
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 5, pp. 2118–2129, DOI 10.12694/scpe.v26i5.4733

A CHALLENGE-RESPONSE BASED AUTHENTICATION APPROACH FOR MULTIMODAL BIOMETRIC SYSTEM USING DEEP LEARNING TECHNIQUES

KHUSHBOO JHA*, ARUNA JAIN, AND SUMIT SRIVASTAVA

Abstract. Multimodal Biometric System (MBS) is an advanced progression of conventional biometric authentication system, which employ multiple biometric traits to enhance security. However, despite their advantages, these systems are vulnerable to presentation attacks, where adversaries use photos, replay videos or voice recordings to deceive the authentication process. Therefore, this paper proposes a challenge-response based approach using texture-based facial features and multidomain speech features. The challenge-response approach requires the user to utter a random word. Next, the system detects the user's facial features (eye and mouth motion) and recognized speech text to confirm whether the authentication request originates from a legitimate user or an imposter. The feature-level fusion via concatenation method is used to combine these image-audio features, to reduce the overlap within the feature spaces and data dimensionality. The fused feature vector is then fed into the deep learning-driven ensemble classifier CNN-BiLSTM to train and test the fused samples for user authentication. The performance evaluation is carried out using a self-built database with 55 users, achieving 96.81% accuracy, 98.20% precision and an Equal Error Rate (EER) of 3.37%. Moreover, the proposed approach surpasses different cutting-edge MBS, deep learning classifiers and image-audio fusion techniques on various performance metrics. Thus, the results underscore the effectiveness of the deep learning-based MBS in ensuring user authentication and spoof detection, demonstrating its considerable potential in bolstering the security of biometric systems against intricate presentation attacks.

Key words: challenge-response, multimodal biometric system, authentication, ensemble classifier, pattern recognition, deep learning

1. Introduction. With the growing need for secure user authentication [1] in digital systems, MBS has emerged as a vital solution. Unlike traditional systems that rely on a single biometric trait, such as speech (speaker) or face recognition, often face challenges related to accuracy and susceptibility to spoofing. MBS [2] integrate multiple modalities to enhance security, reliability and robustness. Such biometric based authentication [1] offers a promising alternative by leveraging unique physiological or behavioral traits for user authentication. Moreover, this makes biometric authentication systems crucial for ensuring security and verifying identities in various applications like access control, attendance monitoring, etc. Among the most common biometric modalities, face and speaker recognition are favored for their contactless nature and user convenience, making them suitable for various real-world applications. However, despite their advantages, these systems are vulnerable to presentation attacks, where adversaries use photos, replay videos, or voice recordings [3] to deceive the authentication process.

The challenge-response [2, 4] technique, a well-established method in cybersecurity, offers a promising solution to this problem by requiring users to provide dynamic inputs in response to randomly generated prompts. Also, it uses a two-way conversation between an authentication server and the user. Under these systems, the verifier creates an arbitrary challenge for the user. The user has to reply with a right response that shows their validity. The authentication server creates and delivers a random challenge to the user upon a client login request. To finish the authentication process, the user must then react appropriately for the difficulty. This approach has great benefits, especially in terms of resilience against sophisticated spoofing attempts, such video attacks, even if it adds some complexity since it depends on active user engagement and incurs extra operating expenses. This approach greatly improves the security of the authentication system by adding dynamic challenges, which introduce uncertainty.

Consequently, this work proposes an enhanced multimodal biometric authentication approach [2] combining face and speech modalities at the feature level, within a challenge-response framework [2, 4]. By leveraging

*Department of Computer Science and Engineering, Birla Institute of Technology, Ranchi, India (kjha.phd@gmail.com).

synchronized eye blink with texture based facial feature [3] and multidomain speech features [5, 10], the system detects liveness and prevents presentation attacks [4]. By integrating these complementary modalities, the proposed method aims to significantly improve the accuracy, robustness, and security of biometric authentication systems, leveraging deep learning techniques for feature extraction and classification to achieve state-of-the-art performance. The proposed approach aims to address several key challenges:

- Authentication Accuracy: By combining multiple biometric modalities, our approach enhances authentication accuracy compared to single-modal systems. Feature-level fusion of face and speech traits enables more robust and reliable user identification, reducing the risk of unauthorized access.
- Resilience to Spoofing Attacks: MBS are inherently more resistant to spoofing attacks compared to single-modal systems. By requiring simultaneous presentation of face and speech traits, our approach mitigates the risk of spoofing attempts, such as using forged images or recordings.
- User Experience: We prioritize user experience in our security approach by leveraging natural and intuitive authentication mechanisms. Users can authenticate themselves using familiar actions like speaking a passphrase while facing a camera, minimizing the cognitive burden and increasing acceptance of security measures.
- Privacy Preservation: Unlike some other biometric modalities, such as fingerprints or iris scans, face and speech traits can be captured without physical contact, preserving user privacy and hygiene. Our approach ensures compliance with privacy regulations and minimizes concerns about intrusive surveillance.
- Adaptability to Environmental Factors: Diverse environments with varying lighting conditions, ambient noise levels and user interactions. This MBS is designed to be robust and adaptable to such environmental factors, ensuring consistent performance across different scenarios.

