Int. J. Adv. Eng. Comput. Sci. 2026; 2(1); 10-17

International Journal of Advances in Engineering and Computer Science

E-ISSN: 3121-6382

Volume 2, Issue 1, March, 2026
(An Academians Publishers of Research & Academic Resources)

Research Article

Multi-Modal Biometric Authentication System On Amazon Web Services Using Deep Learning Techniques

URBANUS, John Bille[1], Dr. Yusuf Musa Malgwi[2], Prof. Garba Joshua Etemi[3]
[1]Computer Science Department, Modibbo Adama University, Yola, Adamawa State, Nigeria
[2]Computer Science Department, Faculty of Computing
Modibbo Adama University, Yola, Adamawa State, Nigeria
[3]Computer Science Department, Faculty of Computing
Modibbo Adama University, Yola, Adamawa State, Nigeria

Abstract: The growing incidence of cyber threats, identity theft, and unauthorized system access has highlighted the limitations of traditional authentication methods such as passwords and PINs. In response, this study presents the design and implementation of a Multi-Modal Biometric Authentication System on Amazon Web Services (AWS) using Deep Learning Techniques, integrating facial recognition and voice biometrics to enhance security and reliability. The system leverages cloud-native services for scalability and real-time processing, utilizing Amazon Rekognition for facial recognition and a deep learning-based speaker verification model, ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network) deployed on AWS SageMaker for robust voice authentication. Additional AWS services including Lambda, API Gateway, Amazon S3, and DynamoDB are used to orchestrate data processing, storage, and communication between system components. A comprehensive dataset comprising facial images and voice recordings from 100 participants was collected under diverse environmental conditions to ensure variability and robustness. The system processes biometric inputs uploaded via a React-based frontend, where facial and voice features are extracted and matched against stored templates. Experimental evaluation was conducted using a dataset of 1000 samples, with 800 samples used for training and 200 for testing. Performance was assessed using standard biometric evaluation metrics, including accuracy, False Acceptance Rate (FAR), and False Rejection Rate (FRR). Results demonstrate that while individual modalities achieve high performance, the fusion of face and voice in a dual authentication framework significantly improves overall system accuracy and reduces error rates. The study concludes that multi-modal biometric authentication offers a more secure and resilient alternative to unimodal systems by mitigating issues such as spoofing, environmental variability, and single-point failure. The integration of the ECAPA-TDNN model with AWS cloud services ensures efficient, scalable, and real-time authentication suitable for deployment in high-security domains such as banking, e-governance, and enterprise access control systems. This research contributes a practical and scalable framework for implementing advanced biometric authentication systems in modern cloud environments.

Keywords: Multi-Modal Biometrics, Face Recognition, Voice Recognition, ECAPA-TDNN, Deep Learning, Amazon Web Services (AWS), Speaker Verification, Biometric Authentication, Cloud Computing, Security Systems.

1. Introduction

The rapid growth of digital services and the increasing sophistication of cyber threats have accelerated the transition from traditional authentication methods, such as passwords and PINs, to biometric-based systems. Biometrics - including facial features, voice, fingerprints, and iris patterns offer secure and non-transferable identity verification, reducing risks associated with credential theft and social engineering attacks (Jain *et al*., 2016). The integration of deep learning has further enhanced biometric systems by enabling robust feature extraction and efficient matching through compact embeddings, thereby supporting scalable and real-time authentication (Deng & Guo, 2020; Taigman *et al*., 2014).

Cloud computing has become a key enabler of modern biometric systems by providing scalable infrastructure, flexible deployment, and reduced operational complexity.

Received 12 January, 2026; Accepted 28 February, 2026; Available Online: 30 March, 2026

Platforms such as Amazon Web Services offer integrated services for biometric processing, storage, and deployment, facilitating real-time and large-scale authentication solutions (Amazon Web Services, 2023; Hashem *et al*., 2022). These capabilities support centralized management and efficient handling of biometric data in distributed environments.

