LAUTECH Journal of Engineering and Technology 19 (5) 2025: 67-75
10.36108/laujet/5202.91.0540

Development of an AI-Driven Secure Safe Box Biometric Authentication System using Speech and Face Recognition Technique

Okomba N. S., Sobowale A. A., Esan, A. O., Omodunbi, B. A. and Awoyemi, T. A.

Department of Computer Engineering, Federal University Oye-Ekiti, Ekiti State, Nigeria

Article Info

Article history:

Received: 11/09/2025
Revised: 26/10/2025
Accepted: 16/11/2025

Keywords:

Biometric authentication, Speech recognition, Face recognition, Convolutional neural network (CNN), Recurrent neural network (RNN), Feature extraction, Secure safe box

Corresponding Author:
nnamdi.okomba@fuoye.edu.ng

ABSTRACT

The rapid advancement of Artificial Intelligence (AI) has enabled multimodal biometric systems that integrate speech and face recognition for more secure and reliable authentication. Unlike Unimodal systems that depend on a single trait and are prone to noise, lighting variations, spoofing, and user variability, multimodal systems combine complementary modalities to enhance robustness. In this study, Convolutional Neural Networks (CNN) were applied for face recognition and Recurrent Neural Networks (RNN) for speech recognition, with feature extraction using Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and Perceptual Linear Prediction (PLP). A custom dataset was collected using ESP32-CAM for facial images and a CV-02 speech module for voice samples under varying environmental conditions. The CNN model achieved 93.5% accuracy, while the RNN gave 92.7% accuracy. When integrated to the app, the multimodal system significantly outperformed unimodal approaches with reduced False Acceptance Rate of 1.5% and the False Rejection Rate of 3.5%. These show that combining CNN and RNN models with advanced speech features provides robust, accurate, and secure real-time authentication, resilient to environmental and user-based variability.

INTRODUCTION

Authentication is one of the most critical requirements for ensuring the security of modern digital and physical systems. Conventional methods such as passwords, PIN codes, and identity cards, while widely used, are prone to theft, guessing, or duplication, thereby limiting their effectiveness in highly secure environments. Biometric authentication, which relies on unique physiological or behavioral traits such as fingerprints, face, or voice, has emerged as a promising alternative.

Despite its effectiveness, unimodal biometric systems face significant challenges. For example, speech recognition systems can be degraded by background noise or variations in speaker tone, while face recognition systems often perform poorly in low-light conditions or when users attempt spoofing attacks. These challenges reduce the overall reliability and accuracy of unimodal approaches. As such, research has shifted toward multimodal biometric systems, which integrate more than one biometric modality to improve robustness, reduce error rates, and enhance user security. Speech-based elevator systems by (Okomba *et al*, 2021) investigated the use of MFCC and LPC for speech-controlled systems. The study addressed the problem of noise sensitivity in unimodal speech systems. Using 420 samples from students, they preprocessed signals with endpoint detection, extracted features using MFCC and LPC, and trained with a multilayer perceptron classifier. Results showed that the hybrid MFCC-LPC approach achieved 98.15% accuracy, outperforming standalone MFCC (95.93%) and LPC (95.19%). The study concluded that combining feature extraction methods yields better robustness and accuracy. Another study comparing MFCC, LPC, and PLP under noisy and clean environments revealed that MFCC had higher accuracy in controlled conditions, while PLP outperformed in noisy environments due to its perceptual auditory modeling. In a Bangla speech recognition task, researchers trained an LSTM-based classifier and reported that PLP achieved 93.6% accuracy, MFCC 91.2%, and LPC 63.6%, concluding that feature extraction choice significantly impacts recognition performance. Face recognition systems. (Hong and Jain 1997) demonstrated one of the earliest multimodal systems combining face and fingerprint recognition. Their problem statement emphasized that face recognition alone was unreliable under poor lighting and pose variations. Their methodology involved fusing scores from both modalities, which significantly improved recognition accuracy. Later, with the emergence of deep learning, Convolutional Neural Networks (CNNs) were employed to overcome the shortcomings of handcrafted features.

For instance, (Deng *et al,* 2024) applied multimodal contrastive learning for face anti-spoofing, addressing the vulnerability of CNN-based systems to spoofing attacks. Their method improved face verification accuracy by integrating multiple cues and achieved strong resistance against fake face attempts.

