Volume 4, Issue III, March, 2025, No. 63, pp. 802-813
Submitted 1/3/2025; Final peer review 30/3/2025
Online Publication 6/4/2025
Available Online at http://www.ijortacs.com

MULTIMODAL BIOMETRIC RECOGNITION SYSTEM USING DEEP CONVOLUTIONAL NEURAL NETWORK

[1*]Ekpo Michael E.

[1*]Department of Computer Science, Ebonyi State University, Abakaliki, Nigeria
Author's Email: [1*] ekpom550@gmail.com,
Corresponding Author's Email and Tel: ekpom550@gmail.com, +234 803 959 0555

Abstract

This study presents a Deep Convolutional Neural Network (Deep CNN) model for multimodal biometric recognition, integrating facial and gait features to enhance person identification, especially in scenarios with limited training data. The proposed system employs advanced preprocessing techniques, including the RetinaFace algorithm for robust Left-Right (LR) face detection and Gait Energy Image (GEI) extraction for effective gait representation. Feature extraction is performed using separate deep CNN-based extractors for facial and gait modalities. Subsequently, feature-level fusion is applied to combine the extracted features into a unified representation for classification.

The model was evaluated using two widely recognized datasets: CASIA-B and Extended Yale-B, encompassing biometric data from 25 individuals under diverse conditions. The proposed system achieved an average accuracy of 92.3%, precision of 91.4%, recall of 93.0%, and an F1-score of 92.2%, demonstrating high reliability and robustness. These results highlight the model's effectiveness in handling variations in body size, clothing, and environmental conditions, making it suitable for real-world applications such as identity verification, security surveillance, and behavioral monitoring. Overall, this work showcases the potential of deep learning-based multimodal biometric systems in improving the accuracy and dependability of automated human recognition technologies.

Keywords: Multimodal Biometric Recognition; Deep Convolutional Neural Network (Deep CNN); Transfer Learning; Gait Energy Image (GE); Left-Right (LR) Face

1. INTRODUCTION

People have been paying more and more attention to information security in recent years as science and technology have advanced (Amine, 2019). Once obtained, impersonation may readily replace traditional authentication techniques such utilising account numbers and passwords. One of the most promising security technologies of the twenty-first century is biometrics. One of the numerous benefits of biometrics over traditional identification is that the technology is not susceptible to loss, theft, or copying.

According to Bailey et al. (2014), biometrics is primarily a technique that uses quantifiable behavioural or physical biometrics to validate identification. Physical and behavioural characteristics are the two categories into which biological traits fall (Ra'Anan et al., 1991).While behavioural traits like gait and keystrokes are primarily learnt via habit, physical traits like fingerprints, retinas, and faces are primarily inborn (Ramirez-Mendoza et al., 2022).

A biometric identification system primarily consists of four steps: individual identification, feature value comparison, image processing feature extraction, and picture capture. Both single-modal and multimodal biometric systems are available. A single biometric characteristic is used to identify people in a single-modal biometric system. Despite the relative maturity of current biometric technologies, including face recognition (Amine, 2019), fingerprint recognition (Jomaa et al., 2020), and iris recognition (Miltra and Gofman, 2016), single biometric technologies, like fingerprint recognition, have faced significant challenges since the COVID-19 outbreak. Fingerprint recognition with protective gloves cannot be unlocked, and face recognition concealed behind masks fails. The deployment of single biometric technologies that use faces and fingerprints is becoming more challenging (Haider et al., 2023).

Multimodal biometric identification is becoming more and more common as biometric identification technology advances and improves. It can achieve the combination of face, fingerprint, vein, iris, voice print, and other biometrics through a meticulous design and fusion algorithm, which can result in complementary information and further increase the accuracy of recognition (Ammour et al., 2023). In order to improve the accuracy and security of the identification process, multimodal biometrics relate to the integration or fusion of numerous human biometrics, utilising each biometric's own advantages and integrating various feature fusion algorithms (Lowe, 2020).High identification accuracy, increased security, and a broader variety of applications are the benefits of multimodal biometric identification technology over single biometric identification technology (Wang et al., 2022).

