \times 3}$ is utilized as input for our proposed method. These pre-processing steps strive to avoid differences in illumination, hence facilitating more accurate feature extraction and subsequent analysis. In the proposed method a precise feature extraction method that produces fixed-length representations in $D_f \subset \mathbb{R}^e$ is used. This stage is denoted as $D_{f\theta_f} : S_f \rightarrow D_f$ and its output is referred to as face feature vector. In case of facial images, CNNs extract fine-grained, localized spatial details (like skin texture and edges), while ViTs model obtain global and long-range dependencies across the entire face.

2.2.2. Voice feature vector formalization and Pre-processing

Let $S_v \subset \mathbb{R}^*$ represent the audio data domain of wave forms digitally denoted by a transitional visual acoustic representation $TS_v \subset \mathbb{R}^{k*}$, like, a spectrogram. These video frames in each sample undergo pre-processing before going to extraction process. Pre-processing addresses variations in the proportions and size of input frames with the objective to optimize model performance. The spectrogram is then created in a sliding window utilizing a Hamming window to improve visibility in low-

light conditions, producing an acoustic depiction ' TS_v ' that correspond to the audio ' S_v '. Average and standard deviation normalization is conducted on every spectrum frequency that in turn aids in ascertaining pivotal areas within frames. The emerging transitional audio representation is utilized as input of the acoustic modality in our proposed method. In the proposed method, a precise feature extraction model that generates fixed-length representations in ' $D_v \subset \mathbb{R}^e$ ' is used. This stage is represented as ' $D_{v\theta_v} : S_v \rightarrow D_v$ ' and its output is referred to as voice feature vector. On one hand using CNN fine-grained localized spectro temporal patterns are captured and on the other hand utilizing the ViT processes these features with the objective of identifying global dependencies across time and frequency.

2.3 Hybrid CNN with Spatio Temporal ViT-based Feature Extraction

The second step in Convolutional Neural Network (CNN)-based audio-visual data is feature extraction. The proposed method applied Hybrid Spatio Temporal Fusion CNN and ViT-based model. The Spatio Temporal Fusion-based Feature Extraction preserves contextual information of both audio and visual data and achieves synchronization. CNNs extract important feature maps from pre-processed audio-visual data based on a stack of convolution, normalization, activation, or pooling layers. Figure 2 shows the structure of Hybrid CNN with Spatio Temporal ViT-based Feature Extraction model.

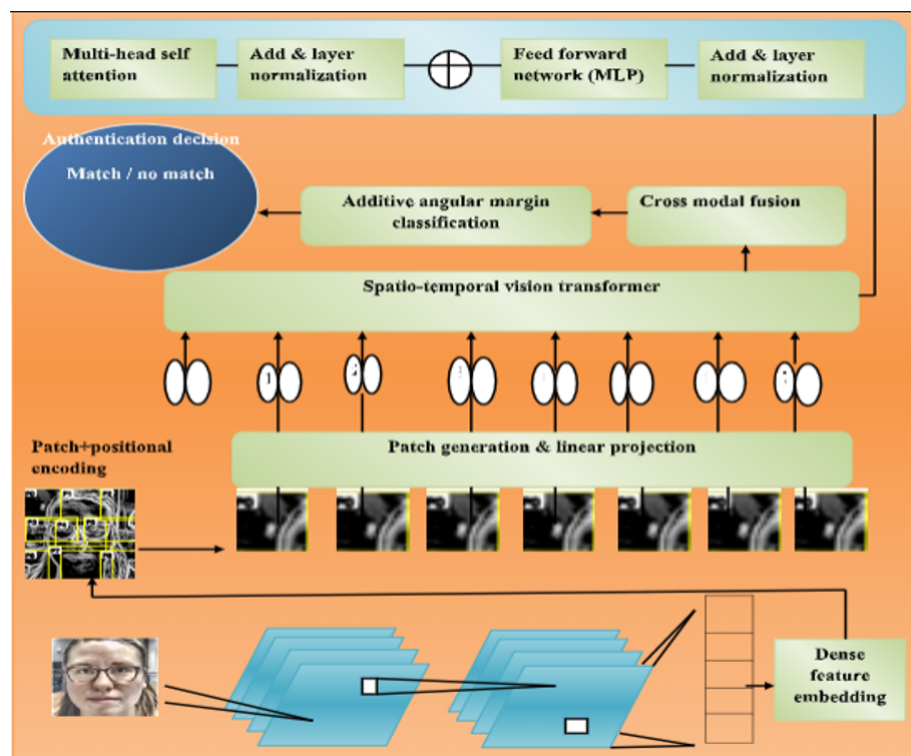


Fig 2. Structure of Hybrid CNN with Spatio Temporal ViT-based Feature Extraction model

As illustrated in the Figure 2, the feature extraction process is split into three parts. They are spatial and local feature extraction, temporal and global modeling and additive angular margin loss function. First, CNNs are applied to obtain spatial and low-level local details or facial features and spectrograms. Second, temporal along with global long-range sequence dependencies like how facial images moves over time via self-attention is modeled. The first convolution layer comprises of kernels that convolve over the pre-processed audio-visual data and kernel detects local feature like, edges. For audio biometric recognition process, the CNN capture localized fine-grained spectro temporal patterns. For visual biometric recognition process, the CNN capture localized fine-grained spatial details.