By addressing these challenges and objectives, this research aims to advance the state-of-the-art in MBS and contribute towards the development of more robust and reliable authentication mechanisms for various real-world applications. Face recognition captures spatial characteristics of an individual's face, while speech recognition analyzes vocal patterns and linguistic features, providing additional layers of security through multimodal fusion. The rationale behind combining these two modalities lies in their complementary nature and potential for greater authentication precision and resistance to spoof attempts. From computational perspective, training and evaluation of the proposed MBAS is done using self-built dataset for Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) classifier. In this context, our research focuses on developing a novel security approach using a MBS based on feature-level fusion of face-speech traits. This research contributes to the advancement of biometric security by introducing a novel MBS [4] based on feature-level fusion [5, 6] of face-speech traits, offering enhanced protection against cyber threats and ensuring the integrity and reliability for authentication based applications.

Contribution of this work is:

- To design and implement a novel MBS that combines multidomain acoustic and texture based facial features with eye blink movement at the feature level, for robust authentication.
- Deep learning based ensemble classifier CNN-BiLSTM for efficient classification of feature vectors thereby improving the performance of the authentication system.
- Challenge-response approach to verify the liveness of the user, thereby improving authentication accuracy and robustness in biometric systems.

Deploying a challenge-response based multimodal biometric system using face and speech in real-world scenarios presents a number of practical implications that need to be carefully considered, particularly concerning scalability and user acceptance. Below are key aspects to address in order to ensure the system's successful deployment and adoption;

Scalability across diverse use cases: The system needs to be adaptable to different scales of implementation, from small-scale applications like personal devices and secure office environments to large-scale implementations such as public venues and national ID systems. The system's architecture should be flexible to scale up or down based on the specific needs of the environment.

User acceptance: For any biometric system to gain widespread user acceptance, it must be easy to use. The challenge-response mechanism should be intuitive and user-friendly, particularly when asking users to

perform tasks such as responding to a spoken prompt or making specific facial gestures. Ensuring that these prompts are simple and non-intrusive is critical for encouraging regular use.

2. Literature review. For liveness detection [3], most of the work either use speech [10] or face recognition [3]. Using speech recognition [4] technology, spoken language is first detected and parse into individual words and sentences. Speech recognition operates on audio data and does not require visual information. Nevertheless, there exists a correlation and complementarity between visual and vocal data, namely in relationship to enhance the performance and security of biometric systems. There exist a limited number of publications dedicated to the multimodal biometric liveness detection [3] of faces and voices. Blanchard et al., [8] developed a memory creation and biometrics-based approach to avoid identity theft. This authentication approach combines ocular biometrics and challenge systems, as well as a new biometric characteristic called pupil memory effect. Credentials can be revoked at any time without any loss to the user, and the approach can be changed for varying levels of security. The technique assesses security and performance and recommends ways to improve deployment.

Addressing photo and video-attacks in face recognition systems, Chou et al., [4] proposed a score-level fusion and challenge-response scenario multimodal presentation attack detection (PAD) method . When presented with a challenge-response situation, the user is asked to say a series of randomly generated words. The liveness of the user is confirmed by detecting both their mouth motion and the specified speech text concurrently. In terms of improving the security of facial recognition systems, experimental results using a selfmade dataset show that the suggested strategy achieves the best half total error rate of 3.64%. Haasnoot et al., [9] present a rigorous and systematic framework and classification system for challenge-response protocols. The classification is validated by comparing it with published literature that specifically describes Biometric Challenge-Response Protocols (BCRP). Lastly, analyze the advantages of robust BCRPs over PAD approaches, particularly in safeguarding individual applications and preventing unintended leaks in BCRP applications.

Hanumanthaiah et al., [2] proposed a multi-modal biometric system that uses face and iris biometrics and challenges users with an emotion-invoking image at random on the screen. Response metrics include changes in the region of interest and distance between iris and facial landmarks. To detect presentation attacks, the response is compared to the expected response, which is determined by the position of an arbitrary emotion-invoking image on screen. In addition, the paper presents a hybrid deep feature that adds security to the recognition process by ensuring authentication failure for fraudulent samples. The proposed technique has a very low error rate of 0.95% in spoof leakage in varied situations when compared to prior works. However, the proposed work employs eye movements in tandem with texture-based face features, speech and speaker recognition as the challenge-response to authenticate the user's identity. For speaker recognition multi-domain based acoustic features [10] are used to enhance both temporal dynamics and spectral characteristics, improving robustness and discrimination between training and testing samples. Enhanced performance is achieved by integrating face and voice information through feature level fusion and ensemble deep learning based CNN-BiLSTM classifier.

3. Problem Formulation. Presentation attacks [4, 9] pose a significant threat to biometric authentication systems, where adversaries attempt to bypass security measures using fake biometric samples. Existing single-modality systems, such as those relying solely on facial recognition, are particularly vulnerable to such attacks. Moreover, traditional liveness detection methods may fail when faced with sophisticated spoofing techniques, such as high-quality deepfake videos or carefully crafted voice recordings. To address these challenges, this research formulates the problem of designing a resilient multimodal biometric authentication system capable of countering both photo and video-based presentation attacks. The proposed solution leverages a challenge-response [2, 4, 9] scenario, where users are prompted to perform specific actions, such as speaking randomly generated words while undergoing real-time facial analysis. By synchronizing speech and facial motion features and employing feature-level fusion [19, 21, 22], the system robustly verifies liveness. Moreover, this allows for more effective utilization of information from each modality and facilitates the extraction of discriminative characteristics that may not be evident when considering modalities in isolation. When processing individual modalities, some features may be redundant or not useful for classification. For instance, certain facial features (like background elements in the image) or irrelevant speech features (such as noise) might not contribute positively to the authentication process. By fusing the two modalities at the feature level, the system is better positioned to discard or minimize the impact of such irrelevant features through dimensionality reduction