Despite their widespread usage in biometric feature extraction from raw data and as recognition classifiers, machine learning (ML) techniques have several limitations when it comes to feature discrimination and selection across a range of domains. In order to extract low-level data to abstract-level features, Artificial Neural Networks (ANN) with several hidden layers have been used to create Deep Learning (DL), a new subcategory of machine learning. DL methods include distributed and parallel data processing, adaptive feature learning, resilient resilience, and dependable fault tolerance (Wang et al., 2021). Recently, biometric recognition systems have made use of deep convolutional neural networks, or deep-CNN (Boucherit et al., 2020). Significant issues with deep CNN models include the lengthy training period, the enormous volume of data, and the need for costly and potent GPUs to meet processing demands. However, by applying the learnt model to new tasks, Transfer Learning (TL) may be able to address these issues.

Additionally, it is a machine learning technique that may apply learnt characteristics

from one job to another. When training a full model from scratch with a modest quantity of data, TL performs well. Additionally, it drastically cuts down on training time and expense, which improves performance on related tasks.

The external environment in real-world applications and the constraints of single-mode biometrics itself affect single-modal biometric recognition technology, notwithstanding its relative maturity. This reduces the accuracy of identity detection and drastically limits the spectrum of applications for single-modal biometrics (Wang et al., 2022). Age and mask wearing will affect facial recognition accuracy. To overcome these challenges, this study proposed a deep Convolutional Neural Network (deep CNN) model that use transfer learning to recognise a person based on a little amount of training data and two biometric traits (face and gait).

2. RESEARCH METHOD

The key stages of the proposed multimodal recognition system include biometric data collecting and pre-processing, feature extraction, feature fusion, and classification components. The entire suggested multimodal system is shown in Figure 1. In the first step, a security camera's walking person frames are obtained, and then images of human silhouettes are extracted using background removal. A person can be seen strolling in various directions in several video sequences. The single sequence of the correctly aligned gait silhouette is averaged to provide a Gait Energy (GE) image. The proposed method concurrently extracts the face area as an input image of the left-right (LR) face and identifies it in a video frame. The suggested deep CNN model then gathers facial and gait characteristics and concatenates them at the feature level fusion. For multimodal recognition, the unified feature vector is sent through the classification phase.

2.1 Data Collection

The proposed approach evaluated the walking category of the human activities video collection (Schuldt et al., 2004). The proposed network was trained using the publically available CASIA-B gait dataset (Yu et al., 2006) for gait recognition and the Extended Yale-B (Georghiades et al., 2001) for face recognition. The walking video dataset includes 25 individuals in four different situations: (d1) outside, (d2) with scale variation, (d3) outside with different types of apparel, and (d4) indoors, as shown in Figure 2. A total of 100 video clips with 25 people are included.