Hybrid CNN with Spatio-Temporal ViT is employed because CNN and ViT complement each other in extracting audio-visual biometric features. CNN effectively captures fine-grained local patterns, such as facial details and spectro-temporal characteristics of voice signals. The Spatio-Temporal ViT learns long-range relationships and global contextual information across spatial and temporal dimensions. Combining both models therefore provides richer and more discriminative feature

representations to improve the integration of face and voice information, and support more reliable multimodal biometric recognition.

Given an input image in ' $S \in \mathbb{R}^{C \times H \times W}$ ', the output of the convolution operation is mathematically represented as given below.

$$Conv_{ij}^{(k)} = \sum_{m=1}^M \sum_{n=1}^N \sum_{c=1}^C K_{mnc}^{(k)} * S_{i+m, j+n+c} + b^{(k)} \quad (1)$$

From the equation (1) the feature map post convolution ' $Conv_{ij}^{(k)}$ ', is generated based on the learnable kernel ' $K_{mnc}^{(k)}$ ' and the bias term of the ' $b^{(k)}$ ' feature map for the respective audio-visual sample data. The CNN processes different hierarchical representations, i.e., proportionately level features or edges of visual data in early layers, while deeper layers of the neural net provide high-level semantic features i.e. Mel Frequency Cepstral Coefficients. Pooling on the other hand allows for relocation stability of the CNN to aspect ratios (i.e. an audio can still be detected even it is slightly changes with respect to temporal or spatial structure) or the aspect ratios of faces. This is because by applying pooling aids in preserving whatever identity in spite of alignment variations. The pooling feature map result is defined as given below.

$$PF_{ij}^{(k)} = \max(m, n) Conv_{i+m, j+n}^{(k)} \quad (2)$$

From the equation (2) the pooled feature map results at position ' i, j ' are obtained ' $PF_{ij}^{(k)}$ ', by using the convoluted value ' $Conv$ ' at location ' $i+m, j+n$ '. Moreover the maximum function ' $\max$ ' selects the highest activation within the pooling window for corresponding audio-visual data. The feature maps from final convoluted output ' $Conv_{ij}^{(k)}$ ' and pooled feature map results ' $PF_{ij}^{(k)}$ ' are processed. These results are then flattened into a one dimension vector including a feature representation that are compressed but are extremely informative. For audio data the vector serves as the bridging gap connecting spatial audio features. For faces, the representation encodes distinguishing features associated to the participants. This is mathematically defined as given below.

$$P = Flatten \left(Conv_{ij}^{(k)}, PF_{ij}^{(k)} \right) \in \mathbb{R}^{1 \times D} \quad (3)$$

The one dimension vector of flattened features is provided as input via multiple dense layers. This is defined by associating to all of the activations from the previous layers. The dense layers obtain feature activation and merge local features into a global association as given below.

$$R = W * P + b \quad (4)$$

Finally with the above obtained results (4) the softmax function is applied to the fully connected layer for recognition purpose. This is mathematically defined as given below.

$$Prob(q=i|P) = \frac{e^{R_i}}{\sum_{j=1}^K e^{R_j}} \quad (5)$$

From the equation (5) ' $Prob$ ' represent the probability that the input ' P ' belongs to class ' i ', i.e. ' R_i ' respectively. The output layer provides a probability score for decision-making. With this the shallow or early layers detect edges and corners for visual data and acoustic edge for audio data. On the other hand, the mid level layers detect face parts ears, nose, eyes for visual data and phonetic segments, rhythm for audio data. Finally, the deeper or layer layers search for facial features that are unique in visual data and recognize complete words or environment sound for audio data. The flattened and embedded feature representations ' P ' are then processed to improve their potentialities as input tokens into Vision Transformer (ViT). By applying the flattened and embedded feature representations improves the method to provide global context modeling.

2.4 Additive Angular Margin-based Vision Transformer

The Additive Angular Margin-based Vision Transformer (ViT) splits up an input image (i.e. flattened and embedded feature representations ' P ') into patches of fixed sizes. This in turn adds positional information to the patches and extracts meaningful global features via transformer encoder blocks incorporating self-attention and MLP layers for multimodal biometric recognition. Additive Angular Margin-based Vision Transformer is employed to capture both local patch-level details and global dependencies from audio-visual biometric data. The patch-based representation and self-attention mechanism is employed to learn important relationships between facial and speech features. Also the Additive Angular Margin-based Vision Transformer significantly increases inter-class separability and intra-class compactness within the audio-visual data. Therefore, AAM-ViT provides a more robust and discriminative feature representation for accurate multimodal biometric recognition.