However, unimodal biometric systems remain susceptible to environmental variability, spoofing attacks, and intra-subject variations. Additionally, concerns regarding demographic bias and fairness, particularly in facial recognition systems, have highlighted performance disparities across population groups (Phillips *et al*., 2018; Patel & Modi, 2021; Buolamwini & Gebru, 2018). These limitations necessitate more robust and inclusive authentication approaches.

Multi-modal biometric authentication, which combines complementary modalities such as face and voice, provides improved accuracy and resilience by leveraging the strengths of each modality (Kumar *et al*., 2019). Recent advances in voice embedding models, such as ECAPA-TDNN, further enhance system performance, especially in noisy environments (Desplanques *et al*., 2020).

This study therefore proposes a cloud-native, deep learning-based multi-modal biometric authentication framework that integrates face and voice recognition to enhance security, scalability, and reliability while addressing the limitations of unimodal systems and ensuring data protection compliance (Subashini & Kavitha, 2011; Almorsy *et al*., 2016; Deng *et al*., 2019; Desplanques *et al*., 2020).

2. Literature Review

2.1 Overview of Biometric Systems

Biometric systems are automated methods for identifying individuals based on unique physiological and behavioral characteristics, offering more secure and reliable authentication than traditional methods such as passwords and tokens (Jain *et al*., 2016). Biometrics are broadly classified into physiological modalities, such as facial features, fingerprints, and iris patterns, which are relatively stable and highly distinctive, and behavioral modalities, such as voice, gait, and keystroke dynamics, which reflect patterns of human activity (Ratha *et al*., 2019; Nautsch *et al*., 2019).

Physiological biometrics provide high accuracy due to their permanence, while behavioral biometrics offer flexibility in real-world applications despite potential variability caused by environmental and contextual factors. For example, voice biometrics analyze speech characteristics such as pitch and tone, and recent advancements in machine learning have significantly improved their robustness and reliability (Deng & Guo, 2020).

Modern biometric systems typically operate through stages including data acquisition, feature extraction, template storage, and matching or decision-making (Jain *et al*., 2016). The integration of deep learning techniques has further enhanced their performance by enabling the extraction of highly discriminative features from complex data, thereby improving accuracy in diverse and unconstrained environments (Taigman *et al*., 2014). As a result, biometric systems are increasingly deployed in applications such as banking, healthcare, and digital identity management, where secure and efficient authentication is essential.

2.2 Deep Learning in Biometrics

Deep learning has significantly transformed biometric authentication by enabling accurate and scalable recognition systems. Unlike traditional approaches that rely on handcrafted features, deep learning models automatically learn hierarchical representations from raw data, improving performance under varying conditions such as illumination, pose, and noise (Deng & Guo, 2020).

Convolutional Neural Networks (CNNs) are widely used for image-based biometrics, particularly in face recognition. They extract complex spatial features and generate compact embeddings for efficient matching. Models such as DeepFace and FaceNet have achieved state-of-the-art performance, demonstrating high accuracy in large-scale facial recognition tasks (Taigman *et al*., 2014; Deng *et al*., 2019).

For voice biometrics, Time Delay Neural Networks (TDNNs), especially ECAPA-TDNN, are commonly employed for speaker recognition. These models capture temporal dependencies in speech signals and produce robust voice embeddings, even in noisy environments (Desplanques *et al*., 2020).

The integration of CNN-based face recognition and TDNN-based voice recognition in multi-modal systems enhances authentication accuracy and robustness by leveraging complementary modalities. As a result, deep learning remains central to modern biometric systems, supporting reliable deployment in security-critical applications (Deng & Guo, 2020; Desplanques *et al*., 2020).

2.3 Multi-Modal Biometric Systems

Multi-modal biometric systems combine two or more biometric modalities, such as face and voice, to enhance the accuracy, robustness, and reliability of identity verification. Unlike unimodal systems, which rely on a single trait and are prone to noise, spoofing, and environmental variations, multi-modal systems leverage complementary information to reduce false acceptance and false rejection rates (Jain *et al*., 2016; Kumar *et al*., 2019).

A key advantage of multi-modal systems is their ability to compensate for the limitations of individual modalities. For instance, facial recognition may be affected by

lighting or pose variations, while voice recognition can degrade in noisy environments. By integrating both modalities, the system maintains reliable performance even when one modality is compromised (Nguyen *et al*., 2020).

