Irish Interdisciplinary Journal of Science & Research (IIJSR)
Volume 10, Issue 1, Pages 113-124, January-March 2026

An Adaptive Wait-and-Watch Framework for Multimodal Biometric Recognition through Raw-Level Data Fusion Integrating Image-Based Iris Modality and Signal-Based Voice Modality for Enhanced Robustness

Dhiman Karmakar[1*], Subarna Sen[2], Debasmita Sen[3] & Swapnadip Mukherjee[4]

[1,2,3,4]*Department of Computer Science, Surendranath College, Kolkata – 700009, West Bengal, India.*
Corresponding Author (Dhiman Karmakar) Email: dhiman.karmakar@gmail.com[*]

DOI: https://doi.org/10.46759/iijsr/2026.10113

Article Received: 19 January 2026 *Article Accepted: 22 March 2026* *Article Published: 28 March 2026*

ABSTRACT

The use of more than one biometric modality, termed Multimodal Biometrics (MMB), has proven to be more efficient, accurate, and robust in automated human recognition than unimodal systems. This study proposes an adaptive human recognition system that fuses iris and voice samples at the sensor level. The novelty of this article is threefold. First, fusion of iris data (an image modality) with voice data (a signal modality) at the raw level is a rare approach in MMB research. This study successfully addresses the challenges of raw-level fusion, such as dealing with unnormalized and noisy data, while leveraging its advantages such as preserving the originality of information to achieve enhanced accuracy. Second, the adaptive nature of the system includes correctly classified test samples within the corresponding training class at each epoch, thereby increasing the training set size and improving the recognition rate. Third, we introduce a unique" wait-and-watch" approach to supplement adaptiveness in the context of the multiclass classification dilemma. We conducted experiments using the well-known databases CORPORA for voice and DOBES for iris traits. Finally, we concluded by using the t-test that the obtained outcome is statistically significant and satisfactory.

Keywords: Multimodal Biometrics; Feature Extraction; Spectrogram; LDA; PCA; KNN; N-D Subspace; Sensor and Score Level Fusion; Adaptive System; Real Time System; Equal Error Rate; False Acceptance Rate.

1. Introduction

Nowadays, biometric recognition, in various forms, is becoming increasingly integrated into everyday life. From facial recognition used to unlock our mobile phones to fingerprint verification for accessing our offices, biometric systems are everywhere. Even soft biometric traits—such as CCTV footage capturing body posture (gait), height, or skin color—are used to help identify criminals.

In the modern digital era, the use of biometric traits—such as facial features, iris patterns, fingerprints, and more—has become an integral aspect of daily life. From securing financial transactions to unlocking smartphones, biometric recognition systems are now embedded across a wide range of applications [1]. Biometrics leverage unique physiological and behavioral characteristics that are distinct to each individual, enabling reliable identification and authentication. The primary advantage of biometric systems lies in their ability to provide a secure, efficient, and user-friendly alternative to traditional methods such as passwords or identity cards, which are more vulnerable to loss, theft, or forgery.

Multimodal biometric systems (MMB), which integrate multiple biometric traits for identification and verification, significantly enhance the overall accuracy, security, and reliability of biometric recognition. Experimental studies have demonstrated that fusion in MMB systems can substantially improve performance by reducing both the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) [2]. Unlike unimodal systems, which depend on a single biometric trait, MMB systems can maintain accuracy even when one modality is compromised owing to environmental conditions or poor data quality [3]. Additionally, the simultaneous forgery of multiple biometric

traits is exceedingly difficult, making multimodal systems inherently more secure than their unimodal counterparts [4, 5]. By incorporating diverse sources of biometric data, MMB systems offer a more robust and comprehensive solution, thereby ensuring an improved recognition accuracy and system resilience.

In addition to the introductory part, the article is organized as follows. Section 2 describes the most recent stage in the development of multimodal analysis for image data and signal fusion. Section 3 discussed the proposed algorithm and its benefits over the traditional age-old techniques. Section 4 illustrates the databases used in the experiment, experimental results and their significance. Finally, Section 5 provides further scope for improvement of the proposed approach.