2.4.1. Patch Linear Embedding and Positional Encoding

The flattened and embedded feature representations ' P ' image is passed and then the patch extraction of the image is split into smaller square-shaped fixed-size patches of ' $16 * 16$ ' pixels. Moreover rather than processing these spectrograms pixel-by-pixel the images are split into square patches. That patch is flattened into a single dimension vector. This is mathematically defined as given below.

$$TP = \frac{H * W}{P_{size}^2} \quad (6)$$

From the equation (6) the total number of patches ' TP ' is obtained based on the height ' H ', weight ' W ' of the flattened and embedded ' P ' image with respect to the patch size ' P_{size}^2 ' respectively. Then the patch linear embedding result is obtained as given below.

$$EV_i^0 = W.P_i + b \quad (7)$$

From the equation (7) the patch linear embedding vector result ' EV_i^0 ' is obtained by using the flattened and embedded ' P_i ' image, learnable weight matrix ' W ' and bias ' b ' respectively. Followed by which a positional encoding vector is assigned to each patch embedding ' EV_i^0 '. The matching patch linear embedding is augmented with the position encoding as given below.

$$Out_{CNN} = EV_i^0 = EV_i^0 + P_i \quad (8)$$

$$EV^0 = EV_1^0 + P_1, \dots, EV_2^0 + P_2, \dots, EV_{TP}^0 + P_{TP} \quad (9)$$

From the equations (8) and (9) the augmented sequences are processed by the transformer encoder.

2.4.2. Cross Modal Fusion-based Transformer Encoder

The transformer encoder employed in our work uses the Cross Modal Fusion-based Self-Attention function instead of processing audio and visual data separately. With this the visual representations straightforwardly influence the audio representations. The Cross Modal Fusion-based Transformer Encoder employed in our work map dependencies across audio-visual data. In other words, the audio tokens act as queries while the visual tokens act as keys and values. This makes certain that the audio speech features fine tunes accurately with subsequent facial expressions. To encapsulate the relationships and dependencies in the input image patches, multi-head self-attention function is used. With the help of attention weights and assessing all tokens in correlation to another enables in capturing relevant patches. Each token here is projected linearly into a query ' Q ', key ' K ' and value ' V ' vectors utilizing learnable query weight matrix ' W_Q ', key weight matrix ' W_K ' and value weight matrix ' W_V ' respectively. Then the output projection matrix incorporating attention output weights is mathematically defined as given below.

$$Attention(Q, K, V) = softmax \left[EV_i^0 \left[\frac{QK^T}{\sqrt{Dim_K}} \right] \right] V \quad (10)$$

Followed by self-attention layer results according to (10) with dimensions ' Dim_K ' the feed forward neural network applied non linear transformations processes to each token. This is mathematically defined as given below.

$$FFNN(EV_i^0) = sigmoid \left[EV_i^0 (W_2.ReLU(W_1.EV_i^0 + b_1) + b_2) \right] \quad (11)$$

From the equation (11) the feed forward neural network is applied to each token ' $FFNN(EV_i^0)$ ' via weighted learning matrices ' W_1, W_2 ' and bias ' b_1, b_2 '. Finally the normalized results by stabilizing gradients and embeddings for heterogeneous features of audio and visual data are defined as given below.

$$Out_{ViT} = Norm(EV) = \frac{EV - \mu}{\sqrt{\sigma^2 + \epsilon}} * \gamma + \beta \quad (12)$$

From the equation (12) the normalized results by scaling heterogeneous features are arrived at by using the mean of input features ' μ ', variance of input features ' σ^2 ', via constant ' ϵ ', learnable scaling ' γ ' and shifting ' β ' factor respectively.

2.5 Additive Angular Margin Classification for multimodal biometric recognition

Each of the feature extracted outputs from the ‘ Out_{CNN} ’ and ‘ Out_{ViT} ’ captures distinct but interrelated features of visual depiction. Finally, additive angular margin is applied that by incorporating loss function aids in maximizing inter-class distance and minimizing intra-class variations. The additive angular margin into a shared hyper-spherical space matches audio-visual data group together, while differently separating from unrelated content. To summarize them into a common pool, all the feature maps are projected into a coarse dimensionality via a linear mapping for multimodal biometric recognition.

This is mathematically defined as given below.

$$Res_{CNN} = W_c Out_{CNN}; Res_{ViT} = W_v Out_{ViT} \quad (13)$$

Next based on the decision fusion representation, the features are normalized and concatenated together. The concatenated decision fused results are obtained as given below.

$$Res_{Fusion} = [Res_{CNN} || Res_{ViT}] \quad (14)$$

From the equation (14) ‘ $||$ ’ denote the concatenation operator, over multimodal features dimension ‘ Res_{CNN} ’ and ‘ Res_{ViT} ’ respectively. Finally to prevent zero padding issues across audio frames and visual sequences layer normalization function is applied to not only stabilize gradients but also to ensure the feature scaling remain manageable in an independent fashion. This is mathematically defined as given below.

$$Res_{Fusion} = LayerNorm \left(\sum_{i \in \{CNN, ViT\}} \frac{e^{w_i}}{\sum_j e^{w_j}} \right) \quad (15)$$