Fusion techniques play a central role in multi-modal biometrics and are typically categorized into feature-level, score-level, and decision-level fusion. Among these, score-level fusion is widely used due to its balance between performance and implementation complexity (Jain *et al*., 2016; Kumar *et al*., 2019). With advances in deep learning, multi-modal systems can further improve accuracy and scalability by learning optimal feature representations and fusion strategies, making them suitable for high-security applications (Deng & Guo, 2020; Kumar *et al*., 2019).

2.4 Research Gap

Despite advancements in biometric authentication, most existing systems remain unimodal, relying on a single trait such as face or voice. These systems are susceptible to environmental variations, spoofing attacks, and demographic biases, which can affect their reliability in real-world applications (Jain *et al*., 2016; Patel & Modi, 2021). Moreover, many studies focus on improving individual modalities without adequately integrating complementary biometric traits to enhance overall system robustness.

Although deep learning has significantly improved biometric performance, its application is often limited to isolated or experimental implementations. Similarly, while cloud platforms such as AWS provide scalable and flexible infrastructure, their potential for supporting real-time multi-modal biometric systems remains underexplored (Hashem *et al*., 2022).

Therefore, a key research gap exists in the development of integrated, cloud-native multi-modal biometric systems that combine deep learning techniques with scalable infrastructure to achieve improved accuracy, security, and efficiency in dynamic environments.

3. Methodology

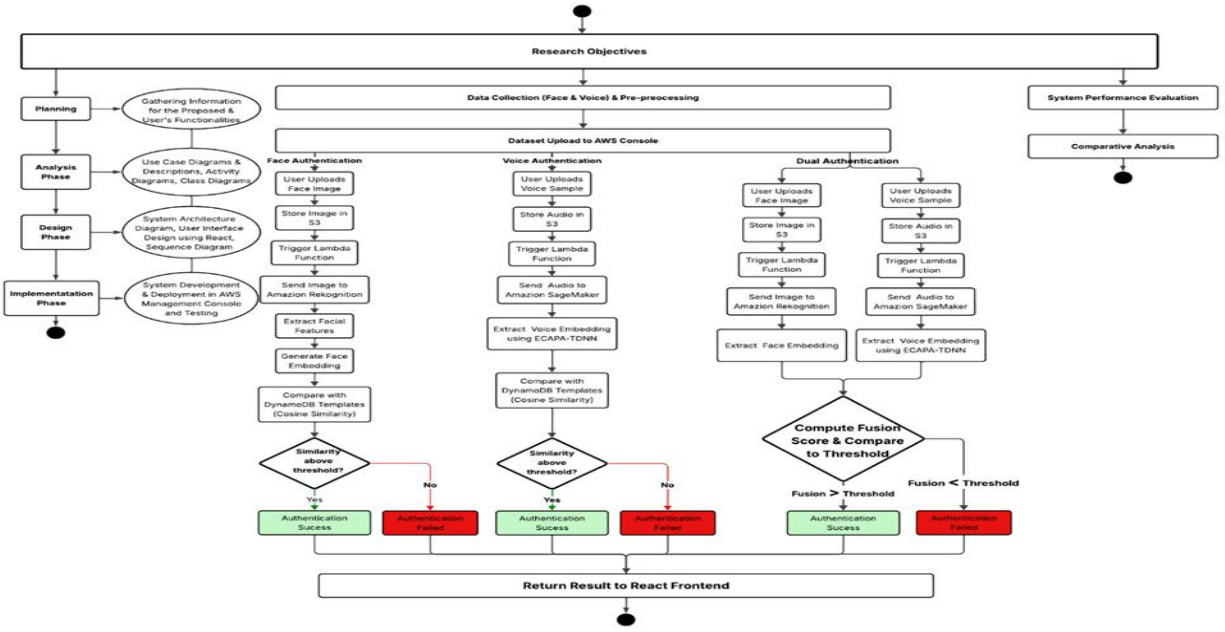

Figure 1: Overall Methodology

3.1 Research Design

This study adopts a system development research design involving the structured design, implementation, and evaluation of a multi-modal biometric authentication system. The process begins with planning and analysis, where system requirements are identified and modeled using use case, activity, and class diagrams. This is followed by the design phase, which defines the system architecture, user interface, and authentication workflows for both face and voice modalities.