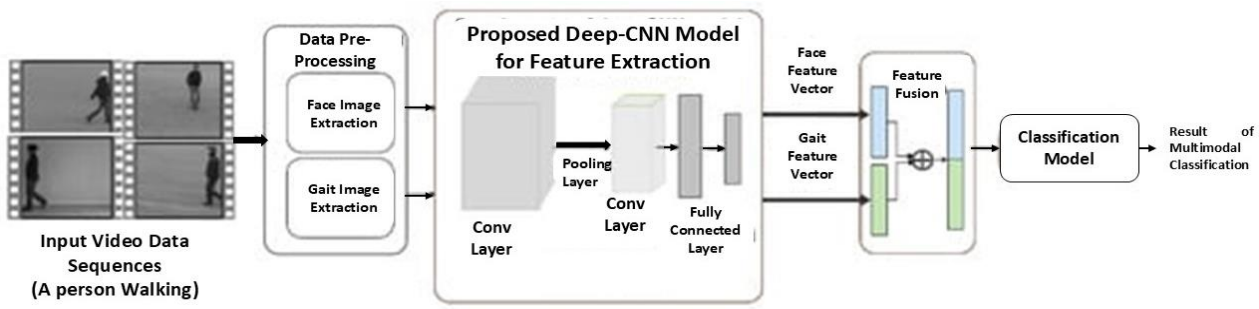

Figure 1: The proposed multimodal biometric recognition architecture

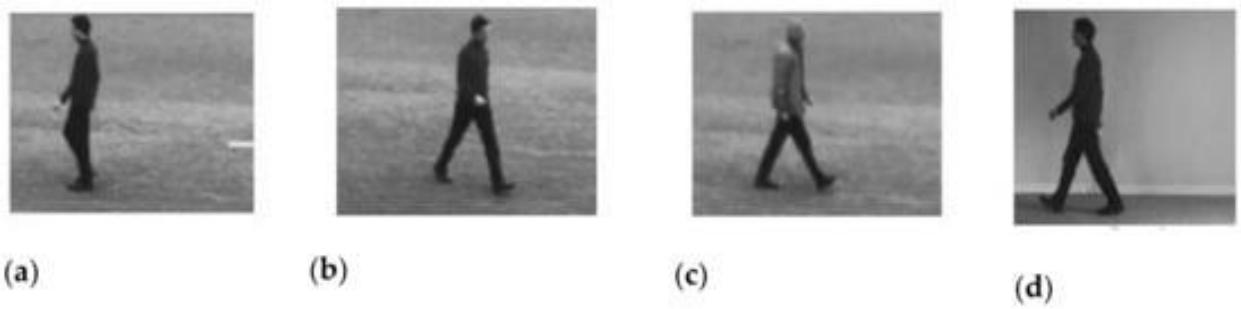

Figure 2: Sample of input videos of a person in four scenarios. (a) Outdoors (d1), (b) Outdoors with scale variation (d2), (c) Outdoors with different clothes (d3), (d) Indoors (d4).

As seen in Figure 6, which represents 25 patients with 45 photos each, the study also included LR facial region images. By identifying foreground items in 100 video sequences that match to 25 people, the human silhouette pictures are recovered. For every individual in a scenario, the suggested approach retrieved four sequences of silhouette photos. As seen in Figure 3, a typical GE picture uses a range of 20 to 25 silhouette frames.

Figure 3: Samples of L-R face region images from video sequences of a person

2.2 Data Pre-Processing through the Extraction of Left Right (L-R) Face and Gait Energy Image

In the first pre-processing stage, a person's face is detected in order to obtain an LR face picture from the input frames. The face is detected using the RetinaFace (Deng et al., 2019), a state-of-the-art deep learning-based facial detector. Face positions and scales on feature pyramids are sampled in this reliable single-stage pixel-wise facial localisation technique. The pre-trained models for face identification are the two trained models, ResNet 50 with a 30M size and MobileNet 0.25 with a 1.7M model size. In addition to the five facial landmarks (right eye, left eye, nose tip, right mouth, and left mouth) on the face, they also predict the face score and face area by bounding box. On the input videos, the face detection accuracy is around 87%. The technique extracts the LR face area pictures from the frames based on the face detection result.

The suggested method extracts and recognises gait features by representing a human walking gesture sequence in a single picture using a GE averaging process. The GE image is less

susceptible to noise in silhouette images and can preserve the original gait sequence data. Foreground moving object detection is the initial stage in GE picture extraction. To detect foreground items from the background model, the simplified Self-Organised Background Subtraction (simplified SOBS) approach (Maddalena and Petrosino, 2008) is used. The first backdrop model is created using the median filter with successive frames. Equation (1) illustrates how the Euclidean distance (Selvarasu et al., 2010) determines the best match by calculating the least distance between the input pixel and the current backdrop model using the image's HSV hexagonal colour space (h, s, v). The criteria automatically established by Otsu's approach must not exceed the necessary distance value. Other pixels are regarded as the foreground object component, whereas the founded best match is defined as a background pixel (Aung et al., 2022).

$$d(b, I(x,y)) = \sqrt{\begin{matrix}(v_b s_b\, cos(h_b) - (v_l s_l\, cos(h_l))^2 + \\ (v_b s_b sin(h_b) - (v_l s_l sin(h_l))^2 + (v_b - v_l)^2\end{matrix}} \quad (1)$$

Where l(x, y) is the intensity of the non-background pixel at position (x, y) and b is the intensity of the background pixel. Figure 4 displays the proposed system's detailed flowchart.