With the above resultant fusion function ‘ Res_{Fusion} ’ as given in equation (15), best balance between accuracy and computation efficiency with respect to multimodal biometric recognition is said to be ensured. Finally, Additive Angular Margin Loss function is applied to enforce deep neural networks to learn extremely different and concise groups of data. As far as audio-visual data is concerned, it is utilized in mapping speech and visual elements into a shared hyper-spherical space. With this function matching audio and visual data naturally cluster together while distinctly separating from unrelated samples. The Additive Angular Margin Loss function is defined as given below.

$$\mathcal{L}_{AAM} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\varphi \cdot (\cos(\theta_{Res_{Fusion_i}} + am))}}{e^{\varphi \cdot (\cos(\theta_{Res_{Fusion_i}} + am))} + \sum_{j \neq i} \theta_{Res_{Fusion_i}} e^{\varphi \cos \theta_{Res_{Fusion_j}}}} \quad (16)$$

With the generated results from equation (16) the Additive Angular Margin Loss applied in our work involving audio-visual data implements a geometric margin in the angular space between feature embeddings. The loss function by mapping both audio and visual data onto a hypersphere in turn generates extremely discriminative groups that enhance multimodal biometric recognition. The pseudo code representation of hybrid CNN-AAViT for audio-visual multimodal biometric recognition is given below.

Algorithm Hybrid convolutional neural network with additive angular vision transformer

Input: Dataset ‘ DS ’, face sample ‘ $S_f = \{S_{f1}, S_{f2}, \dots, S_{fM}\}$ ’, voice sample ‘ $S_v = \{S_{v1}, S_{v2}, \dots, S_{vN}\}$ ’

Output: Robust multimodal biometric recognition

- 1: **Initialize** ‘ $M = 50$ ’, ‘ $N = 50$ ’
- 2: **Begin**
- 3: **For** each dataset ‘ DS ’ with face sample ‘ S_f ’ and voice sample ‘ S_v ’
- //Pre-processing**
- 4: Formulate Face feature vector formalization and Pre-processing
- 5: Formulate Audio feature vector formalization and Pre-processing
- //Feature extraction**
- //Convolution layer**

```

6: Apply convolution function according to (1)
//Pooling layer
7: Generate pooled feature map results according to (2)
8: Flatten convoluted and pooled feature map into a one dimension vector according to (3)
9: Merge local features into a global association according to (4) and (5)
//Vision transformer
//Patch Linear Embedding and Positional Encoding
//Patchification
10: Patch global association feature into a single dimension vector according to (6)
//Patch Linear Embedding
11: Generate flattened and embedded feature representations according to (7)
//Positional encoding
12: Generate positional encoding vector to each patch embedding results according to (8) and (9)
//Cross Modal Fusion-based Transformer Encoder
13: Generate output projection matrix incorporating attention output weights according to (10)
//Feed forward neural network
14: Formulate per token processing using feed forward neural network according to (11)
15: Evaluate normalization function for heterogeneous features according to (12)
//Additive Angular Margin Classification for multimodal biometric recognition
16: Generate coarse dimensionality via a linear mapping for multimodal biometric recognition according to (13)
17: Generate resultant fusion results according to (14)
18: Generate resultant fusion normalized results according to (15)
19: Generate Additive Angular Margin Loss function results according to (16)
20: Return classified results
21: End for
22: End

```

Hyperparameters are implemented in Table 1.

Table 1. Hyperparameters details	
Hyperparameter	Value
Input image size	$224 \times 224 \times 3$
CNN filters	32
CNN kernel	3×3
Pooling	2×2 Max Pooling
CNN dropout	0.4
ViT embedding dimension	256
Patch size	16×16
Transformer blocks	6
Attention heads	8
MLP dimension	1024
Attention dropout	0.1
Temporal sequence length	16 frames
Fusion dimension	512
AAM margin	0.3
AAM scale	30
Optimizer	Adam
Learning rate	0.00001
Batch size	32
Epochs	100

Cross-validation is used for handling the class imbalance issue to compute the performance of the model ability to carry out unseen data. By using cross-validation, the dataset is divided into three sets such as training and testing and validation. Most data samples (70%) were used for training and the lesser data samples (20%) were taken for testing and remaining samples (10%) were taken for validation. In our work, the 10- fold cross-validation is utilized for measuring the results. To conduct and validate and ensure fair benchmark comparison on the MSU-AVIS dataset, proposed CNN-AAViT and conventional Multimodal BRS [1], ANN+fuzzy-NL-SVM [2] and CNNs and RNNs [3] are implemented in Python language. Several parameters such as PSNR, precision, sensitivity, specificity and accuracy are utilized. To conduct the experimental purpose, the samples in the range from 226-2260 are considered.

3 Results and Discussion

Comparison study is provided to analyze and validate the performance of four different learning methods, hybrid CNN-AAViT, Multimodal BRS [1], ANN+fuzzy-NL-SVM [2] and CNNs and RNNs [3] in solving the multimodal biometric recognition using MSU-AVIS dataset in Python.

3.1 Case Scenario 1: PSNR analysis

PSNR involved in audio-visual data is validated. The PSNR rate is evaluated as

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (17)$$

$$MSE = \frac{1}{m} \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^n [I(i, j) - K(i, j)]^2 \quad (18)$$

Where, peak signal-to noise ratio ‘*PSNR*’ is evaluated based on squared sum of the mean square error ‘*MSE*’ between original and reconstructed image.