The implementation phase utilizes a cloud-based architecture on AWS. Biometric datasets are collected, preprocessed, and stored in Amazon S3. Facial images are processed using AWS Lambda and Amazon Rekognition to generate embeddings, which are matched against templates stored in DynamoDB. Similarly, voice samples are processed using AWS Lambda and Amazon SageMaker, where the ECAPA-TDNN model extracts voice embeddings for verification.

The system also supports dual authentication, where face and voice embeddings are combined using a score-level fusion mechanism to improve accuracy and robustness. Finally, system performance is evaluated and results are returned to the frontend interface, ensuring a scalable and reliable authentication framework.

3.2 Dataset Collection and Preprocessing

The dataset used in this study comprises 1000 biometric samples collected from 100 participants, including both facial images and voice recordings. Each participant contributed multiple samples to capture variations in real-world conditions such as lighting, facial expressions, background noise, and recording environments, thereby enhancing system robustness.

Facial images were collected under varying conditions, including changes in illumination, pose, and appearance, and were preprocessed through face detection, alignment, normalization, and cropping to ensure consistency. Similarly, voice recordings were obtained in both quiet and moderately noisy environments and processed using noise reduction, segmentation, and normalization techniques. Feature extraction was then applied to prepare the data for deep learning models, enabling effective representation of facial and speech characteristics for accurate authentication.

3.3 System Architecture

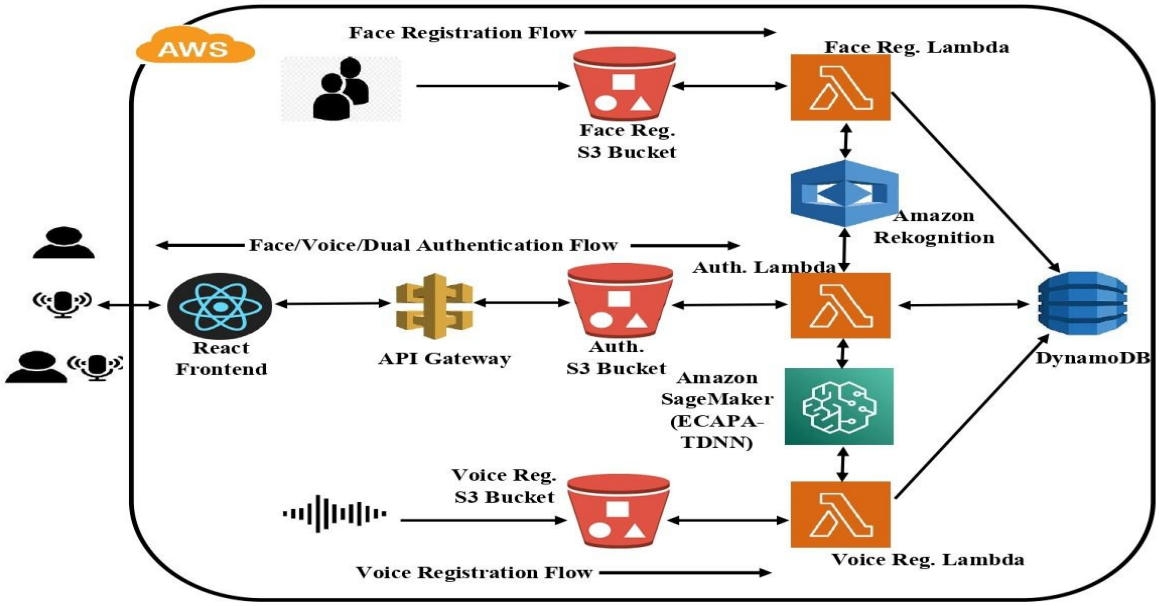

Figure 2: System Architecture

The system architecture, as illustrated in Figure 2, is a cloud-native, deep learning-based multi-modal biometric authentication framework designed to enhance security, scalability, and efficiency. It is deployed on Amazon Web Services, leveraging serverless computing, managed AI services, and distributed storage to overcome the limitations of traditional on-premise systems.