1.1. Study Objectives

The objectives of the present study are as follows:

(1) To achieve a significant improvement in recognition accuracy for multimodal biometric systems. (2) To develop and implement a novel strategy for the fusion of image-based traits (iris) and signal-based traits (voice). (3) To investigate the adaptive capability of the system by dynamically increasing the volume of the training dataset during runtime. (4) To address the challenges associated with processing noisy and unnormalized data acquired directly from sensors. (5) To incorporate a mechanism for deferring multiclass classification in cases of ambiguity when classifying a test sample. (6) To provide a new perspective for computer vision researchers by enabling the application of existing image analysis algorithms to speech data.

2. Literature Review

In multimodal biometric systems, fusion strategies are typically implemented at various stages of the recognition pipeline, including sensor-level, feature-level, score-level, and decision-level fusion. The choice of fusion level plays a crucial role in determining the system performance, as each level offers distinct advantages and challenges in integrating biometric modalities. Sensor level fusion (also referred to as raw-level fusion) [6] combines data from different modalities at the initial, preprocessed stage. This approach preserves the richest form of information, because fusion occurs before any significant processing or feature extraction. Consequently, this minimizes information loss and can lead to significant performance improvements. However, despite its potential to enhance model accuracy, sensor-level fusion has rarely been adopted in practical systems because of challenges related to synchronization, scalability, and data compatibility. Feature-level fusion [5] involves the combination of features extracted from multiple biometric traits into a single, unified feature vector. This method allows the system to work with compact and informative data representations, facilitating more efficient learning and decision making.

Score-level fusion integrates the matching scores generated independently by each modality, thereby offering flexibility and modularity. However, decision-level fusion aggregates the final decisions made by each modality, often using majority voting or rule-based strategies [7]. Each of these fusion strategies has its trade-offs, and selecting the appropriate fusion level is essential for optimizing system performance based on the application context and computational constraints.

Among the commonly used biometric traits, voice and iris are particularly popular because of their accessibility and user-friendliness. These modalities support contactless and non-intrusive data acquisition, making them ideal for real-time and user-centric applications. Voice recognition is especially suited for remote or hands-free environments such as phone banking. It captures both physiological attributes (e.g., vocal tract characteristics) and behavioral features (e.g., speech patterns), thereby providing rich identity-specific information. However, performance can be influenced by factors such as background noise, illness, or emotional variability. Iris recognition is known for its high accuracy and stability, although controlled conditions are often required for optimal image acquisition. When integrated into a multimodal biometric (MMB) system, voice and iris modalities complement each other by offsetting their individual limitations. Voice adds dynamic and behavioral context to more static, high-precision iris features, resulting in a more robust and reliable identification process.

Moreover, facial recognition can be used to offer visual verification, further enhancing the multimodal framework. The integration of these traits improves the system accuracy and significantly reduces both false acceptance and false rejection rates, leading to a more resilient and secure biometric recognition system. Abozaid et al. [8] proposed a multimodal biometric recognition system that integrated facial and voice modalities. In their approach, key features were independently extracted from facial images and voice signals, followed by fusion at both feature and score levels. Their results showed that score-level fusion yields a higher recognition rate than feature-level fusion.

In a related study, Dali et al. [9] applied feature-level fusion using three distinct schemes on pre-normalized facial and voice features to construct a multimodal biometric system. An Artificial Neural Network (ANN) was employed for classification, and the system's performance was assessed using Equal Error Rate (EER) and recognition rates, with comparisons drawn against recent benchmark approaches.

Byahut et al. [3] extracted facial features using Log-Gabor filters and Local Binary Patterns (LBP), while voice features were obtained using Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coefficients (LPC). These features were fused at the feature level, and classification was performed using the K nearest neighbors (KNN) algorithm. The effectiveness of the model was evaluated using standard performance metrics.