Table 2. Results of PSNR for face image and voice signals

Samples	PSNR (dB) for face image				PSNR (dB) for voice signals			
	CNN-AAViT	Multimodal BRS [1]	ANN+fuzzy-NL-SVM [2]	CNNs and RNNs [3]	CNN-AAViT	Multimodal BRS [1]	ANN+fuzzy-NL-SVM [2]	CNNs and RNNs [3]
226	62.15	59.35	56.25	53.14	69.35	59.25	54.15	51.35
452	68.35	55.35	52.15	49.25	72.55	62.15	57.85	54.55
678	72.55	57.95	55.45	52.15	75.85	65.35	60.35	57.85
904	81.35	68.35	66.25	63.15	79.25	69.15	64.25	59.35
1130	85.89	70.98	68.15	65.25	83.35	73.25	68.15	63.25
1356	89.35	75.65	73.25	71.55	88.15	78.55	73.15	68.55
1582	92.15	78.56	76.15	73.25	90.45	80.25	75.55	70.25
1808	95.45	80.35	78.35	75.35	93.25	82.15	77.23	72.15
2034	96.35	82.45	80.55	77.15	94.55	84.35	79.25	74.55
2260	98.25	85.65	83.25	81.35	97	87	82	77

Table 2 shows PSNR. PSNR of CNN-AAViT of voice signals improved by 12%, 18% 23% and 15%, 12% and 22% for face images over [1], [2], [3]. When the experiment was conducted with 226 samples, PSNR was obtained as 62.12dB, 59.35dB, 56.25dB and 53.14dB using face images for CNN-AAViT, [1], [2] and [3]. Sensitivity was obtained as 69.35dB, 59.35dB, 54.15dB and 51.35dB using voice signals for the CNN-AAViT, [1], [2] and [3].

3.2 Case Scenario 2: Precision analysis

Precision concentrates on the method's positive prediction. It answers to the question out of all positive predictions, how many samples were actually correct.

$$Pre = \frac{TP}{TP + FP} \quad (19)$$

Where, 'Pre' is result of precision, 'TP' is true positive and 'FP' is false positive value.

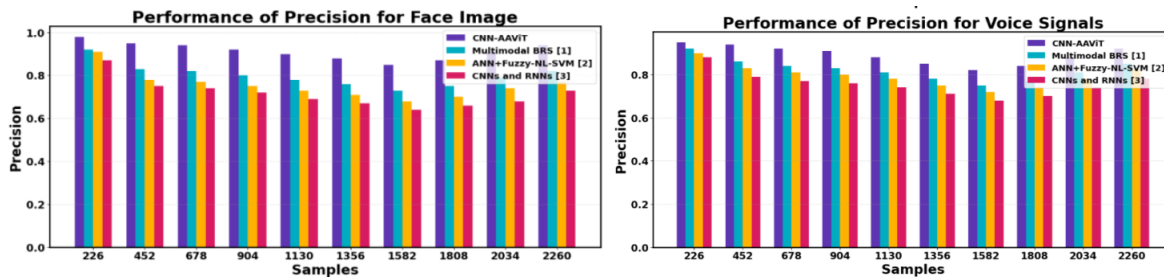


Fig 3. Precision versus Samples (using face image and voice signals)

Figure 3 shows precision. Precision of CNN-AAViT enhanced by 13%, 18% and 22% (for face images), 8%, 11% and 15% (for voice signals) than [1], [2], [3]. When the experiment was conducted with 226 samples, precision was obtained as 0.98, 0.92, 0.91 and 0.87 using face images for CNN-AAViT, [1], [2] and [3]. Precision was obtained as 0.95, 0.92, 0.9 and 0.88 using voice signals for the CNN-AAViT, [1], [2] and [3].

3.3 Case Scenario 3: Sensitivity analysis

Sensitivity or recall is utilized in measuring the performance of a classification method. It is defined as the ratio of actual positive samples that are identified correctly by the method.

$$Sen = \frac{TP}{TP + FN} \quad (20)$$

Where, 'Sen' is sensitivity. Table 2 lists sensitivity for face image and voice signals.

Table 3. Results of Sensitivity for face image and voice signals

Samples	Sensitivity for face image				Sensitivity for voice signals			
	CNN-AAViT	Multimodal BRS [1]	ANN+fuzzy-NL-SVM [2]	CNNs and RNNs [3]	CNN-AAViT	Multimodal BRS [1]	ANN+fuzzy-NL-SVM [2]	CNNs and RNNs [3]
226	0.98	0.97	0.95	0.94	0.95	0.9	0.85	0.82
452	0.95	0.85	0.8	0.77	0.93	0.85	0.8	0.78
678	0.93	0.83	0.78	0.75	0.92	0.8	0.75	0.73
904	0.91	0.81	0.76	0.73	0.91	0.78	0.73	0.71
1130	0.88	0.78	0.73	0.7	0.88	0.75	0.7	0.68
1356	0.85	0.75	0.7	0.67	0.85	0.71	0.66	0.64
1582	0.83	0.73	0.68	0.65	0.82	0.68	0.63	0.61
1808	0.84	0.74	0.69	0.66	0.85	0.72	0.67	0.65
2034	0.88	0.78	0.73	0.7	0.88	0.77	0.72	0.7
2260	0.92	0.82	0.77	0.74	0.93	0.83	0.8	0.78