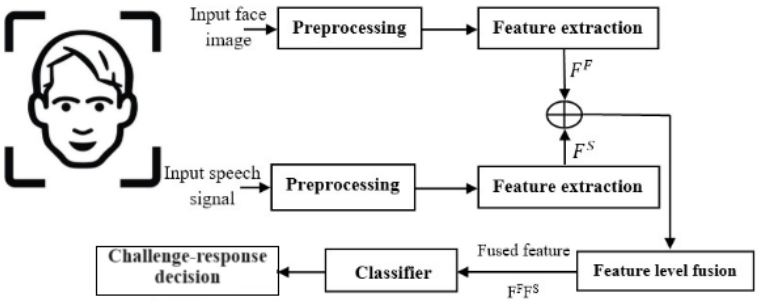

Fig. 4.1: Multimodal biometric system.

technique like Linear Discriminant Analysis (LDA) [5]. LDA focus on maximizing class separability by emphasizing features that are most relevant for distinguishing between individuals, thereby reducing feature overlap between modalities and improving the clarity of the final decision-making. Therefore, the goal is to enhance security without compromising usability, ensuring that the authentication process remains both effective and user-friendly in diverse application scenarios. The research aims to achieve a high level of accuracy, precision, and resilience against presentation attacks while maintaining a seamless user experience.

4. Proposed Method. A challenge-response [2, 4] scenario forms the foundation of the proposed approach. Upon user login request to a MBS, the proposed system generate a random word (challenge) which the user needs to respond (speak as response) accordingly within a specified duration. The dynamic texture based facial features (including mouth movement, eye blink) and the speech based user's answer (the requested word) is computed. The fused image-audio template undergo training/testing phase and compared (with the stored template in the reference database) by using deep learning ensemble classifier to ascertain the authenticity of the login user, as shown in Figure 4.1. As per the threshold value. the challenge-response decision i.e., whether the respondent is alive/legitimate or not is taken. If not, then the session expires. Else, the login requested is accepted and the user is successfully authenticated.

Two pivotal stages, training and testing [6, 10], are crucial components in the user authentication process for any MBS. In the training phase, users commence the procedure by completing the registration application. This approach entails recording a short video utilizing an intelligent sensor camera (in-built laptop camera with microphone) and the speech input is preprocessed to acquire reliable speech characteristics. Simultaneously, we extract a facial image frame from the motion picture to facilitate efficient feature extraction. In order to achieve feature-level fusion, the feature vectors obtained from both modalities are combined, with the goal of using the complimentary information offered by the aural and visual cues. Furthermore, a classifier is used to train these combined feature templates. These templates that have undergone training are then saved in a database for the following testing phase. During the testing step, which replicates the training procedure, the genuine identity as well as liveness of the claimed user is verified. The user inputs their visual-audio biometric template as per the given challenge which is then fused and compared to the trained benchmarks stored in the database. This comparison is conducted via a one-to-one approach [6, 10], where the claimed user's template is compared to all registered templates in order to establish its authenticity. The ultimate determination of the user's legitimacy is established by evaluating the level of resemblance between the request template and the source data, as assessed by predetermined performance standards.

4.1. Pre-processing. In face recognition-based image processing, high-quality pre-processing procedures precede feature extraction stage and frequently encompass many steps aimed at enhancing the quality and accuracy of the image data. This study employs data augmentation methods including scaling, flipping and rotation to enhance dataset diversity, hence improving model robustness and generalization. These pre-processing methods improve feature extraction and analysis by reducing the effects of illumination, perspective, and other environmental factors. Similarly, speech pre-processing improves speech quality and discrimination. Therefore, cutting-edge noise reduction technique is employed using Python programming prior to feature extraction. It

eliminates noise, filters out extraneous frequencies, and normalizes amplitudes to improve signal-to-noise ratios.

4.2. Feature extraction.

4.2.1. Facial image. For face recognition [11], the Active Appearance Model (AAM) [12] stands as a pre-eminent algorithm for facilitating model-assisted object tracking and detection within the realm of computer vision. AAM, renowned for its efficacy, constructs representations of both texture and shape for a given object by synthesizing a compilation of real images. This synthesis culminates in a comprehensive depiction of detailed texture features, grounded in the patterns of intensity and color inherent to the object.

The crux of AAM's functionality (using 20 landmark points) lies in its adeptness at minimizing disparities between newly acquired images and synthesized counterparts. Operating under the premise that interpretation can be framed as an optimization endeavor, AAM endeavors to reduce differences through iterative adjustments. This process is encapsulated in the definition of a difference vector, denoted as δI is defined in equation (1), which delineates discrepancies between the grey level value (I_i) vectors of the preprocessed input face image (I_p^F) and I_m is the grey level vector of present model parameter. Through the iterative manipulation of the model parameter (c), AAM endeavors to ascertain the optimal alignment between model and image (I_p^F). Such that it matches each other the best by minimizing the size of difference vector $\Delta = |\delta I|^2$.

$$\delta I = I_i - I_m \tag{4.1}$$

Moreover, in this work AAM method is combined with Eye Aspect Ratio (EAR) [13], which serves as an effective tool for liveliness detection due to its sensitivity to eye movements, particularly opening and closing actions. Employing six facial landmarks around the eyes, EAR exhibits significant variations during these movements, rendering it robust for blink identification. Through the flashing technique, EAR values fluctuate noticeably, offering insights into dynamic facial expressions. This technique underscores the reliability of EAR in assessing facial liveliness, thereby contributing to advancements in biometric systems and human-computer interaction.