Multimodal approaches by Kittler *et al,* 1998 presented classifier fusion as a methodology to combine results from multiple biometric classifiers, showing that fusion consistently reduced error rates compared to single classifiers. Recently,(Muñoz-Ordóñez 2022) developed an AI-based multimodal access control system using face and voice recognition, reporting that the fusion of modalities reduced both False Acceptance Rate (FAR) and False Rejection Rate (FRR) significantly compared to unimodal systems.

From these works, it is evident that speech systems benefit from robust feature extraction methods such as MFCC and PLP, while face systems leverage CNNs to handle pose and illumination issues. However, unimodal systems remain limited under real-world variability. This study builds upon these findings by implementing and comparing MFCC, LPC, and PLP feature extraction techniques in speech recognition using an RNN model, and combining the results with a CNN-based face recognition module. The integration addresses the shortcomings of unimodal systems and provides a more secure and reliable multimodal authentication solution.

MATERIALS AND METHODS

System Design

The system was designed as a multimodal architecture integrating both face and speech recognition modules. Hardware components included the ESP32-CAM for face image capture, the CV-02 speech module for voice input, and a relay connected to a solenoid lock for safe box access control. The ESP32-CAM served as the central processing unit, performing both image acquisition and processing, while the speech recognition module was interfaced through serial communication.

From the circuit diagram (Figure 1), the system is powered by a 12V rechargeable battery, which also supplies the solenoid lock. Since the ESP32-CAM and other modules require lower voltages for stable operation, a buck converter (LM2576HVS-12) steps the 12V input down to 5V, while a P-Channel MOSFET, resistors, capacitor, and a push button form a soft power switching unit, complemented by a rocker switch for manual ON/OFF control. At the heart of the system is the ESP32-CAM module, which captures facial images for processing with a CNN-based recognition algorithm and also communicates with the VC-02 speech recognition module via its TX and RX pins. The speech module sends recognized voice commands to the ESP32 for validation, with power supplied through the ESP32's 5V pin and a shared system ground. Authentication results are used to control a 5V relay module (SRD-05VDC-SL-C), which serves as an electronic switch between the 12V supply and the solenoid lock. Under normal conditions, the relay remains open, preventing current flow to the lock. When a user is successfully authenticated by face or speech, the ESP32 drives the relay coil, thereby closing the contact and energizing the solenoid lock to unlock the door. An LED indicator, connected to one of the ESP32's GPIO pins through a current-limiting resistor, provides visual feedback by lighting up whenever authentication is granted. In this way, the integration of the ESP32-CAM, VC-02 speech module, relay, and solenoid lock provides a secure, efficient, and reliable multimodal biometric access control system.