Table 3 shows sensitivity for face image and voice signals. When the experiment was conducted with 226 samples, sensitivity was observed as 0.98, 0.97, 0.95 and 0.94 using face images for the CNN-AAViT method, [1], [2] and [3]. Sensitivity was observed as 0.95, 0.9, 0.85 and 0.82 using voice signals for the CNN-AAViT, [1], [2] and [3]. Sensitivity is enhanced using CNN-AAViT method by 10%, 16% and 19% (for face images) and 13%, 18% and 21% (for voice signals) over [1], [2], [3].

3.4 Case Scenario 4: Specificity 'Spe' analysis

It is a paramount performance metric, in the realm of multimodal biometric recognition tasks. It is referred to as the ratio of negative samples that were accurately identified as negative.

$$Spe = \frac{TN}{TN + FP} \quad (21)$$

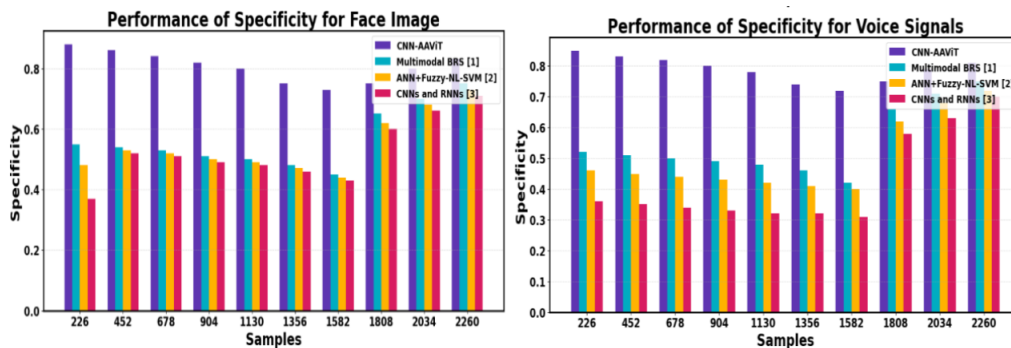


Fig 4. Specificity versus Samples (using face image and voice signals)

Figure 4 shows specificity. Specificity of CNN-AAViT method enhanced by 30%, 32% and 35% (for face images) and 30%, 36% and 46% (for voice signals) than [1], [2], [3]. Let us consider 226 samples, specificity was obtained as 0.88, 0.55, 0.48 and 0.37 using face images for CNN-AAViT, [1], [2] and [3]. Specificity was obtained as 0.85, 0.52, 0.46 and 0.36 using voice signals for the CNN-AAViT, [1], [2] and [3].

3.5 Case Scenario 5: Accuracy 'Acc' analysis

It is defined as ratio of correctly identified samples.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (22)$$

Where, 'TP, TN' is true positive and negative, 'FP, FN' is false positive and negative 'FN'.

Table 4. Results of Accuracy for face image and voice signals

Samples	Accuracy for face image				Accuracy for voice signals			
	CNN-AAViT	Multimodal BRS [1]	ANN+fuzzy-NL-SVM [2]	CNNs and RNNs [3]	CNN-AAViT	Multimodal BRS [1]	ANN+fuzzy-NL-SVM [2]	CNNs and RNNs [3]
226	0.97	0.9	0.88	0.84	0.94	0.9	0.85	0.83
452	0.95	0.93	0.83	0.78	0.93	0.95	0.9	0.87
678	0.94	0.82	0.72	0.67	0.91	0.83	0.88	0.85
904	0.92	0.8	0.7	0.65	0.88	0.8	0.75	0.72
1130	0.9	0.78	0.68	0.63	0.85	0.77	0.72	0.69
1356	0.87	0.75	0.65	0.6	0.83	0.75	0.7	0.67
1582	0.85	0.73	0.63	0.58	0.8	0.72	0.67	0.64
1808	0.89	0.77	0.67	0.62	0.82	0.74	0.69	0.66
2034	0.92	0.8	0.7	0.65	0.84	0.76	0.7	0.67
2260	0.94	0.82	0.72	0.67	0.86	0.78	0.73	0.7

Table 4 shows accuracy. Accuracy of CNN-AAViT method improved by 12%, 22% and 27% for facial images and 13%, 16% and 16% for voice signals than [1], [2], [3]. Let us consider 226 samples, accuracy was obtained as 0.97, 0.9, 0.88 and 0.84 using face images for CNN-AAViT, [1], [2] and [3]. Accuracy was obtained as 0.94, 0.9, 0.85 and 0.83 using voice signals for the CNN-AAViT, [1], [2] and [3].