Therefore, AAM [24] with EAR serves as a cornerstone in the realm of facial feature analysis, by virtue of its capability to delineate and align intricate features with precision. Its scholarly approach underscores its significance in enabling nuanced interpretations and facilitating sophisticated applications within the domain of computer vision.

4.2.2. Speech signal. In the realm of speaker recognition [5, 6, 10, 14], multi-domain [10] feature extraction techniques play a vital role in capturing various aspects of the speech signal. Therefore, inspired by this, our previous work [10] has been selected for the purpose of implementing a comprehensive speech feature. The unique aspect of this study is in its investigation and incorporation of multidomains acoustics, namely, strategies for extracting features in the cepstral, frequency and time domains [10]. The objective is to augment the discriminative capability of the deep learning classifier by integrating several domains, thereby enhancing the performance of the biometric authentication system.

Thus, let us expound upon the work and analyze the benefits of each domain-specific characteristics: The first component is cepstral features, Power Normalized Cepstral Coefficient (PNCC) [6, 10], which produces the most precise modeling of human hearing by using an asymmetric noise suppression module to reduce the influence of ambient noise and reduce computing complexity. This allows PNCC to better represent the spectral characteristics of speech, particularly in noisy environments, resulting in improved robustness and discrimination. This study has considered the first thirteen PNCC features, which are sufficient for expressing the spectral characteristics of speech signals. The second frequency domain characteristic, known as the spectral peak-based Formant feature [6, 10], pertains to the resonance frequencies of the vocal tract that are unique to each speaker. These frequencies include discriminative information for correct articulation and perception, which are resistant to noise. By extracting these frequencies, the characteristic features of vowels and consonants can be captured, providing valuable information for tasks such as speaker authentication. We employed a sliding Hamming window with a duration of 30 milliseconds, an overlap of 20 milliseconds, and 128 Mel-filter bins spanning frequencies from 50Hz to 4 kHz. We have extracted three formant features. Zero Crossing Rate (ZCR) [6, 10], the third time domain feature, analyzes a signal in the time domain to identify its salient characteristic. ZCR quantifies the rate at which the speech signal changes sign (i.e., crosses the zero axis) over time. ZCR is particularly useful for differentiating between voiced and unvoiced speech segments and can serve as an indicator

of speech rate and energy distribution. We have used the Librosa library function in Python 3 to extract one zero crossing feature.

Moreover, these feature extraction approaches are cutting-edge in their respective domain [10] and, after amalgamation, a resilient and effective feature set is acquired as speech feature. By combining PNCC, Formant frequencies and ZCR, a comprehensive representation of the speech signal consisting of 17 features [10] is obtained, capturing ceptral, frequency and time domain characteristics. These multi-domain features [10] are instrumental in improving the performance of speaker recognition tasks.

4.3. Feature-level fusion. The Feature-level fusion [15, 16, 24] effectively combines the extracted features of facial image and speech modalities, capturing their complementary nature. This enables a more comprehensive depiction of the information beneath, improving the system's capacity to identify intricate connections among various modalities. Additionally, it decreases the number of dimensions in the combined feature set relative to alternative fusion like decision level [16], potentially resulting in enhanced computational efficiency and decreased complexity in later processing stages. In addition, it allows for smooth integration of various feature sets, allowing for the inclusion of specialized knowledge and providing flexibility in designing models. The concatenation method used in feature-level fusion for integrating speech-image features improves the system's reliability, discriminative capability and adaptability in MBS..

4.4. Classification. To get the final classified output (approved or unauthorized) for the user authentication, an ensemble classification technique is used. The particular features influence the selection of the CNN [3, 6] and Bi-LSTM [17]: The CNN classifier analyze the spatial characteristics of the fused feature vector. Convolutional layers selectively capture significant patterns and correlations present in the feature space. The BiLSTM classifier [17] captures temporal dependencies by bidirectionally processing sequential information. Moreover, for the temporal nature of both facial movements (e.g., blinking, nodding) and speech patterns, the Bi-LSTM network is employed to model sequential dependencies in the data. This helps to capture subtle variations in both face and speech, leading to more accurate classification even when there are changes in environmental conditions or user demographics. As a consequence, the ensemble classifier improves the robustness by combining the outputs of face-speech, each trained on different environmental conditions or demographic data. This mitigates the biases introduced by training on a specific set of conditions or users. Thus, the ensemble CNN-BiLSTM classifier prioritises the correlations between spatial and temporal features, which are crucial for effective classification processing. This technique takes the fused feature vector $F^F F^S$ and feeds it into ensemble CNN-BiLSTM classifier.

CNN layer description
- Layers: 3 convolutional layers.
- Filter Sizes: (3x3), (5x5).
- Pooling: Max pooling after each convolution.
- Activation: ReLU.
- Dropout: 0.3 to avoid overfitting.

BiLSTM layer description
- Layers: 2 BiLSTM layers.
- Units: 128 units per layer.
- Directionality: Bidirectional (forward and backward).
- Activation: tanh.

Ensemble Strategy
- Fusion Method: Weighted combination of CNN and BiLSTM outputs.
- Weights: Tuned based on validation performance.