Zhang et al. [10] developed an Android-based multimodal biometric (MMB) authentication system utilizing facial and voice traits. Facial features were extracted using the LBP method, and voice features were obtained using Voice Activity Detection (VAD). An adaptive fusion strategy was employed to combine these pre-extracted features. The accuracy of the system was evaluated on an Android-based smart terminal. Tokhy et al. [11] introduced a Feature Fusion Center (FFC) to integrate biometric data from the fingerprint, iris, and voice modalities. Specific methods were used to extract key features from each trait, and classification was conducted using Support Vector Machines (SVM) along with the sum FFC approach. System performance was measured using Equal Error Rate (EER) and Receiver Operating Characteristic (ROC) curves, demonstrating robust performance and high authentication accuracy. Alsaade et al. [12] proposed a multimodal biometric system that utilized dynamic feature-level fusion. Facial features were extracted using Principal Component Analysis (PCA) and Gabor filters, whereas voice features were obtained via MFCCs. The fused features were tested across multiple

benchmark databases and yield promising results. Alharbi et al. [13] presented a score-level fusion-based user identification model, where voice recognition employed a Gaussian Mixture Model (GMM) and facial recognition used FaceNet. The integration of these models through score-level fusion leads to a significant reduction in the error rate.

In [14], a robust Multimodal Biometric Attendance System (MBAS) was introduced by combining facial and speech modalities at the fusion level. The system incorporates dynamic speech input for liveness detection. Linear Discriminant Analysis (LDA) was applied for dimensionality reduction, and classification was performed using a Bidirectional Long Short-Term Memory (Bi-LSTM) network. The performance of the model is evaluated using publicly available benchmark datasets.

Merit et al. [15] proposed a Bimodal Deep Learning Network (BDLN or BNet) that creates a unified feature vector by averaging the extracted features from the facial and voice data. Identification was conducted using Softmax activation and related methods, contributing to improved biometric security. In a separate study, Kailas et al. [16] developed a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model trained on features extracted from Electrocardiogram (ECG) and iris modalities. The fused feature vector was used for both training and testing, and system performance was evaluated using the established performance metrics.

In [19], Garg et al. developed a multimodal recognition system by integrating key features from three biometric modalities—fingerprint, iris, and ECG. They employed a Swin Transformer architecture for feature extraction, and their results demonstrated enhanced resistance to spoofing attacks. In a separate study, Riaz et al. [20] proposed a multimodal recognition system based on hand biometrics, specifically dorsal finger crease (DFC) and finger knuckle print (FKP). They utilized the circular shift combination local binary pattern (CSC-LBP) method for feature extraction. The extracted features were fused at the feature level, and a support vector machine (SVM) classifier was employed to evaluate the effectiveness of the developed model.

In [21], Abdul-Al et al. developed a hybrid multimodal biometric system by integrating visible and infrared facial images. They employed VGG16 for feature extraction, Principal Component Analysis (PCA) for dimensionality reduction, and Sequential Neural Networks (SNNs) for classification. Additionally, a weighted score-level fusion technique was used to enhance the overall accuracy of the system. Further, Chitrapu et al. [22] developed an explainable multimodal biometric system by integrating deep learning, trust-adaptive fusion, and encrypted domain matching to enhance privacy. They employed a trust-adaptive fusion strategy to improve the system's robustness against noisy inputs.

However, the cited studies treat audio signals and images as entirely distinct data types and apply separate algorithms to process each modality independently. This approach often results in increased computational cost and system complexity. By incorporating the visibility of audio data the transformation of audio signals into visual representations such as spectrograms, offers a promising alternative with several advantages.

First, by converting audio into image-like formats, established image processing algorithms can be leveraged to analyze audio data, potentially streamlining the overall processing pipeline. Second, visualizing audio facilitates

human interpretation of complex sound structures, enabling a clearer understanding of the patterns. This aids both manual inspection and the design of more effective algorithms. Third, in biometric recognition systems, visual representations of audio features help validate whether the extracted features are meaningful and discriminative. This can significantly enhance feature quality, leading to improved model accuracy and robustness.

In summary, incorporating the visibility of audio data bridges the gap between abstract sound and both human and machine interpretation, while offering a more unified and efficient approach to multimodal data processing.

3. Methodology

As previously mentioned, audio signals are seldom fused with images at the raw level. The major difficulty in handling raw audio is the substantial presence of noise. Raw facial images, for example, are more susceptible to noise than iris images because of the differences in sensor type and quality. Hence, the combined noise level in face-voice fusion is typically higher than that in iris-voice fusion. Therefore, in this particular case, the iris is a better alternative to the face. Because we do not treat the traits separately, a common noise reduction algorithm should be applicable to both. Fortunately, beyond dimensionality reduction, PCA is also capable of removing noise, because the eigenvectors associated with smaller eigenvalues are more likely to represent noise and are thus eliminated. We coined the term VOIRIS to refer to the image vector formed by concatenating the iris image with the image representation of voice. It is worth noting that applying PCA to VOIRIS not only projects the data onto the major variance region in n-dimensional space, but also implicitly removes unwanted data.