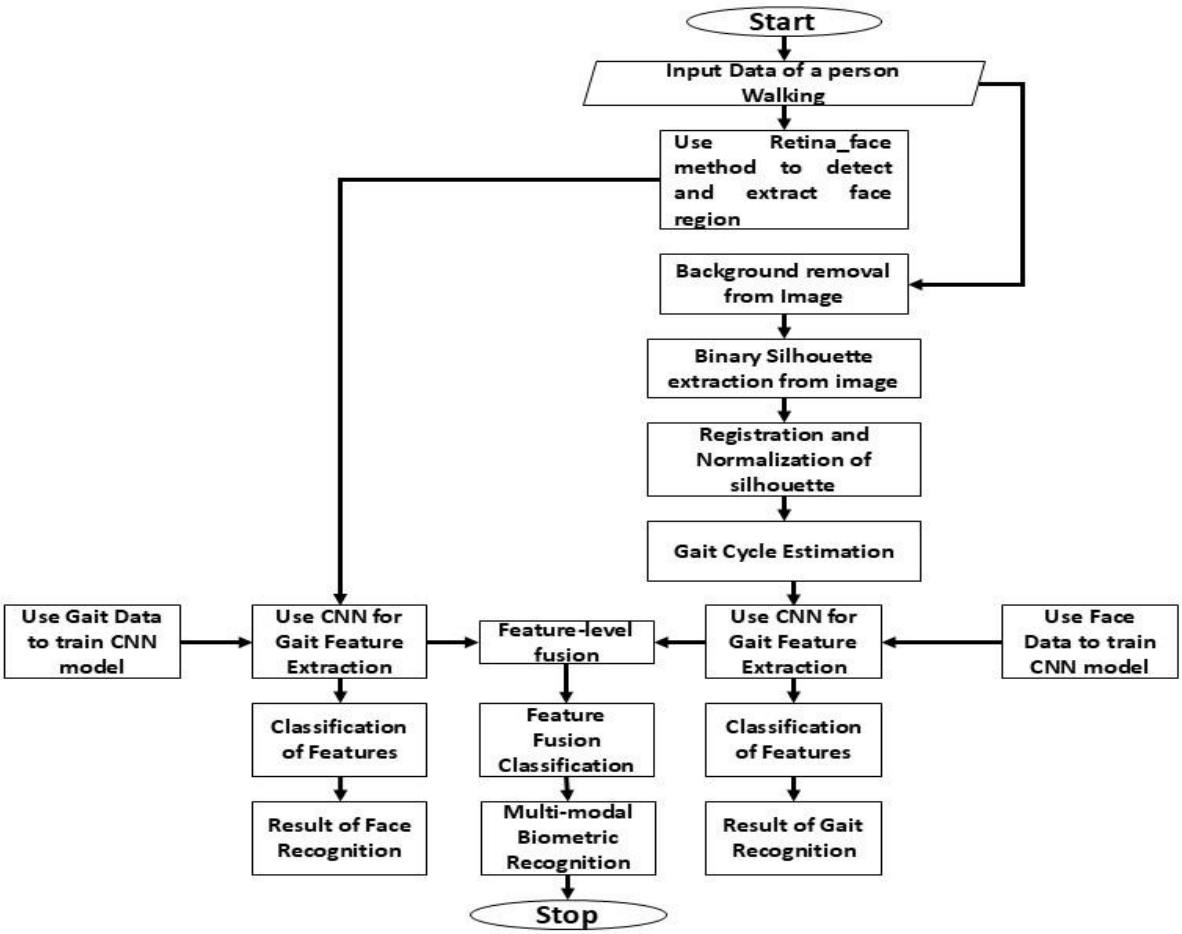

Figure 4: Flowchart of the Proposed Multimodal Biometric Recognition System

The proposed method creates size-normalized and horizontally centre-aligned silhouette pictures after obtaining the human walking binary silhouette image sequences. An effective spatiotemporal gait representation method for human walking characteristics in a whole gait cycle is the grey level of the GE picture, which allows for individual gait detection. The space-normalized energy image is described by each walking human silhouette picture. The time-normalized accumulated energy image, or GE, is the average cycle of the silhouette pictures into a single embodiment. The stance and the swing are the two stages that make up a human gait cycle. The heel strike of one foot initiates the step, which continues until the heel strike of the same foot is ready for the subsequent step. It is referred to as a whole gait cycle. Equation 2 defines a GE picture from gait silhouette image sequences (Aung et al., 2022).

$$G(x,y) = \frac{1}{N}\sum_{l=1}^{N} B_l(x,y) \qquad (2)$$

where x and y are values in the 2D image coordinate, t is the frame number in the image series, and N is the number of frames in a cycle's silhouette gait sequence. The suggested solution took into account a range of N values between 20 and 25 frames each cycle. The gait silhouette picture frame t in the series is denoted by Bt(x, y). Figure 5 displays the final LR face area and GE images.