The architecture comprises three main workflows: face registration, voice registration, and authentication. During registration, biometric data is stored in Amazon S3 and processed using AWS Lambda. Facial images are analyzed with Amazon Rekognition to generate embeddings, while voice samples are processed using deep learning models on Amazon SageMaker to produce speaker embeddings. These templates are then stored in Amazon DynamoDB for efficient retrieval during authentication.

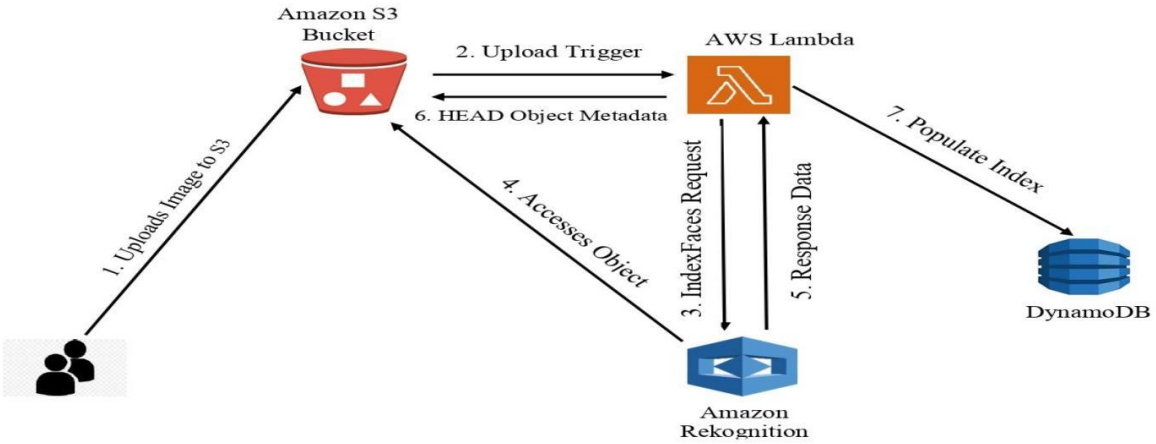

Figure 3: Face Registration Flow

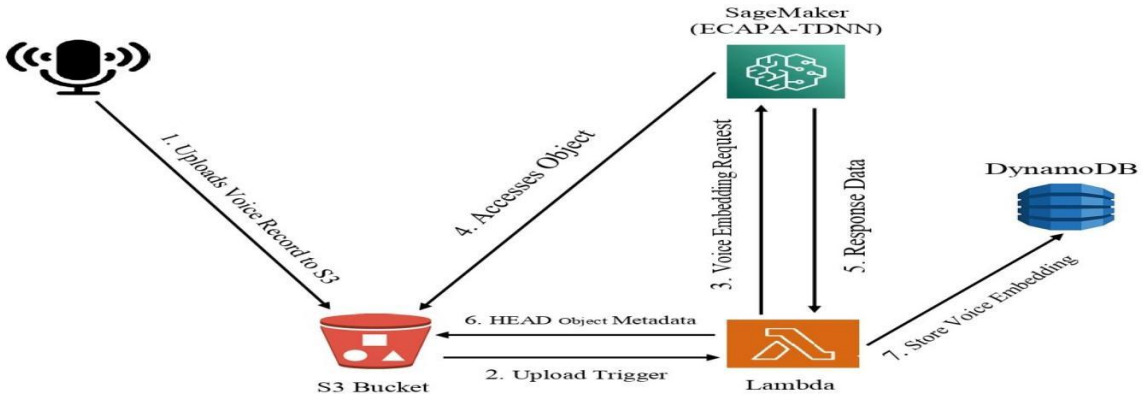

Figure 4: Voice Registration Flow

The authentication process supports three modes: face-only, voice-only, and dual-modality authentication. This approach improves robustness by allowing one modality to compensate for limitations in the other. The dual-modality mode further enhances security by requiring successful verification of both traits, thereby reducing false acceptance rates and improving resistance to attacks.