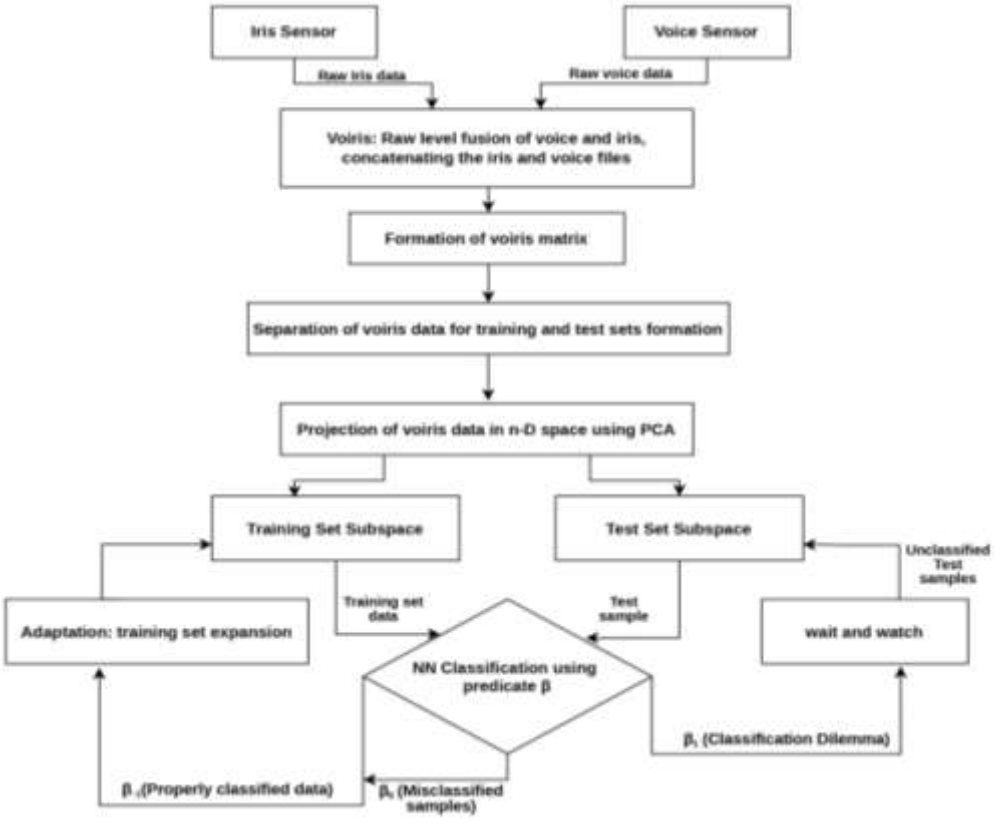

Figure 1. Architecture of a Multimodal Biometric Recognition System Based on Raw-Level Iris–Voice Fusion (Voiris), PCA Feature Projection, and Nearest Neighbour Classification with Adaptive Feedback.

The flow-graph in Figure 1 represents an abstract view of the proposed algorithm. First, the raw audio file was converted to an image form. To date the formation of spectrogram images from audio has been the most widely accepted conventional method. Some sample audio files from our database were converted to a spectrogram image form as shown in Figure 2. However, the short time Fourier transform applied to the digital audio sample incurred a substantial loss in raw data owing to the formation of complex numeral values. This data loss prevented the reversibility of the process.

To combat the aforementioned loopholes, we simply store the character-wise numeric values of the audio file in a vector of length l. This vector upon reshaping to matrix form can be treated as a voice-image. To fuse a voice image, vertically beneath an iris image of size m×n, we simply append $p = \lceil l / n \rceil * n - l$ zeros to the end of the vector to make the new extended length $L = l + p$ of the vector a multiple of n. Now the voice vector is reshaped to $l / n \times n$ matrix form. This voice-image is vertically concatenated to the iris image to form a fused voice-iris image matrix, coined as VOIRIS of size $(m + l / n) \times n$.