Fully Connected (Dense) Layer
- Units: 64.
- Activation: ReLU.
- Dropout: 0.3.

Output Layer
- Units: 2 (binary classification: authenticated or not).

- Activation: Softmax for probability output.
- Loss Function: Categorical Cross-Entropy.
- Optimizer: Adam.
- Learning Rate: 0.001.

5. Result and Discussion.

5.1. Database. There are several well-known databases available for presentation attack detection [4] focusing on video-attacks or photo-attacks. Unfortunately, the video data do not have speech/voice information and they are not compatible with the challenge-response [2, 4] method being proposed in this research. Therefore, a self-made video database is prepared for this research work. The database is compiled using data collected from 55 subjects. A laptop equipped with a high-quality camera with a microphone is used to capture user's videos (with required audio) from a distance of approximately 40cm. Videos are captured in a controlled indoor environment with consistent lighting conditions at a frame rate of 36 frame per second. Total 550 videos, consisting of 385 authentic clips and 165 spoof videos. All users were asked to speak a simple word from the predefined word database to ensure authenticity of the clips. Each segment lasts for 6 to 8 seconds, with each word being repeated twice. There are two video-attack scenarios for the spoof clips: (1) gazing into the camera without uttering a word; (2) gazing into the camera and speaking words that are not included in the predetermined word database. The experiments are conducted using various tools and libraries such as Python, PyAudio, Dlib, Google translate API, OpenCV and scikit-learn. Few deep learning classifiers are chosen as baseline for this research: Support Vector Machines (SVM) [18], Random Forests (RF) [18], CNN [3, 6] and BiLSTM [17]. These classifiers will be compared to determine their effectiveness. A carefully curated database is used to divide the legitimate and spoofed videos into a training subset and a testing subset, ensuring a balanced representation of both. For experimental observation, 80% of the final database used for training and remaining 20% for testing of the proposed approach using ensemble classifier.

5.2. Assessment of proposed MBS with UBS using speech and face. In this comparative analysis, the performance of an Unimodal Biometric System (UBS) for speech and face, specifically speaker recognition [6] and face recognition [7], is compared against the proposed MBS that combines speaker and face recognition modalities [26] for user authentication. The evaluation as shown in Table 5.1 focuses on accuracy [10] and EER metrics [10] to assess the efficacy of each system. By examining the accuracy of individual biometric modalities and their fusion in the MBS, insights into the system's robustness and reliability are garnered. Moreover, the EER metric [10] offers a comprehensive measure of the system's ability to balance false acceptance [10] and false rejection rates [10].

Through this comparative analysis, a comprehensive understanding of the performance advantages and limitations of unimodal versus MBS is elucidated, providing valuable insights for the advancement and optimization of authentication technologies in real-world applications.

5.3. Assessment of proposed approach with different levels of fusion of face-speech feature. The analysis of our proposed approach encompasses various levels of fusion, including rank, sensor, decision, feature, and score fusion [16, 26] of speech-face features. Through meticulous examination and experimentation, we evaluate the performance and efficacy of each fusion level in enhancing authentication accuracy, resilience to spoofing attacks and user experience. By scrutinizing the outcomes of different fusion strategies as shown in Table 5.2 using various standard metrics [10] such as such as accuracy [10], sensitivity [10], specificity [10],

Table 5.1: Comparison of proposed MBS with UBS.

Biometric system	Features used	Accuracy (in %)	EER (in %)
UBS using speech	PNCC+Format+ZCR	94.67	10.14
UBS using face	AAM+EAR	93.81	15.29
MBS (proposed)	**Feature level fusion of the above features**	**96.81**	**3.37**

Table 5.2: Assessment of proposed approach with different levels of fusion (in %).

Performance metrics	Decision	Sensor	Rank	Score	**Feature (proposed)**
Accuracy	93.09	92.70	89.41	90.24	**96.81**
Sensitivity	96.62	93.72	88.92	87.35	**98.22**
Specificity	87.83	89.81	82.58	86.85	**95.00**
Precision	97.81	95.48	90.14	87.27	**98.20**
F-measure	96.51	94.16	89.43	88.75	**98.03**
MCC	89.52	91.53	88.50	86.52	**97.62**
NPV	86.33	84.26	89.54	90.64	**93.05**
FPR	10.16	9.05	11.22	15.64	**5.52**
FNR	6.37	8.73	8.60	7.87	**1.23**
EER	8.76	8.89	9.91	11.75	**3.37**

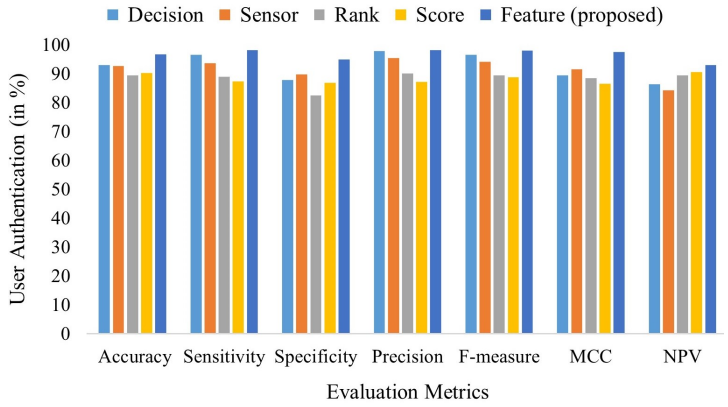

Fig. 5.1: Graphical presentation of proposed approach compared with different levels of fusion (in %).

precision [10], F-measure [10], Mathew's Correlation Co-efficient (MCC) [10], Negative Predictive Value (NPV) [10], False Positive Rate (FPR) [10], False Negative Rate (FNR) [10] and EER [10] are evaluated. Therefore, the aim of this research is to identify optimal configurations that maximize security and usability while maintaining compliance with regulatory standards. This comprehensive analysis provides valuable insights into the strengths and limitations of each fusion level, guiding future advancements in security.