3.6 Statistical Test

ANOVA (Analysis of Variance) is a statistical test employed to evaluate whether there is a major dissimilarity among multiple student data classes within a dataset. In the context of student performance prediction, ANOVA test is applied to assess whether the mean values of a dependent variable (accuracy) differ significantly across different groups defined by a categorical variable (face image and voice signals). Anova test is conducted for proposed LKMILBDC and the existing [1] and [2] and [3] using MSU-AVIS dataset based on accuracy value.

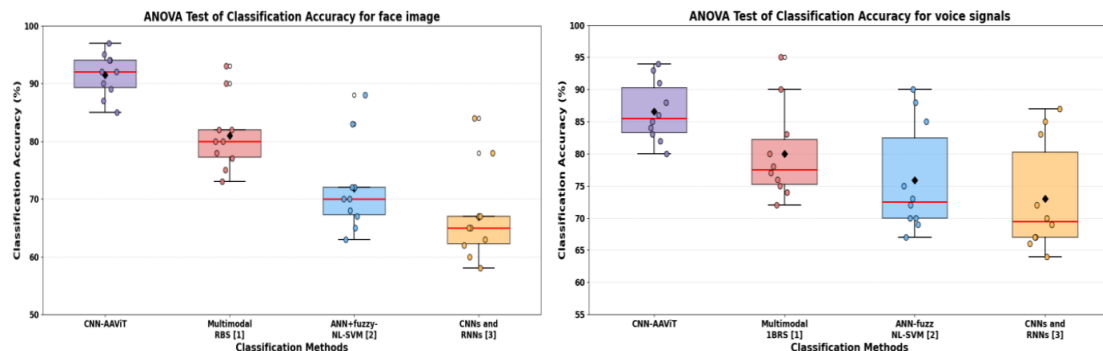


Fig 5. Paired t-tests Outcome

In Figure 5, paired t-tests outcome is described. CNN-AAViT method achieved statistically significant improvements over [1], [2] and [3] with all p-values below 0.05. The highest t-statistic was observed for indicating the strongest performance difference. The paired t-test results show that CNN-AAViT performed significantly better than all compared models.

4 Conclusion

Proposed hybrid CNN-AAViT is introduced for obtaining audio-visual multimodal biometric recognition with maximum accuracy. The integration of these complementary biometric modalities helps overcome the limitations associated with relying on a single biometric trait and consequently improves the reliability and security of biometric verification. Novelty of proposed method lies in the incorporation of CNN with spatio temporal ViT for multimodal biometric recognition. This combination along with the Additive Angular Margin Loss function allows the method to verify whether the biometric data originates from a single source, significantly enhancing security. Moreover, the study incorporates Spatio Temporal ViT-based Feature Extraction to enhance classes parability, followed by feature-level fusion and classification using an Additive Angular Margin classifier.

For face images, the quantitative results indicate of CNN-AAViT enhancement in PSNR by 84dB, precision by 0.91, sensitivity by 0.90 and specificity by 0.80 and accuracy by 0.91. For voice signals, proposed CNN-AAViT enhanced PSNR by 84dB, precision by 0.89, sensitivity by 0.89 and specificity by 0.87 and accuracy by 0.88 when compared to traditional baseline techniques. The proposed method has used in several applications in secure authentication, access control, identity verification, and other multimodal biometric security systems where improved robustness and recognition reliability are required. Nevertheless, proposed method considers only one dataset. Various metrics such as time and memory was not considered. In future work, the proposed work will be extended to apply adaptive multimodal fusion, and privacy-preserving biometric learning to improve computational efficiency and scalability. Also, the several metrics such as time and memory will be focused for enhancing the audio-visual multimodal biometric recognition.

5 Acknowledgements

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

Author details: S. Preethi is a Research Scholar; P. Joseph Charles is an Assistant Professor.

References

- 1) Garg R, Pathak P, Singh MP. A multimodal biometric recognition system based on Fingerprints, Iris and ECG via Swin Transformer and CNN Model. *Systems and Soft Computing*. 2025;7:1–12. <https://doi.org/10.1016/j.sasc.2025.200369>.