After segregating the entire VOIRIS dataset to form the training and test sets, PCA was applied to project the entire dataset in n-dimensional subspace. For classification purposes, an unknown VOIRIS test sample was quantified using predicate β capable of producing three output parameters: β_{-1}, β_0 and β_1. We simply used the NN algorithm for the classification cases.

Let *k* be equal to i for which $ed(s, t_i)$ is the minimum for all *i*, where *s* is the sample test VOIRIS, t_i is the *i-th* trained VOIRIS and ed is the Euclidean distance between them. Let *class(j)* signifies the class label of the *j-th* VOIRIS. If $class(t_k) = class(s)$ we conclude that the classification is correct, otherwise wrong. The rate of recognition is calculated as $\frac{number\ of\ correctly\ classified\ VOIRIS}{total\ number\ of\ test\ VOIRIS} \times 100$. To this point, in addition to the formation of voice-image and VOIRIS the process is straightforward. However, a twist occurs in this juncture. The entire classification process is governed by the predicate β based on a predefined threshold θ. Predicate β yields three output parameters namely, β_{-1}, β_0 and β_1 based on the following constraints.

case β_{-1}*(*t_k *is correctly classified):* $ed(s, t_k) < \theta$ *and*

$class(s) = class(k)$

case β_0*(*t_k *is unclassified, wait − n − watch):* $ed(s, t_k) > \theta$

and $class(s) = class(k)$

case β_1*(*t_k *is misclassified):* $ed(s, t_k) < \theta$ *and*

$class(s) \neq class(k)$

The intuition is that the process never becomes stuck although it fails to find a nearby neighbor of the test class sample. Instead, the process waits in a loop for re-classification and keeps monitoring the current scenario. A correctly classified test sample, during run time, was included in the training set to expand its size and subspace area. This is what we call the adaptation property of the system, as it helps increase the recognition rate

significantly. Owing to the gradual extension of the subspace area with time, for some unclassified sample *s*, the probability of obtaining tk such that $ed(s, t_k) < \theta$ is enhanced. However, the adaptiveness property also suffers from the drawback of including an incorrect entity in the growing training set (Case β_1).

4. Result and Discussion

Our model utilizes two biometric modalities: the iris and voice. For experimentation, each dataset was divided into training and testing subsets. Because iris and voice are uncorrelated traits and there has been a lack of availability of a single multimodal database of iris and voice, we make the following assumption. The left and right iris images of an individual (class A) from the iris dataset, and the corresponding voice samples of the same individual from the voice dataset, all represent the same class (person A).

We used the DOBES [17] iris database and Corpora Bangla [18] voice database. For each individual, the entire set consisted of six irises (three left and three right instances) and ten voice samples. Figure 3 shows a snapshot of VOIRIS formation from the above-mentioned databases. Each row in the figure represents a single person. Snapshots of five arbitrarily chosen individuals are shown. Two left irises and two right irises fused with different voice-images of a unique person are displayed in each row.

To construct the multimodal dataset, we pair two voice samples with each instance of the left iris and two different voice samples with each instance of the right iris of each person, concatenating them vertically to form 12 multimodal VOIRIS images per person. Note that, as the total number of voice instances for a single class is ten, two voice samples need to be repeated while fusing with the iris instances. To ensure a gradual increase in the number of training samples, during the experiment, we added one VOIRIS trained sample in different runs of the process. This result is shown in Figure 4.