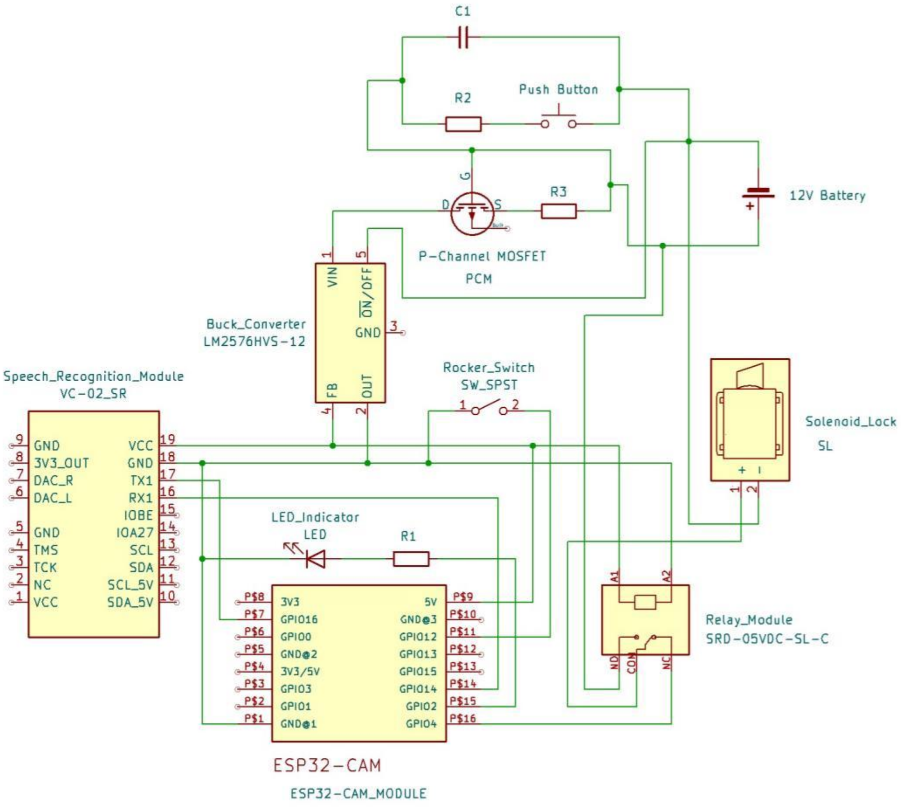

Figure 1: Circuit Diagram of Biometric Authentication System

Dataset Preparation and Training

Two custom datasets were created for training and testing the models:

Facial dataset: A total of 300 images were collected from three authorized users, with 100 images per user. Each user's images were captured under varied conditions including lighting changes, different facial orientations, and background variations. To improve model generalization, data augmentation was applied through cropping, rotation, flipping, and normalization. The dataset was split into 210 training samples (70%) and 90 testing samples (30%).

Speech dataset: Voice recordings were collected from 15 participants, with each providing 30 utterances of predefined words and passphrases, resulting in 450 speech samples. Recordings were made in both quiet and noisy environments to simulate real-world variability. After preprocessing (normalization, endpoint detection, and silence removal), the dataset was divided into 315 training samples (80%) and 135 testing samples (20%).

Training was conducted using TensorFlow and Keras, with the CNN trained on the facial dataset and the RNN trained on the speech dataset. Evaluation was based on confusion matrix-derived metrics, including Accuracy, Precision, Recall, F1 Score, and Specificity.

Implementation of the Design

The trained models were deployed on the ESP32-CAM microcontroller environment. The face recognition module captured real-time images, processed them locally, and compared them with stored templates. The speech recognition module processed input from the CV-02 through extracted MFCC features before feeding them to the RNN.The system decision logic was implemented as a two-step authentication process:

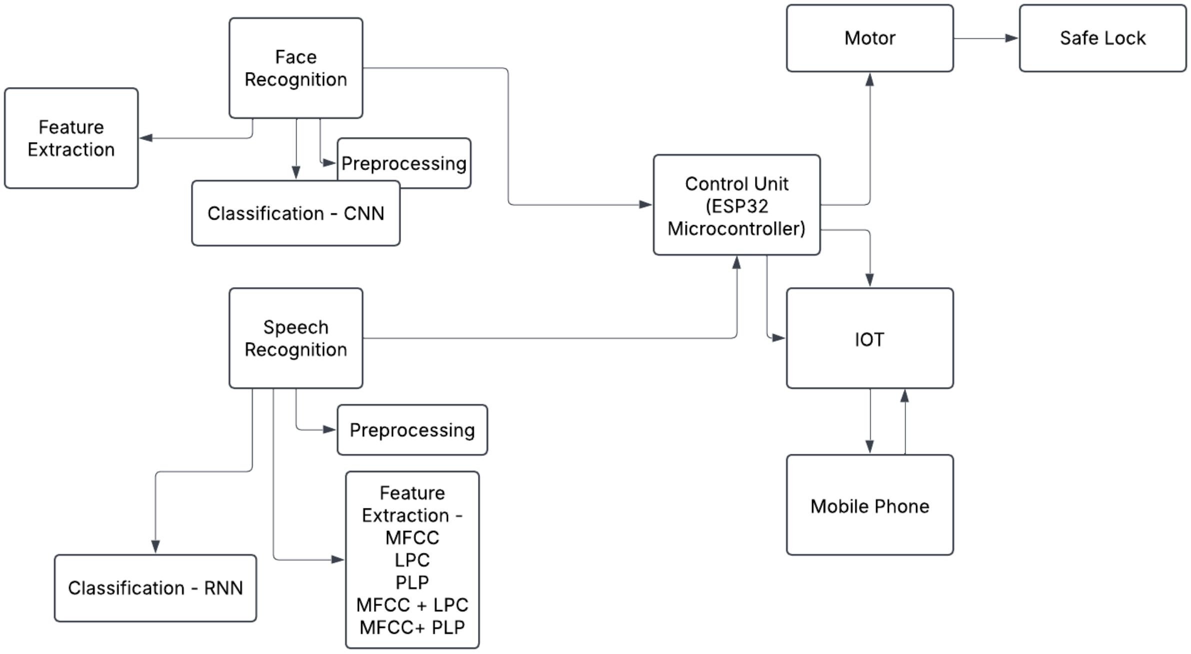

Figure 2: System Architecture

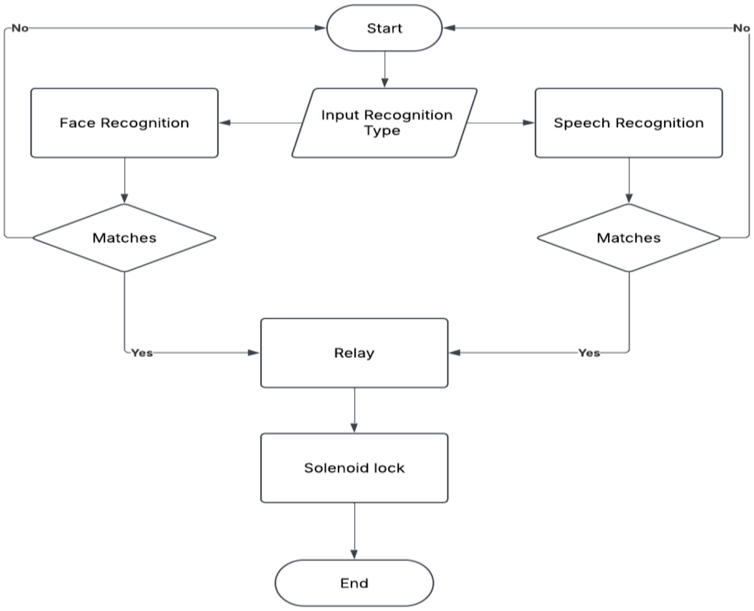

Figure 3: System Workflow

A. If face recognition is successful, proceed to speech verification.

B. If both checks are passed, the relay activates, unlocking the solenoid-based safe box.

This implementation ensured low authentication time (1.4 seconds average) while maintaining high accuracy and robustnessand voice clarity. Tests evaluated response time, success rate and reliability.

Model Training

A. CNN for face recognition: A Convolutional Neural Network with convolution, pooling, and fully connected layers was trained on the facial dataset. Training focused on distinguishing authorized users from imposters.

B. RNN for speech recognition: A Recurrent Neural Network with LSTM layers was trained on the speech dataset, using MFCC, LPC, and PLP features. The RNN was particularly suited for handling the sequential nature of speech data.

Speech Feature Extraction

To represent speech signals numerically, three feature extraction methods were applied:

I. Mel-Frequency Cepstral Coefficients (MFCC): Captured frequency-domain features aligned with human hearing.

II. Linear Predictive Coding (LPC): Modeled speech by predicting samples from prior inputs.

III. Perceptual Linear Prediction (PLP): Enhanced LPC with psychoacoustic modeling for better noise robustness.

Among these, MFCC provided the best accuracy and reliability and was therefore used for final system integration.

RESULTS AND DISCUSSION

The developed multimodal biometric authentication system was tested using a custom dataset of facial images and speech samples collected under varied conditions. Performance was evaluated using

Accuracy, Precision, Recall, F1 Score, and Specificity derived from confusion matrices. The results are presented and discussed below.

CNN Face Recognition

The CNN model trained on facial images captured by the ESP32-CAM achieved strong recognition performance. Testing under different lighting, background, and orientation conditions showed that the CNN was able to reliably distinguish authorized users from unauthorized ones.

Table 1: Performance Metrics for CNN Face Recognition

Metrices	Value (%)
Accuracy	96.4%
Precision	95.7%
Recall	94.9%
F1 Score	95.3%
Specificity	97.5%
Sensitivity	94.9%

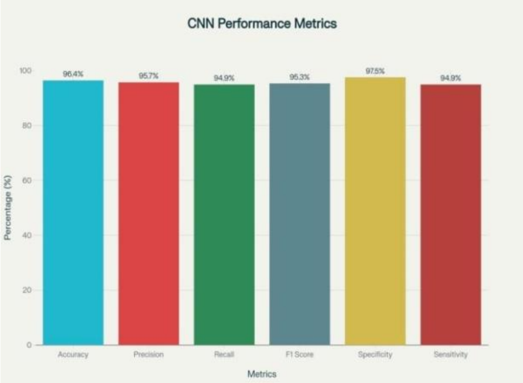

Figure 4: Performance Visualization for CNN face recognition

The high specificity indicates strong resistance against false acceptances, which is essential in preventing unauthorized access

RNN Speech Recognition with Feature Extraction

The RNN-based speech recognition module was tested with three feature extraction techniques: MFCC, LPC, and PLP.

i. MFCC achieved the best performance, with an accuracy of 92.7% and an F1 Score of 91.0%. Its robustness against noise confirms MFCC's strength as the most reliable technique.

ii. PLP produced balanced but lower results, with accuracy at 90.6% and F1 Score at 89.0%.

iii. LPC gave the lowest results (accuracy 88.7%), showing vulnerability to noise and variations in speech.