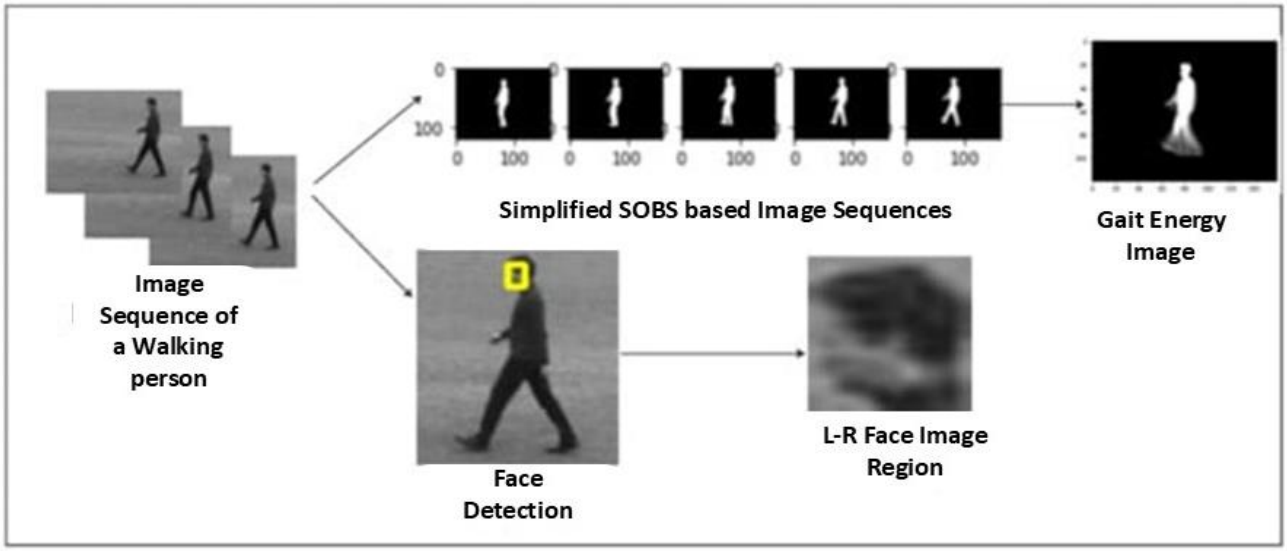

Figure 5: Sample of GE image and low-resolution face image extraction

2.3 ARCHITECTURE OF THE PROPOSED CNN MODEL

The study used LR faces and GE photographs as the input for the deep convolutional neural network, rather than HR faces and raw gait sequences. The particular design of our proposed CNN model is shown in Table 1. The model contains nine learning layers and uses a fixed input picture size of 224 × 224 pixels per channel. The convolutional layer, which includes 32 filters, is the initial layer and the main component of the network. These 3 x 3 filters are convolved with the input volume to produce an output feature map. Each layer's feature map size is determined by Equation 3 (Aung et al., 2022).

$$FeatureMap_{output} = \frac{W-F+2P}{S} + 1 \qquad (3)$$

Where S is the stride size, P is the padding boundary, F is the filter size, and W is the input volume size. To avoid overfitting and slower learning, a dropout with 0.5 rates and a Rectified Linear Unit (ReLU) transfer function are used as activators after the first layer. Each layer's trainable parameters are the quantity of learnable parameters that are impacted by the back propagation process. Each layer's learnable parameters are determined using the formula in Equation (4). Only the network's convolutional and fully linked layers have these parameters. To lower the dimensionality of the feature map and preserve the most important information, each pooling layer of the suggested model carries out the maximum sub-sampling operation (Aung et al., 2022).

$$Trainable\ Parameters = \left((filter_{size} + Depth) + 1\right) * num_{of_{filters}} \quad (4)$$

Prior to joining the final Fully Convolutional (FC) layers, the output from the preceding convolutional layers is flattened. These layers include the neurones for linking neurones between two layers, as well as the weight and bias. To avoid the source of overfitting problems, the flattened vector is then run through an FC layer and a 0.5-dropout layer, which randomly removes 50% of the network's nodes. An FC layer is also the final layer in the suggested model. The output classification result of the classes of the connected datasets is represented by the output layer of the suggested network.

2.4 CNN Classification Model Architecture and Multimodal Biometric Feature Fusion

Relatively big datasets have been recognised by the suggested model. We also explore its application to various jobs involving modest amounts of data. In order to outperform other methods, deep learning models need a lot of data, costly GPUs, and a lengthy training period, all of which raise the computational cost. By employing a learnt model on one task as part of the training process for another, the Transfer Learning (TL) methodology helps get around these drawbacks of deep learning approaches. Figure 6 illustrates how the suggested deep CNN model was used as a feature extractor in this stage by passing information from the base model to the classification model. Two separate feature extractors that do not share their weight values are one for the face and one for gait. The suggested technique preserves learning generic features in feature extraction by freezing the pre-trained layers of a base model. A new, smaller classification model is then fed the learnt characteristics. In order to learn more about a new dataset and carry out a classification procedure for unimodal identification, the model was made up of two FC layers. Figure 6 illustrates the proposed two-channel CNN feature fusion structure used in this study, where the fusion takes place at the feature layer.