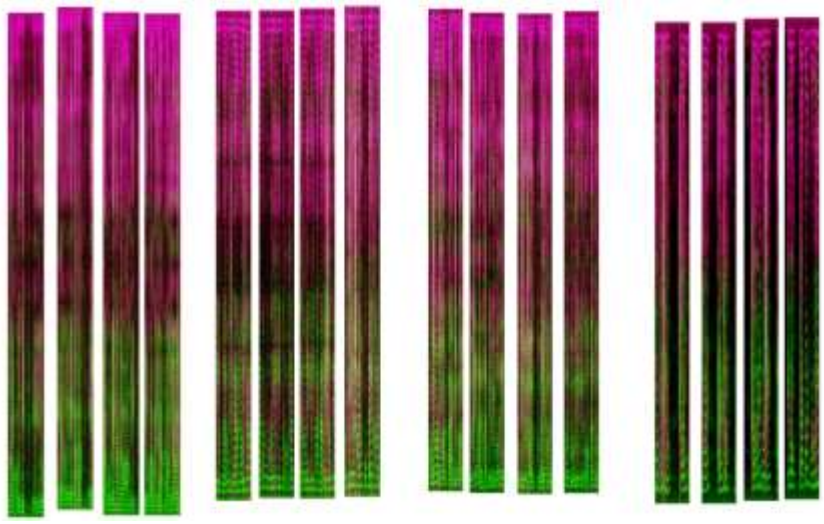

Figure 2. Spectrograms are constructed for four distinct classes corresponding to different individuals, with each class consisting of four independent voice samples to account for variations within the same speaker.

We plotted the number of training samples and recognition rate across the X and Y axes respectively. The graph compares three different approaches: score-level fusion of iris and voice using spectrogram, sensor-level fusion without using the adaptiveness trick, and the proposed approach.

In Figure 4, a remarkable improvement in the recognition rate is observed with the introduction of the adaptiveness property. This improvement is quite obvious for the following reasons. Suppose that at a point *x*, the *y−values* in the traditional and adaptive sensor levels are respectively y_1 and y_2. It is clear that because of the presence of adaptation, unlike the traditional curve, the number of training images is not actually *x*, rather much greater than *x* which in turn boils down to yield the result $y_2 > y_1$. In both cases, raw-level fusion seems to dominate the traditional score-level fusion using the spectrogram approach. This occurs because of the lack of preservation of data originality and truncating large complex values in the latter case.

The results of our experiments were found to be statistically significant using a well-established *t-test*, specifically Welch's *t-test*, which addresses the well-known Behrens-Fisher problem. This test was applied to assess the performance differences between the proposed model and the traditional approaches. To perform the analysis, 20 randomly selected pairs of samples were created. Each pair consisted of one misclassification rate from our model and one error rate from the traditional method.

The following formulas were used for the *t-test.*

$$g_p = S^2_p/n_p$$

$$g_q = S^2_q/n_q$$

$$t = (\underline{p} - \underline{q}) / \sqrt{g_p + g_q}$$

$$df = \frac{(g_p + g_q)^2}{g_p{}^2 / (n_p - 1) + / g_q{}^2 (n_q - 1)}$$

where, p and *q* denote two samples, S^2_p and S^2_q are the variance of sample *p* and *q*, respectively, n_p and n_q denote the size of the sample *p* and *q*, respectively, *t* and *df* stand for t-statistics and degrees of freedom, respectively, variables g_p and g_q are used for temporary storage purposes.

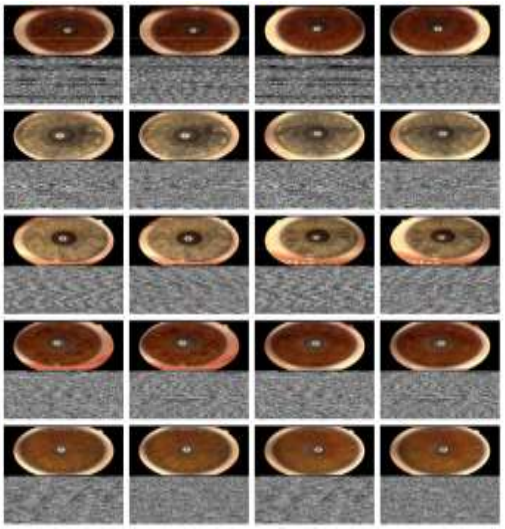

Figure 3. Arbitrary VOIRIS samples from five different individuals are displayed in a row-wise arrangement, such that each row corresponds to a specific person, enabling easy visualization and comparison of inter-person variations.