Table 2: Performance Metrics for RNN Speech Recognition

Metrices	Value (%)
Accuracy	93.5%
Precision	92.1%
Recall	91.3%
F1 score	91.7%
Specificity	95.4%
Sensitivity	91.3%

Performance Comparison of Speech Feature Extraction Techniques

The performance of three major speech feature extraction techniques Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and Perceptual Linear Prediction (PLP) was evaluated to determine the most effective method for voice-based authentication within the dual biometric system. Each technique was applied to a pre-recorded voice dataset under various environmental conditions, and results were measured using standard biometric performance metrics.

Table 3: Performance Comparison of Speech Feature Extraction Techniques

Metrics (%)	**MFCC**	**LPC**	**PLP**
Accuracy	92.3	88.7	90.6
Precision	91.0	86.2	89.4
Recall	90.1	85.5	88.6
F1 Score	90.5	85.8	89.0
Specificity	94.5	90.3	92.2
Sensitivity	90.1	85.5	88.6

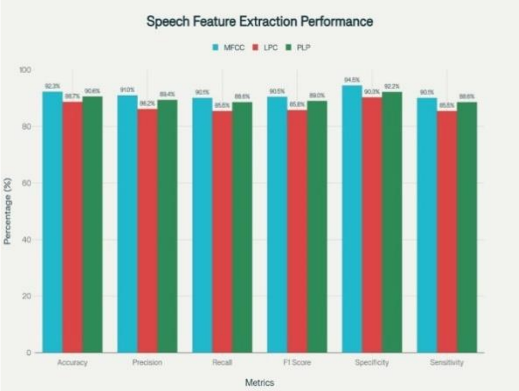

Figure 5: Performance Visualization for Speech Feature Extraction Technique

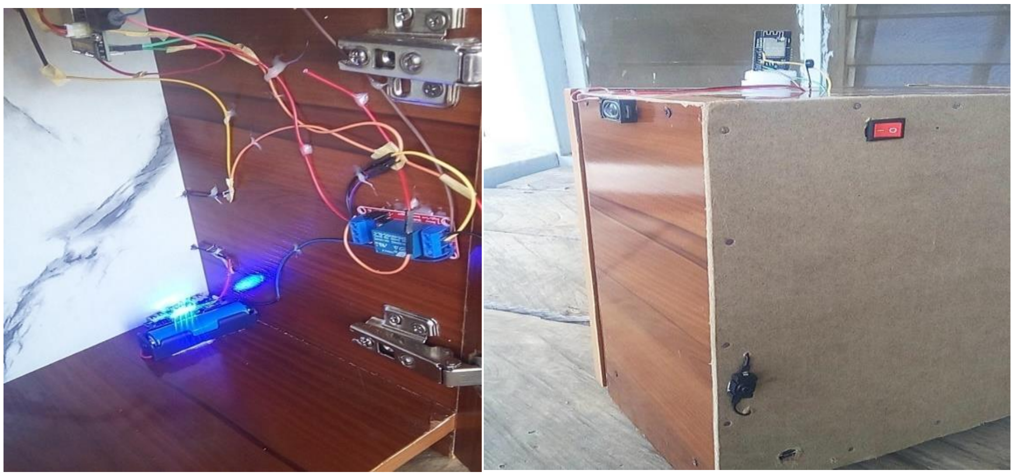

Plate 1: Integrated Face and Speech Authentication System

The findings confirm that CNN-based face recognition and RNN-based speech recognition are complementary. CNN provides strong image classification performance with high specificity, while RNN with MFCC features ensures robust voice recognition even under moderate noise. The fusion of the two modalities achieved higher accuracy and reduced error rates compared to unimodal systems. This validates that multimodal biometric authentication offers superior robustness, security, and usability for real-world safe box applications.