In order to produce the greatest number of concatenated feature vectors, the suggested system performs a feature-level fusion procedure on every feasible combination of face and gait features after fusing the retrieved features from the two feature extractors for multimodal biometric recognition. For the multimodal recognition process, the classifier layer processes the integrated feature vector.

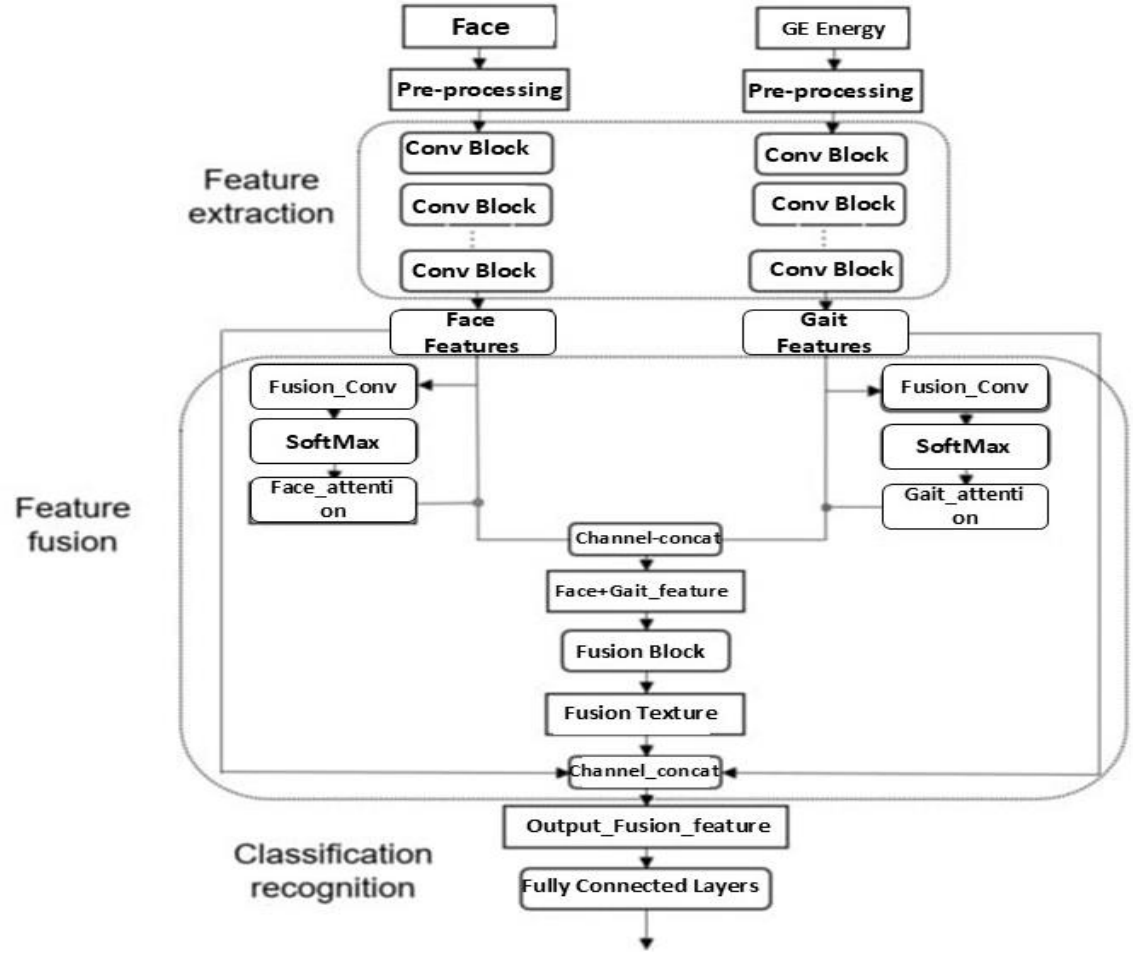

Figure 6: Feature Fusion Framework for the Multimodal Biometric Model

3. SYSTEM IMPLEMENTATION

MATLAB was used to create the multimodal biometric identification system, which entails a number of clearly defined phases. The goal of the pre-processing step is to extract useful information from facial photos and gait recordings. Pre-trained models, such as the Viola-Jones detector, are used to recognise faces. The discovered face areas are then cropped and resized to a common format for deep learning models. Human silhouettes are extracted from video sequences using background removal techniques such as the Gaussian Mixture Model (GMM) for gait recognition. Gait Energy Images (GEIs), which indicate a distinct gait characteristic for every person, are created by aligning and averaging these silhouettes. For the Convolutional Neural Network (CNN) to receive high-quality input, these processed datasets are essential.