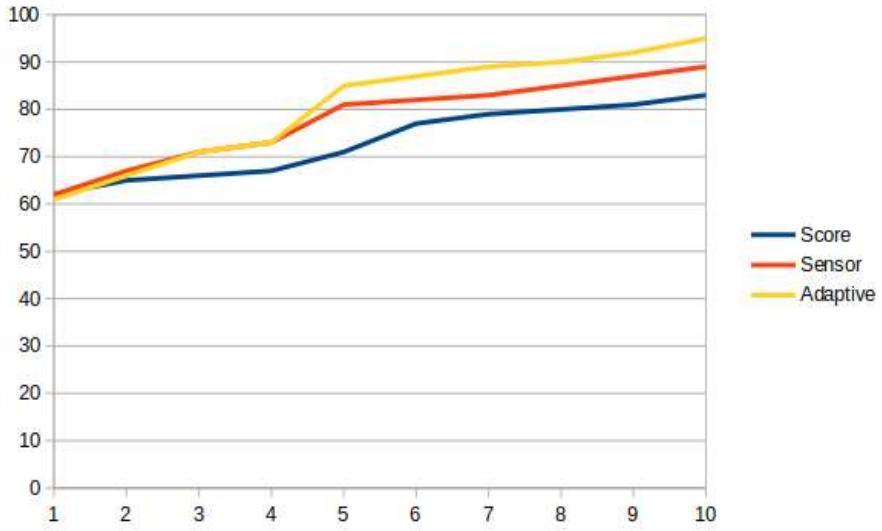

Figure 4. A comparison graph illustrating iris–voice fusion using spectrogram representations is presented across three approaches: (i) score-level fusion, where matching scores are combined; (ii) traditional sensor-level fusion, where raw sensor outputs are integrated; and (iii) adaptive sensor-level fusion, where fusion is dynamically adjusted based on data characteristics.

In our analysis, the sample standard deviations are found to be *17.2835* and *23.1280* respectively. The computed degree of freedom *(df)* was *35.1763*, and the test statistic *(t)* was *1.0314*. This value was compared with the critical *t−values* from a standard *t−distribution* table at a significance level of *0.05*. For reference, the critical *t−values* were approximately *1.697* for *df = 30* and *1.684* for *df = 40*. Because our calculated *t−value (1.0314)* is less than both critical values, the null hypothesis stating that the traditional and proposed methods perform equivalently, can be rejected. Thus, our method was statistically proven to outperform traditional approaches.

5. Conclusion and Future Recommendation

Undoubtedly, the increase in training set size, owing to adaptiveness, increases the rate of recognition in the majority of cases. However, the wait-and-watch situation may worsen because of repeated misclassifications. In such a scenario, the inclusion of a faulty training sample in the training set may radically deteriorate the system. In this context, the proper selection of predefined threshold θ, holds the stability of the system. This model operates on the assumption that iris and voice samples labeled as class A belong to class A, despite being sourced from uncorrelated databases.

Although this assumption is acceptable for initial experimentation, future work could benefit from (1) using real-world multimodal datasets, referred to as VOIRIS, where both the iris and voice truly originate from the same individual (e.g., person A). (2) Expanding this research to larger and more diverse datasets captured in real-world settings would help validate the robustness of the system. (3) Incorporate anti-spoofing techniques to address vulnerabilities in voice biometrics. (4) Further enhance system robustness and reliability for practical, real-world biometric applications. (5) Develop adaptive threshold selection mechanisms that can dynamically adjust θ based on system performance and environmental variations. (6) Explore advanced machine learning or deep learning techniques to improve feature extraction and multimodal fusion accuracy.

Declarations

Source of Funding

This study did not receive any grant from funding agencies in the public, commercial, or not–for–profit sectors.

Competing Interests Statement

The authors have not declared any conflict of interest.

Consent for publication

The authors declare that they consented to the publication of this study.

Authors' contributions

This article was prepared by four authors. The first author, Dr. Dhiman, was involved in idea generation, conceptual design, intellectual content development, partial coding, and data analysis. The second author, Ms. Subarna, was responsible for drafting the paper and conducting the majority of the literature review. The third author, Ms. Debasmita, programmed the voice data to be displayed as images. The fourth author, Mr. Swapnadip, handled the coding related to spectrogram formation and raw-level fusion. All authors agreed to be accountable for all aspects of the work.

Informed Consent

Not applicable for this study.

Availability of data and material

The datasets generated and/or analyzed during the current study are available in the http://phoenix.inf.upol.cz/iris/ and http://www.isca-speech.org/archive.

Institutional Review Board Statement

Not applicable for this study.

Ethical Approval

Not applicable for this study.

Acknowledgments

The authors acknowledge the support of Surendranath College, Kolkata, India where they are affiliated.

Declaration of Artificial Intelligence

The authors declare that no artificial intelligence (AI) tools or AI-assisted technologies were used in conducting the research or preparing this manuscript.