To enhance the comprehension of the data reported in Table 5.2, we have included a comparative analysis in Figure 5.1. This figure demonstrates the performance of the proposed methodology (in %), for various levels of fusion.

5.4. Assessment of proposed MBS with various state-of-the-art classifiers. The analysis of our proposed approach extends to the evaluation of various classifiers [24, 25] to ascertain their efficacy in the context of our MBS. Through rigorous experimentation and comparative assessment [19, 20], we scrutinize the performance of different classifiers, including but not limited to Support Vector Machines (SVM) [18], Random Forests (RF) [18], CNN [8, 9], Bi-LSTM and CNN-BiLSTM classifiers as shown in Table 5.3. By examining metrics such as accuracy, Equal Error Rate (EER), and other computational efficiency, we aim to identify the classifier that offers the optimal balance of performance, robustness, and scalability for our proposed system. Furthermore, our analysis delves into the classifier's ability to handle the intricacies of multimodal data fusion, including feature-level integration of speech and facial features, and its resilience against spoofing attacks. Through this comprehensive evaluation, we seek to provide valuable insights into the suitability of different classifiers for enhancing security and mitigating cyber threats [1]. Ultimately, our findings

Table 5.3: Comparison of proposed approach (in %) with cutting-edge classifier based MBS.

Performance metrics	RF	SVM	CNN	BiLSTM	**CNN-BiLSTM (proposed)**
Accuracy	88.11	91.08	94.15	95.30	**96.81**
Sensitivity	89.62	92.72	93.92	95.52	**98.22**
Specificity	85.83	88.61	89.58	92.85	**95.00**
Precision	90.81	94.33	94.14	95.07	**98.20**
F-measure	89.02	93.16	93.43	95.75	**98.03**
MCC	79.65	77.35	91.05	95.17	**97.62**
NPV	80.89	83.26	89.54	92.61	**93.05**
FPR	10.16	9.23	8.61	7.46	**5.52**
FNR	8.37	6.73	4.07	4.87	**1.23**
EER	9.26	7.98	6.34	6.16	**3.37**

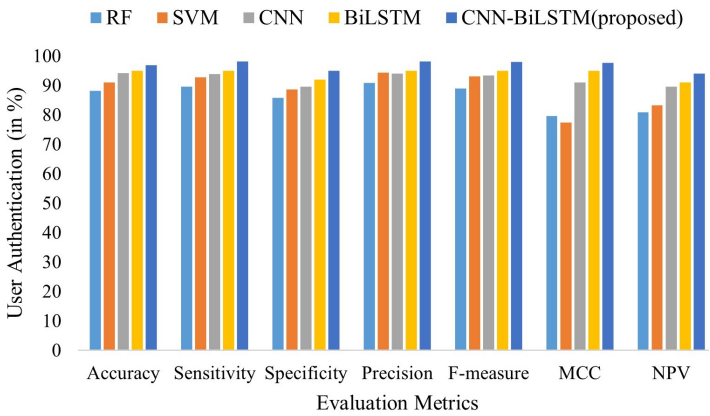

Fig. 5.2: Graphical presentation of proposed approach with different cutting-edge classifier based MBS (in %).

contribute to the advancement of biometric authentication systems for safeguarding critical infrastructure in digital authentication environments.

To enhance the comprehension of the data reported in Table 5.3, we have included a comparative analysis in Figure 5.2. This figure demonstrates the performance of the proposed methodology (in %), for different classifier based MBS.

5.5. Assessment of MBS with a few cutting-edge methods. The proposed MBS illustrates significant advancements in the field of MBS [26], by investigating a challenge-response approach that combines facial features and multidomain audio evaluation for user authentication. The existing literature primarily emphasizes the integration of facial expressions with other modalities apart from speech, or vice versa, as demonstrated in Table 5.4. Abinaya et al., [19] developed an MBS that integrates various modes of behavior, such as keystroke and speech features. This system utilizes advanced DL algorithms to precisely recognize individuals. The attributes from the two modalities were combined via a weighted linear method of feature-level fusion. The combined features were trained using a CNN classifier. Likewise, Abdulbaqi et al., [20] have proposed a system that employs Awica Wavelet Transform (AWT) algorithms to analyze the uniqueness of an individual's ECG signal in combination with their face features in order to verify users. Nevertheless, the achieved classification accuracy was only 94%. Rahman et al., [21] recently demonstrated a method for feature-level fingerprint and electrocardiogram (ECG) fusion using CNN classifier. Experiments show that the proposed technique has a 94.5% accuracy rate.

Vekariya et al., [22] presented a technique for multi-biometric authentication that also incorporates feature-

Table 5.4: Comparison of the proposed MBS with a few cutting-edge methods.