The system makes use of MATLAB's Deep Learning Toolbox to use CNNs for feature extraction and classification. Metrics including accuracy, precision, recall, and F1-score are used to examine the system's performance, guaranteeing a thorough evaluation of its efficacy. The development process is further enhanced by MATLAB's robust visualisation and optimisation tools, which enable iterative improvements for a

biometric identification system that is incredibly precise and dependable.

4. SYSTEM RESULTS AND DISCUSSION

The input picture is 224 pixels in height and width with a channel, and the suggested model contains nine depth layers totalling 33.73 MB in size. The suggested model is a small network with fewer learnt layers and fewer trainable parameters. The enormous network issue that causes slowness to result in overfitting and other issues may be avoided by the tiny network. The model became more sophisticated as a result of the weights being raised by several depth levels. Overfitting and decreased accuracy were the results of a tiny percentage of the training sets in the large network.

The biometric data was successfully prepped for feature extraction during the pre-processing stage. The RetinaFace model effectively identified Left-Right (LR) face regions in a variety of settings, including changes in size, background, and illumination, with an accuracy of 87% for face identification. Clean silhouette sequences were effectively retrieved for gait analysis using the GMM-based background removal approach. Gait Energy Images (GEIs) were created by averaging these sequences, preserving unique gait characteristics while reducing noise. In order to properly prepare the face and gait data for feature extraction, this step was necessary.

The system successfully captured distinct biometric features by using a bespoke CNN for gait data and a pre-trained ResNet-50 model for face data. The advantages of both modalities were then combined by fusing these retrieved characteristics at the feature level. By using this method, the model was able to address issues that are frequently encountered in unimodal systems, such as occlusion and data loss. The fusion approach produced strong feature vectors for classification while preserving the discriminative strength of both modalities.

After a classification layer was applied to the fused features, a remarkable 92% identification accuracy was achieved across 25 patients. Important indicators like recall (93%) and accuracy (91%) demonstrated the system's ability to correctly identify and categorise people. The model maintained a solid trade-off between precision and recall, as seen by its balanced performance, as indicated by its 92% F1-score. Furthermore, compared to developing models from scratch, using transfer learning greatly shortened training time by around 60%, increasing computing efficiency.

Following a 10-fold cross-validation, the findings of the suggested multimodal biometric identification system are compiled in the Table 1below. To illustrate overall performance, the assessment metrics include Accuracy, Precision, Recall, and F1-Score for each fold along with their average values.

Table 1: Validation of the Results

Iteration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	91.8	90.7	92.3	91.5
2	92.5	91.6	93.0	92.3
3	91.2	90.4	92.0	91.2
4	93.0	92.2	93.8	93.0

Iteration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
5	92.7	91.9	93.4	92.6
6	91.5	90.6	92.2	91.4
7	92.0	91.1	92.7	91.9
8	92.8	91.9	93.5	92.7
9	93.3	92.5	94.0	93.2
10	92.4	91.5	93.1	92.3
Average	**92.3**	**91.4**	**93.0**	**92.2**

The findings of the 10-fold cross-validation show that the proposed multimodal biometric identification system is very effective and dependable across a variety of evaluation criteria. With an average accuracy of 92.3%, the technique demonstrates excellent reliability in correctly identifying individuals based on the combination of face and gait biometric characteristics. Additionally, the accuracy, recall, and F1-scores, which averaged 91.4%, 93.0%, and 92.2%, respectively, suggest that the system is well-balanced, lowering false positives and false negatives. This balance shows that the model can handle challenging biometric data, such as variations in facial features or gait patterns, while maintaining high classification performance.

The model's consistent performance across all folds further demonstrates its ability to generalise to unknown inputs. This achievement was made possible by the combination of feature-level fusion, transfer learning, and sophisticated pre-processing methods as LR face identification and Gait Energy Image (GE). Notably, the system's resilience is demonstrated by its capacity to adjust to a variety of situations, such as changes in settings, scale, and attire. These findings demonstrate that the suggested system is appropriate for real-world uses including access control, security, and surveillance where dependability and flexibility are crucial.