References

[1] Lawless, S., & Lawless, C. (2024). Biometrics, presents, futures: The imaginative politics of science–society orderings. Science and Public Policy, 51: 274–284.

[2] El Rahman, S.A., & Alluhaidan, A.S. (2024). Enhanced multimodal biometric recognition systems based on deep learning and traditional methods in smart environments. PLoS ONE, 19: e0291084.

[3] Byahatti, P., & Shettar, M.S. (2020). Fusion strategies for multimodal biometric system using face and voice cues. IOP Conference Series: Materials Science and Engineering, 925: 012031.

[4] Mai, S., Zeng, Y., & Hu, H. (2023). Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. IEEE Transactions on Multimedia, 25: 4121–4134.

[5] Byeon, H., et al. (2024). Artificial intelligence-enabled deep learning model for multimodal biometric fusion. Multimedia Tools and Applications, 83: 80105–80128.

[6] Bhuvana, J., Barve, A., Shah, P.K., & Dikshit, S. (2024). Image sensor fusion for multimodal biometric recognition in mobile devices. Measurement: Sensors, 36: 101309.

[7] Kazi, M., et al. (2024). Face, fingerprint, and signature-based multimodal biometric system using score level and decision level fusion approaches. IETE Journal of Research, 70: 3703–3722.

[8] Abozaid, A., Haggag, A., Kasban, H., & Eltokhy, M. (2019). Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. Multimedia Tools and Applications, 78: 16345–16361.

[9] Cherifi, D., Boushaba, S., & Nait-Ali, A. (2020). Feature level fusion of face and voice biometrics systems using artificial neural network for personal recognition. Informatica, 44: Article 2596.

[10] Zhang, X., Cheng, D., Jia, P., Dai, Y., & Xu, X. (2020). An efficient Android-based multimodal biometric authentication system with face and voice. IEEE Access, 8: 102757–102772.

[11] El Tokhy, M.S. (2021). Robust multimodal biometric authentication algorithms using fingerprint, iris and voice features fusion. Journal of Intelligent & Fuzzy Systems, 40: 647–672.

[12] Alsaedi, N.H., & Jaha, E.S. (2022). Dynamic audio-visual biometric fusion for person recognition. Computational Materials and Continua, 71: Article 1608.

[13] Alharbi, B., & Alshanbari, H.S. (2023). Face-voice based multimodal biometric authentication system via FaceNet and GMM. PeerJ Computer Science, 9: e1468.

[14] Jha, K., Jain, A., & Srivastava, S. (2024). Feature-level fusion of face and speech based multimodal biometric attendance system with liveness detection. AIP Advances, 14: Article 11011.

[15] Merit, K., & Beladgham, M. (2024). Enhancing biometric security with bimodal deep learning and feature-level fusion of facial and voice data. Journal of Telecommunications and Information Technology, Pages 31–42.

[16] Kailas, A., & Murthy, G.N.K. (2024). Deep learning based biometric authentication using electrocardiogram and iris. International Journal of Artificial Intelligence, 13: 1091.

[17] Machala, L., & Dobeš, M. (n.d.). UPOL iris image database. Available at: http://phoenix.inf.upol.cz/iris/. (Accessed 31 Jan 2026).

[18] Alam, F., Habib, M., Sultana, D., & Khan, M. (n.d.). Development of annotated Bangla speech corpora. Available at: http://www.isca-speech.org/archive. (Accessed 31 Jan 2026).

[19] Garg, R., Pathak, P., & Singh, M.P. (2025). A multimodal biometric recognition system based on fingerprints, iris and ECG via Swin transformer and CNN model. Systems and Soft Computing, 200369.

[20] Riaz, I., et al. (2025). Multimodal biometric recognition system based on feature-level fusion of dorsal finger crease and finger knuckle print. IEEE Transactions on Artificial Intelligence.

[21] Abdul-Al, M., et al. (2026). Fusion-enhanced hybrid multimodal biometric system: Integrating visible and infrared facial recognition for robust authentication. IEEE Access, 14: 6006–6028.

[22] Chitrapu, P., Morampudi, M.K., & Kalluri, H.K. (2026). A secure and explainable multimodal biometric system using trust adaptive fusion for face and fingerprint. Scientific Reports.