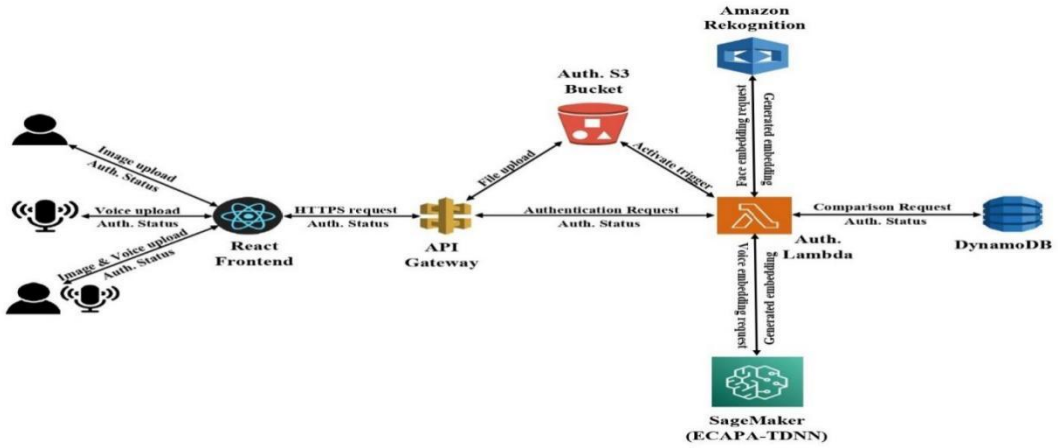

Figure 5: Authentication Flow

The system adopts a modular design that separates registration and authentication while maintaining seamless integration. This enhances scalability, simplifies maintenance, and supports independent model updates, providing a flexible and secure cloud-based solution.

3.4Tools and Technologies

a) Amazon Simple Storage Service (Amazon S3)
b) AWS Lambda
c) Amazon Rekognition
d) Amazon SageMaker
e) Amazon DynamoDB
f) Amazon API Gateway
g) React Frontend Interface

4. Results

4.1 Performance Evaluation

Table 1: Performance Evaluation Metrics

Mode	Accuracy (%)	FAR (%)	FRR (%)	Precision (%)	Recall (%)	F1 Score (%)	Response Time (s)	Usability (/10)
Face	94.5	3.2	2.3	94	94.5	94.2	1.8	8.9
Voice	92.3	4.1	3.6	92	92.3	92.1	2.1	8.5
Dual	98.1	1.2	0.7	98	98.1	98	2.4	9.3