- 2) Nosrati L, Bidgoli AM, Javadi HHS. Machine Learning and Metaheuristic Algorithms for Voice-Based Authentication: A Mobile Banking Case Study. *International Journal of Computational Intelligence Systems*. 2024;17:1–21. <https://doi.org/10.1007/s44196-024-00690-7>.
- 3) Vatchala S, Yogesh C, Ganesan YGVPA, Raja MK, Arul AV, Ramesh D. Multi-Modal Biometric Authentication: Leveraging Shared Layer Architectures for Enhanced Security. *IEEE Access*. 2025;13:28029–28041. <https://doi.org/10.1109/ACCESS.2025.3534223>.
- 4) Ito Y, Ishizawa C, Kageyama Y. Multimodal confusion state detection in Japanese text reading using Vision Transformers, facial video, and eye tracking. *Displays*. 2026;91:1–15. <https://doi.org/10.1016/j.displa.2025.103267>.
- 5) Xu C, Yin X, Zhou Z, Liu B. ACNN-Transformer Hybrid Model for Real-Time Recognition of Affective Tactile Biosignals. *Computers, Materials and Continua*. 2026;87(1):1–12. <https://doi.org/10.32604/cmc.2026.074417>.
- 6) Soudy AH, Sayed O, Tag-Elser H, Raga R, Mohsen S, Mostafa T, et al. Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications*. 2024;32:19759–19775. <https://doi.org/10.1007/s00521-024-10181-7>.
- 7) Gao R, Bourla T. On Designing a SwinIris Transformer Based Iris Recognition System. *IEEE Access*. 2024;12:30723–30737. <https://doi.org/10.1109/ACCESS.2024.3369035>.
- 8) Kuang S, Shia J, van der Heijden K, Mehrkanoona S. BAST-Mamba: Binaural Audio Spectrogram Mamba Transformer for binaural sound localization. *Neurocomputing*. 2025;650:1–9. <https://doi.org/10.1016/j.neucom.2025.130804>.
- 9) Shafizadegan F, Naghsh-Nilchi AR, Shabaninia E. Multimodal vision-based human action recognition using deep learning: a review. *Artificial Intelligence Review*. 2024;57(7):1–85. <https://doi.org/10.1007/s10462-024-10730-5>.
- 10) Zaman K, Unok M, Sah M, Cemdirekolu. A Survey of Audio Classification Using Deep Learning. *IEEE Access*. 2023;11:106620–106649. <https://doi.org/10.1109/ACCESS.2023.3318015>.
- 11) Ünver B, Yoldar MT, Özkan S, Murakkoklu. Vision Transformers in Person Re-Identification: A Review. *IEEE Access*. 2026;14:92748–92764. <https://doi.org/10.1109/ACCESS.2026.3704666>.
- 12) Zhang P, Cui M, Ma Q, Li Y. MSKFaceNet: A Lightweight Face Recognition Neural Network for Low-Power Devices. *IEEE Access*. 2025;13:120533–120546. <https://doi.org/10.1109/ACCESS.2025.3584814>.
- 13) Indolia S, Nigam S, Singh R, Singh VK, Singh MK. Micro Expression Recognition Using Convolution Patch in Vision Transformer. *IEEE Access*. 2023;11:100495–100507. <https://doi.org/10.1109/ACCESS.2023.3314797>.
- 14) Jha K, Jain A, Srivastava S. Feature-level fusion of face and speech based multimodal biometric attendance system with liveness detection. *AIP Advances*. 2024;14:1–11. <https://doi.org/10.1063/5.0234430>.
- 15) Salturk S, Kahraman N. Deep learning-powered multimodal biometric authentication: integrating dynamic signatures and facial data for enhanced online security. *Neural Computing and Applications*. 2024;36(19):11311–11322. <https://doi.org/10.1007/s00521-024-09690-2>.
- 16) Gimba UA, Ariffin NABM, Udzir NI, Sani NFM. Enhancing biometric authentication through multimodal approach combining face and fingerprint recognition using convolutional neural networks (CNN). *Discover Computing*. 2025;28(246):1–15. <https://doi.org/10.21203/rs.3.rs-7048774/v1>.
- 17) Sharhrah S, Divya G, Hareesh KN, Prasad BR, Anusuya KP. Multimodal Biometric Authentication using Convolutional Neural Network, Swin Transformer, and Multi-Head Attention with Global Max Pooling. *ITM Web of Conferences*. 2025;79:1–7. <https://doi.org/10.1051/itmconf/20257901052>.
- 18) Li X, Xu C, Yagi Y. OU-CA: A Face-Voice Multimodal Benchmark for Continuous User Authentication. *IEEE Transactions on Biometrics*. 2015;14:1–20. <https://doi.org/10.1109/TBIOM.2026.3690283>.
- 19) Alshardan A, Kumar A, Alghamdi M, Maashi M, Alahmari S, Alharbi AAK, et al. Multimodal biometric identification: leveraging convolutional neural network (CNN) architectures and fusion techniques with fingerprint and finger vein data. *PeerJ Computer Science*. 2024;p. 1–42. <https://doi.org/10.7717/peerj-cs.2440>.
- 20) Han KU, Yun SJ, Song C, Kim KM, O CJ. An Improved Adaptive Angular Margin Loss Function for Deep Face Recognition. *American Journal of Neural Networks and Applications*. 2025;11(1):1–10. <https://doi.org/10.11648/j.ajnn.20251101.11>.
- 21) Hui Z. Research Review on Transformer Model in Computer Vision. *Proceedings of the 5th International Conference on Big Data, Information and Computer Network*. 2026;p. 643–648. <https://doi.org/10.1145/3801228.3801327>.
- 22) Garg R, Pathak P, Singh MP. A multimodal biometric recognition system based on fingerprints, iris and ECG via Swin Transformer and CNN model. *Systems and Soft Computing*. 2025;7:1–12. <https://doi.org/10.1016/j.sasc.2025.200369>.
- 23) Kumari CS, Nagapushpa KP, Jayalaxmi H, Asha CN, Harakannanavar SS, Jakati JS. Experimental analysis of face and iris biometric traits based on the fusion approach. *Indian Journal of Science and Technology*. 2025;16(31):2388–2397. <https://doi.org/10.17485/IJST/v16i31.314>.