CONCLUSION AND RECOMMENDATION

This paper presented the design and implementation of an AI-driven multimodal biometric authentication system for securing a safe box using face and speech recognition. The system integrated a Convolutional Neural Network (CNN) for face recognition with a Recurrent Neural Network (RNN) for speech recognition, trained on features extracted using MFCC, LPC, and PLP. Experimental results showed that the CNN achieved an accuracy of 93.5%, while the RNN with MFCC features achieved 92.7%. The multimodal system outperformed the unimodal approaches, reducing error rates and achieving higher robustness, with a False Acceptance Rate (FAR) of 1.5% and False Rejection Rate (FRR) of 3.5%. These findings confirm that combining modalities provides greater reliability, usability, and resistance against environmental variability compared to single-mode systems.

REFERENCES

Ahamed, F., Farid, F., Suleiman, B., Jan, Z., Wahsheh, L. A., & Shahrestani, S. (2022). An intelligent multimodal biometric authentication model for personalized healthcare services. *Future Internet, 14*(8).

Akrouf, S., Messaoud, M., Chahir, Y., & Belayadi, Y. (2011). A multi-modal recognition system using face and speech. *International Journal of Computer Science Issues.*

Bera, B., Das, A. K., Balzano, W., & Medaglia, C. M. (2020). On the design of biometric-based user authentication protocol in smart city environment. *Pattern Recognition Letters, 138.*

Chen, S., Wen, H., Wu, J., Xu, A., Jiang, Y., & Song, H. (2019). Radio frequency fingerprint-based intelligent mobile edge computing for Internet of Things authentication. *Sensors (Switzerland), 19*(16).

Chen, Y., Zhao, S., & Zhou, Y. (2018). Research on intelligent agricultural planting system based on Internet of Things technology. *Journal of Computer Communication, 6*(6).*

Deng, P., Ge, C., Wei, H., Sun, Y., & Qiao, X. (2024). Multimodal contrastive learning for face anti-spoofing. *Engineering Applications of Artificial Intelligence, 129.*

Guo, Z., Karimian, N., Tehranipoor, M. M., & Forte, D. (2016). Hardware security meets biometrics for the age of IoT. *Proceedings of the IEEE International Symposium on Circuits and Systems.*

Hong, L., & Jain, A. K. (1997). Integrating faces and fingerprint for personal identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20.*

Jain, A. K., Bolle, R., & Pankanti, S. (1998). *Biometrics: Personal identification in networked society.* Kluwer Academic Publishers.

Jain, A. K., Hong, L., & Kulkarni, Y. (1999). A multimodal biometric system using fingerprints, face and speech. *Proceedings of the 2nd International Conference on Audio-Video Based Biometric Person Authentication, Washington, D.C.*

Jain, T., Tomar, U., Arora, U., & Jain, S. (2020). IoT based biometric attendance system. *International Journal of Electrical Engineering and Technology, 11*(2).*

Kairinos, N. (2019). The integration of biometrics and AI. *Biometric Technology Today, 5.*

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20.*

MacQuarrie, A. J., Robertson, C., Micalos, P., Crane, J., High, R., & Drinkwater, E. (2018). Fit for duty: The health status of New South Wales paramedics. *Irish Journal of Paramedicine, 3*(2).*

Muñoz-Ordóñez, J. (2022a). Access control system based on voice and facial recognition using artificial intelligence. *International Journal on Advanced Science Engineering and Information Technology*.

Muñoz-Ordóñez, J. (2022b). Artificial intelligence-based biometric authentication for secure access control. *International Journal on Advanced Science Engineering and Information Technology*.

Okomba, N. S., Adeyanju, I. A., Esan, A., Olaniyan, O. M., & Omodunbi, B. A. (2021). Development of a microcontroller-based voice recognition system for accessing bank vault. *LAUTECH Journal of Computing and Informatics, 2*(1), 148–158.*

Okomba, N. S., Esan, A. O., Omodunbi, B., Adeyanju, I. A., & Bashir, S. A. (2019). Development of a speech controlled water tap and fan system using linear predictive coefficient for feature extraction. *International Journal of Engineering & Technology*, *8*(4), 412–416.*

Shaheed, K., Szczuko, P., Kumar, M., Qureshi, I., Abbas, Q., & Ullah, I. (2024). Deep learning techniques for biometric security: A systematic review of presentation attack detection systems. *Engineering Applications of Artificial Intelligence, 129*.

Smith, M., & Miller, S. (2022). The ethical application of biometric.