4.2 Confusion Matrix

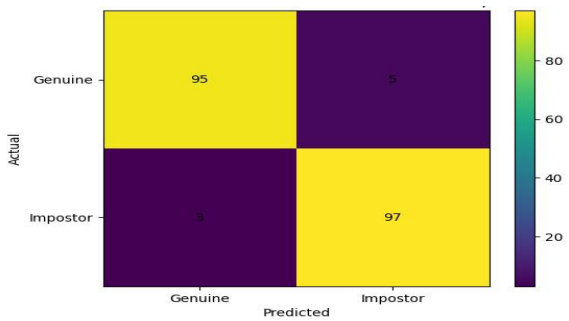

Figure 6: Face Authentication Confusion Matrix

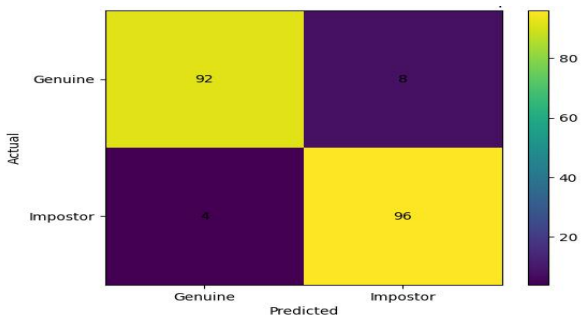

Figure 7: Voice Authentication Confusion Matrix

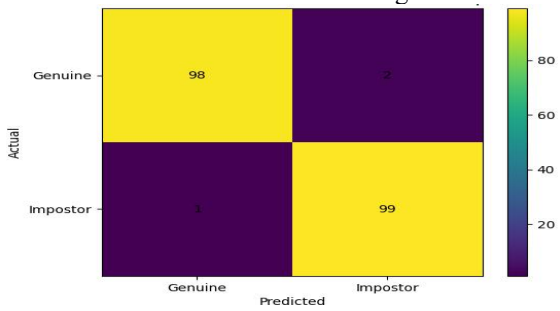

Figure 8: Dual Authentication Confusion Matrix

The confusion matrices as shown in figure 6, 7 & 8 above are based on the testing dataset of 200 samples, assuming a balanced distribution of 100 genuine users and 100 impostors. For face authentication, the system correctly classifies 95 genuine users and 97 impostors, with 5 false rejections and 3 false acceptances. Voice authentication shows slightly lower performance, correctly identifying 92 genuine users and 96 impostors, but with higher misclassification (8 false rejections and 4 false acceptances).

In contrast, the dual authentication system demonstrates the best performance, correctly identifying 98 genuine users and 99 impostors, with minimal errors (2 false rejections and only 1 false acceptance). This further validates that combining face and voice biometrics significantly enhances system accuracy, reduces error rates, and improves overall authentication reliability.

4.3 ROC Curve

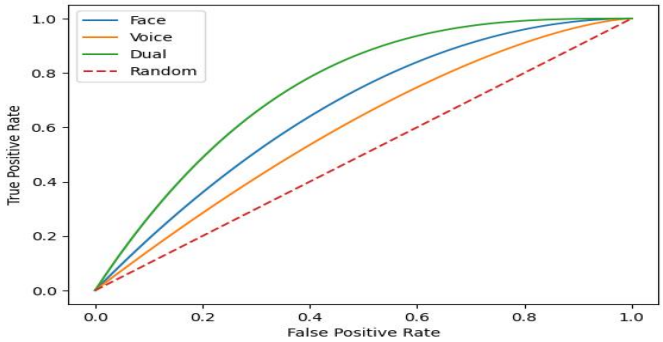

Figure 9: Receiver Operating Characteristic (ROC) Curve

The ROC curves in Figure 9 compare the performance of face, voice, and dual authentication models. The dual model shows the best performance, with its curve closest to the top-left corner, indicating high true positive and low false positive rates. Face authentication performs moderately, outperforming voice, while the voice model is closest to the diagonal baseline, indicating lower discrimination capability. Overall, all models outperform random classification, with the dual approach achieving the highest accuracy and reliability.

4.4 FAR and FRR Analysis

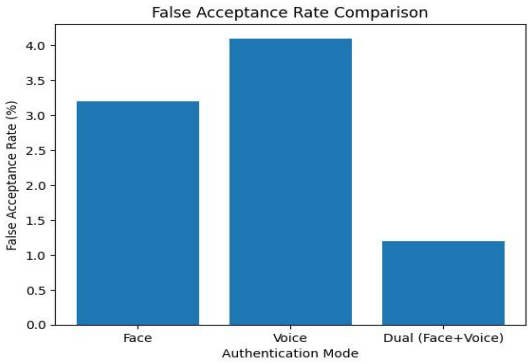

Figure 11: False Acceptance Rate Chart

The FAR results in Figure 11 indicate that voice authentication has the highest false acceptance rate (≈4.1%), followed by face authentication (≈3.2%), while dual authentication achieves the lowest FAR (≈1.2%), demonstrating improved security against unauthorized access.

Similarly, the FRR results in Figure 12 show that voice authentication has the highest rejection rate (≈3.6%), followed by face (≈2.3%), whereas dual authentication records the lowest FRR (≈0.7%). These findings indicate that multi-modal authentication reduces both false acceptance and rejection rates, achieving a better balance between security and usability.

5. Discussion

The results show that the multi-modal authentication system outperforms unimodal approaches across all metrics. The dual model achieved the highest accuracy (98.1%), compared to face (94.5%) and voice (92.3%), demonstrating the effectiveness of combining complementary modalities to improve reliability.