5. CONCLUSION

The study successfully developed a deep convolutional neural network (Deep CNN) model for multimodal biometric recognition based on transfer learning using face and gait information, obtaining high accuracy and resilience with minimal training data. By using advanced pre-processing techniques, such as Left-Right (LR) face detection and the Gait Energy Image (GE) representation for gait analysis, the system was able to successfully identify and combine significant biometric characteristics for improved identification performance. Transfer learning significantly reduced processing costs without compromising classification accuracy, facilitating the reuse of previously trained models.

The suggested method showed consistent and dependable results after a thorough examination using 10-fold cross-validation. It achieved good precision, recall, and F1-scores, with an average accuracy of 92.3%.These results highlight the system's potential for use in security, surveillance, and identity verification applications by demonstrating its ability to manage real-world difficulties such as scale, clothing, and ambient fluctuations. By highlighting the possibility of merging several features with cutting-edge deep learning techniques to improve the efficiency and reliability of

identification systems, this study lays a strong platform for future research in multimodal biometrics.

6. REFERENCES

Amine, N.-A. (Ed.). (2019). *Hidden Biometrics: When Biometric Security Meets Biomedical Engineering*. Springer.

Ammour, N., Bazi, Y., &Alajlan, N. (2023). Multimodal approach for enhancing biometric authentication. *Journal of Imaging, 9*, 168. https://doi.org/10.3390/jimaging9090168

Aung, H. M. L., Pluempitiwiriyawej, C., Hamamoto, K., &Wangsiripitak, S. (2022). Multimodal biometrics recognition using a deep convolutional neural network with transfer learning in surveillance videos. *Computation, 10*, 127. https://doi.org/10.3390/computation10070127

Bailey, K. O., Okolica, J. S., & Peterson, G. L. (2014). User identification and authentication using multi-modal behavioral biometrics. *Computers & Security, 43*, 77–89.

Boucherit, I., Zmirli, M. O., Hentabli, H., &Rosdi, B. A. (2020). Finger vein identification using deeply-fused convolutional neural network. *Journal of King Saud University - Computer and Information Sciences, 34*, 346–656.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2019). Retinaface: Single-stage dense face localisation in the wild. *arXiv* 2019, arXiv:1905.00641.

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 643–660.

Haider, S. A., Ashraf, S., Larik, R. M., Husain, N., Muqeet, H. A., Humayun, U., Yahya, A., Arfeen, Z. A., Khan, M. F. (2023). An improved multimodal biometric identification system employing score-level fuzzification of finger texture and finger vein biometrics. *Sensors, 23*, 9706. https://doi.org/10.3390/s23249706

Jomaa, R. M., Mathkour, H., Bazi, Y., & Islam, M. S. (2020). End-to-end deep learning fusion of fingerprint and electrocardiogram signals for presentation attack detection. *Sensors, 20*, 2085.

Lowe, J. (2020). Ocular motion classification for mobile device presentation attack detection. *University of Missouri-Kansas City*.

Maddalena, L., & Petrosino, A. (2008). A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Transactions on Image Processing, 17*, 1168–1177.

Mitra, S., & Gofman, M. (Eds.). (2016). *Biometrics in a Data-Driven World: Trends, Technologies, and Challenges*. CRC Press.

Ra'Anan, Z., Sagi, A., Wax, Y., Karplus, I., Hulata, G., &Kuris, A. (1991). Growth, size rank, and maturation of the freshwater prawn, Macrobrachiumrosenbergii: Analysis of marked prawns in an experimental population. *Biological Bulletin, 181*, 379–386.

Ramírez-Mendoza, R. A., Lozoya-Santos, J. D. J., Zavala-Yoé, R., Alonso-Valerdi, L. M., Morales-Menendez, R., Carrión, B., Cruz, P. P., Gonzalez-Hernandez, H. G. (Eds.). (2022). *Biometry: Technology, Trends and Applications*. CRC Press.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, 26 August 2004 (pp. 32–36).

Selvarasu, N., Nachiappan, A., & Nandhitha, N. (2010). Euclidean distance basedcolor image segmentation of abnormality detection from pseudo color thermographs. *International Journal of Computer Theory and Engineering, 2*, 514.

Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters, 141*, 61–67.

Wang, Y., Shi, D., & Zhou, W. (2022). Convolutional neural network approach based on multimodal biometric system with fusion of face and finger vein features. *Sensors, 22*, 6039. https://doi.org/10.3390/s22166039

Yu, S., Tan, D., & Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, China, 20–24 August 2006 (pp. 441–444).