Ref.	Biometric trait	Fusion type	Database used	Method	Accuracy (in %)	EER (in %)
[19] 2022	Speech & keystroke	Feature	BioChaves	MFCC and press/release timestamp feature with CNN classifier	91.50	N/A
[23] 2022	Face & ear	Score	ORL (face) & IIT Delhi (ear)	DWT technique with ANFIS classifier	96.24	N/A
[20] 2023	Face & ECG	Decision	Self-made	AWT technique with DNN classifier	94.00	52.96
[21] 2024	ECG & fingerprint	Feature	MIT-BIH(ECG) & FVC2004 (fingerprint)	Deep embedding with CNN classifier	94.50	N/A
[22] 2024	Face & fingerprint	Feature	SDUMLA	BCOAK based feature with SVM classifier	96.00	N/A
Our	**Face & speech**	**Feature**	**Self-made**	Multi-domain speech feature with AAM-EAR with CNN-BiLSTM classifier	**96.81**	**3.37**

Abbreviations: N/A denotes Not Applicable, meaning that no such parameter was evaluated by the method.

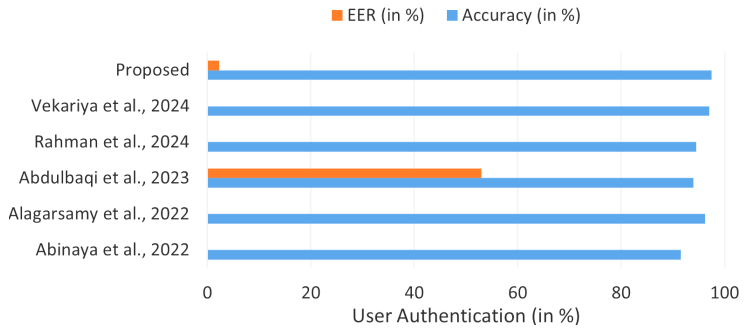

Fig. 5.3: Graphical assessment of proposed MBS with a few cutting-edge MBS.

level fusion. The proposed method employs a cutting-edge Binary Chimp Optimized Adaptive Kernel (BCOAK) SVM to identify the most impactful features. As per the experimental findings the system achieved an accuracy of 97%. Alagarsamy and Murugan [23] proposed a MBS combining ear and facial recognition using innovative approaches to improve identification accuracy. The system starts with preprocessing, ring projection, and data normalization. Discrete Wavelet Transform (DWT) feature extraction is used followed by Adaptive-Network- based Fuzzy Inference System (ANFIS) classifier to complete the identification process. The method calculates individual scores based on ear and facial feature matching results and fuses them to improve biometric identification. However, throughout history, people have traditionally relied on identifying others based on facial or speech signals, rather than using characteristics such as ears, hands, signatures, or fingerprints. Although prior studies have provided a solid foundation for applications of MBS, there has been limited research on the combination of face and speech interaction. This study's contribution is therefore original and significant.

In order to enhance the comprehension of the data shown in Table 5.4, we have included a comparative analysis in Figure 5.3. The following illustration indicates the efficacy of the proposed methodology in conjunction with several cutting-edge MBS. In a challenge-response MBS using face-speech, advanced deepfake techniques [27] present several vulnerabilities that could undermine the system's security. Deepfakes, which use artificial intelligence [28] to synthetically generate highly realistic facial images and voices, pose a significant threat to biometric authentication [1]. Therefore, to mitigate these vulnerabilities, combining several liveness

checks across modalities like ours (challenge-response i.e., simultaneous face and voice interactions) makes it harder for deepfakes to spoof [3] both aspects accurately. Moreover, the pilot study findings show that the proposed technique obtains impressive reliability and accuracy.

6. Conclusion. This research propose a novel authentication approach using MBS and integrating challenge-response technique using texture based facial and multidomain speech features. Leveraging facial and speech feature being fused at feature level and classified using deep learning based ensemble CNN-BiLSTM classifier, our approach demonstrates robustness and adaptability in securing various applications. Extensive experimentation and evaluation using the self-made database showcased remarkable results, with a computational perspective revealing a high accuracy rate of 96.81%, precision of 98.20% and an impressively low EER of 3.37%. Moreover, our approach surpasses different cutting-edge MBS, deep learning classifiers and image-audio fusion techniques on various performance metrics, validating its effectiveness in real-world scenarios. The proposed approach addresses the critical challenges of authentication accuracy, resilience against spoofing attacks and user experience enhancement while upholding privacy and regulatory compliance standards. These findings underscore the potential of our proposed methodology to effectively mitigate security risks and safeguard critical infrastructure within authentication environments. Thus, this research contributes to advancing the security landscape of internet based authentication environment, paving the way for enhanced protection and resilience against cyber threats in the rapidly evolving digital landscape. In the future, we plan to analyze and implement the proposed approach using dataset with users having speech or visual impairments.