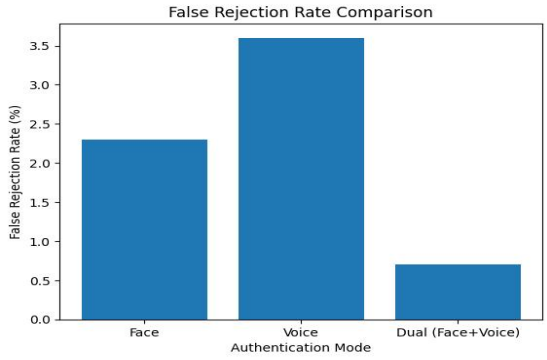

Figure 12: False Rejection Rate Chart

In terms of security, the dual system recorded the lowest FAR (1.2%) and FRR (0.7%), indicating improved resistance to unauthorized access and fewer errors in rejecting legitimate users. Voice authentication showed the highest error rates due to environmental sensitivity, while face authentication performed moderately. Precision, recall, and F1-score further confirm the robustness of the dual approach.

ROC and confusion matrix analyses also indicate that the dual model achieves the best balance between true and false classifications. Although it introduces slightly higher response time, it remains within acceptable limits and achieves the highest usability score, reflecting strong user confidence.

Overall, the findings confirm that multi-modal fusion enhances accuracy, security, and usability, making the system suitable for real-world deployment.

6. Conclusion

This study developed and evaluated a cloud-native, deep learning-based multi-modal biometric authentication system integrating face and voice

recognition. The results show that the dual authentication model outperforms unimodal systems, achieving higher accuracy (98.1%) while reducing both False Acceptance Rate (FAR) and False Rejection Rate (FRR).

The use of CNN-based facial recognition and ECAPA-TDNN for voice verification enabled effective feature extraction, while deployment on AWS provided a scalable and efficient infrastructure for real-time processing. The integration of cloud services ensured seamless data handling and system performance.

Overall, multi-modal authentication improves robustness against environmental variations, spoofing, and bias, while maintaining usability. Despite a slight increase in response time, the system offers a secure, scalable, and reliable solution for modern authentication, providing a strong foundation for future research and real-world applications.

References

Almorsy, M., Grundy, J., & Müller, I. (2016). An analysis of the cloud computing security problem. *Future Generation Computer Systems, 29*(6), 1317–1328.

Amazon Web Services. (2023). *AWS overview of cloud services*. https://aws.amazon.com

Bonneau, J., Herley, C., Van Oorschot, P. C., & Stajano, F. (2012). The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. *IEEE Symposium on Security and Privacy*, 553–567.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1–15.

Deng, J., & Guo, J. (2020). Retrospective analysis of deep learning for face recognition. *arXiv preprint arXiv:2003.01510*.

Deng, J., Guo, J., Niannan, X., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4690–4699.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *Proceedings of Interspeech 2020*, 3830–3834.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2022). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115.

Jain, A. K., Ross, A., & Nandakumar, K. (2016). *Introduction to biometrics*. Springer.

Kumar, A., Wong, D. C. M., Shen, H. C., & Jain, A. K. (2019). Personal verification using palmprint and hand geometry biometric. *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication*, 668–678.

Nautsch, A., Wang, X., Evans, N., Kinnunen, T., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., & Lee, K. A. (2019). ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8.

Nguyen, D. T., Yamagishi, J., & Echizen, I. (2020). Capsule-forensics: Using capsule networks to detect forged images and videos. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2307–2311.

Patel, V. M., & Modi, C. (2021). Face recognition challenges in unconstrained environments: A survey. *ACM Computing Surveys, 54*(3), 1–36.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2018). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception, 15*(2), 1–11.

Ratha, N. K., Connell, J. H., & Bolle, R. M. (2019). Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal, 40*(3), 614–634.

Subashini, S., & Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications, 34*(1), 1–11.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1701–1708.

Urbanus, U., & Yusuf, Y. (2024). Cloud-based face recognition system using Amazon Rekognition. *Journal of Cloud Computing and Security*, 12(1), 45–58.