REFERENCES

[1] Jha, K., Jain, A., & Srivastava, S. (2024). A Secure Biometric-Based User Authentication Scheme for Cyber-Physical Systems in Healthcare. Int. J. Exp. Res. Rev, 39, 154-169.
[2] Hanumanthaiah, A. K., & Eraiah, M. B. (2022). Challenge Responsive Multi Modal Biometric Authentication Resilient to Presentation Attacks. International Journal of Intelligent Engineering & Systems, 15(2).
[3] Jha, K., Srivastava, S., & Jain, A. (2024). A Novel Texture based Approach for Facial Liveness Detection and Authentication using Deep Learning Classifier. International Journal of Computational and Experimental Science and Engineering, 10(3).
[4] Chou, C. L. (2021). Presentation attack detection based on score level fusion and challenge-response technique. The Journal of Supercomputing, 77(5), 4681-4697.
[5] Jha, K., Jain, A., & Srivastava, S. (2024). Analysis of Human Voice for Speaker Recognition: Concepts and Advancement. Journal of Electrical Systems, 20(1), 582-599.
[6] Jha, K., Jain, A., & Srivastava, S. (2023, March). An Efficient Speaker Identification Approach for Biometric Access Control System. In 2023 5th International Conference on Recent Advances in Information Technology (RAIT) (pp. 1-5). IEEE.
[7] Jha, K., Srivastava, S., & Jain, A. (2023, March). Integrating Global and Local Features for Efficient Face Identification Using Deep CNN Classifier. In 2023 International Conference on Device Intelligence, Computing and Communication Technologies,(DICCT) (pp. 532-536). IEEE.
[8] Blanchard, N. K., Kachanovich, S., Selker, T., & Waligorski, F. (2020). Reflexive memory authenticator: a proposal for effortless renewable biometrics. In Emerging Technologies for Authorization and Authentication: Second International Workshop, ETAA 2019, Luxembourg City, Luxembourg, September 27, 2019, Proceedings 2 (pp. 104-121). Springer International Publishing.
[9] Haasnoot, E., Spreeuwers, L. J., & Veldhuis, R. N. (2022). Presentation attack detection and biometric recognition in a challenge-response formalism. EURASIP Journal on Information Security, 2022(1), 5.
[10] Jha, K., Srivastava, S., & Jain, A. (2025). A novel speaker verification approach featuring multidomain acoustics based on the weighted city block Minkowski distance. ETRI Journal, 47(2), 227-243, DOI 10.4218/etrij.2023-0485.
[11] Stasiak, L. A., & Pacut, A. (2010). Face Tracking and Recognition with the Use of Particle-Filtered Local Features. Journal of Telecommunications and Information Technology, (4), 26-36.
[12] Ueda, Y., Nakamura, K., Saegusa, C., & Ito, A. (2023). Recent advances and future directions in facial appearance research. Frontiers in Psychology, 14, 1154703.
[13] Hutamaputra, W., Utaminingrum, F., Budi, A. S., & Ogata, K. (2023). Eyes gaze detection based on multiprocess of ratio parameters for smart wheelchair menu selection in different screen size. Journal of Visual Communication and Image Representation, 91, 103756.
[14] Jain, P., Kasture, N. R., & Kumar, T. (2020). Comparative study of speaker recognition techniques in IoT devices for text independent negative recognition. Scalable Computing: Practice and Experience, 21(3), 359-368, DOI 10.12694/scpe.v21i3.1704
[15] Wang, Y., Zhou, Y., & Wang, B. (2024). Dimension Extraction of Remote Sensing Images in Topographic Surveying Based on Nonlinear Feature Algorithm. Scalable Computing: Practice and Experience, 25(5), 4246-4254.
[16] Dalila, C., Saddek, B., & Amine, N. A. (2020). Feature level fusion of face and voice biometrics systems using artificial neural network for personal recognition. Informatica, 44(1).

[17] Srikanth, J., & Shanmugam, A. D. (2023). A Deep LSTM-RNN Classification Method for Covid-19 Twitter Review Based on Sentiment Analysis. Scalable Computing: Practice and Experience, 24(3), 315-326.
[18] Bharathi, V. (2024). Vulnerability detection in cyber-physical system using machine learning. Scalable Computing: Practice and Experience, 25(1), 577-591.
[19] Abinaya, R., Indira, D. N. V. S. L. S., & Swarup Kumar, J. N. V. R. (2022, February). Multimodal Biometric Person Identification System Based on Speech and Keystroke Dynamics. In International Conference on Computing, Communication, Electrical and Biomedical Systems (pp. 285-299). Cham: Springer International Publishing.
[20] Abdulbaqi, A. S., Turki, N. A., Obaid, A. J., Dutta, S., & Panessai, I. Y. (2023). Spoof Attacks Detection Based on Authentication of Multimodal Biometrics Face-ECG Signals. In Artificial intelligence for smart healthcare (pp. 507-526). Cham: Springer International Publishing.
[21] A. El-Rahman, S., & Alluhaidan, A. S. (2024). Enhanced multimodal biometric recognition systems based on deep learning and traditional methods in smart environments. Plos one, 19(2), e0291084.
[22] Vekariya, V., Joshi, M., & Dikshit, S. (2024). Multi-biometric fusion for enhanced human authentication in information security. Measurement: Sensors, 31, 100973.
[23] Alagarsamy, S. B., & Murugan, K. (2022). Multimodal of ear and face biometric recognition using adaptive approach Runge–Kutta threshold segmentation and classifier with score level fusion. Wireless Personal Communications, 124(2), 1061-1080.
[24] Jha, K., Jain, A., & Srivastava, S. (2024). A Contactless Speaker Identification Approach Using Feature-Level Fusion of Speech and Face Cues with DCNN. Proceedings on Engineering Sciences, 6(3), pp. 1047–1056.
[25] Agrawal, S. S., Jain, A., & Sinha, S. (2016). Analysis and modeling of acoustic information for automatic dialect classification. International Journal of Speech Technology, 19, 593-609.
[26] 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(11), 115007.
[27] Sharma, V. K., Garg, R., & Caudron, Q. (2024). A systematic literature review on deepfake detection techniques. Multimedia Tools and Applications, 1-43.
[28] Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems With Applications, 124260.

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Aug 27, 2024
Accepted: Oct 2, 2024