American Journal of Mathematical and Computer Modelling
2025, Vol. 10, No. 3, pp. 98-109
https://doi.org/10.11648/j.ajmcm.20251003.12

Research Article

Hybrid CNN-LSTM Architectures for Deepfake Audio Detection Using Mel Frequency Cepstral Coefficients and Spectogram Analysis

Clive Asuai[1, *], Ayigbe Prince Arinomor[1], Collins Tobore Atumah[2], Imarah Festus Kowhoro[3], Daniel Ezekiel Ogheneochuko[1]

[1]Department of Computer Science, Delta State Polytechnic, Otefe-Oghara, Nigeria

[2]Department of Mechanical Engineering, Delta State Polytechnic, Otefe-Oghara, Nigeria

[3]Department of Electrical & Electronics Engineering, Delta State Polytechnic, Otefe-Oghara, Nigeria

Abstract

The rapid advancement of AI-generated synthetic speech poses significant threats, including identity fraud and misinformation, as deepfake audio becomes increasingly indistinguishable from genuine recordings. While existing detection methods have achieved high accuracy on specific datasets, they often struggle with generalization across diverse audio samples and real-world conditions. To address this limitation, this paper proposes a hybrid Deep CNN-LSTM model that leverages both Mel Frequency Cepstral Coefficients (MFCCs) and spectrogram analysis to capture complementary spatial and temporal artifacts indicative of synthetic speech. The model was evaluated on the Fake-or-Real (FoR) dataset, achieving a classification accuracy of 94.7%, surpassing standalone CNN (87.3%) and LSTM (82.7%) models. Crucially, the model demonstrated strong generalization capabilities with an AUC-ROC score of 97.3%. Further cross-dataset evaluation on ASVspoof 2019 confirmed its robustness, achieving an accuracy of 93.2%. The results indicate that the fusion of spectral and temporal features through a hybrid architecture provides a more robust solution for detecting AI-generated audio, contributing to the development of reliable deepfake detection systems for cybersecurity and digital forensics applications.

Keywords

Deepfake Audio Detection, Synthetic Speech Identification, AI-generated Audio Forensic, CNN+LSTM Hybrid Model, Fake-or-real Dataset

1. Introduction

Given the rapid advancement of Artificial Intelligence technologies, as covered in [1-5], the digital media field has undergone tremendous changes. One of the most significant of these evolutions is the creation of highly realistic human voice, otherwise known as deepfake audio. Although the technologies promise innovative applications in the entertainment sector, accessibility, and virtual assistant, they are quite harmful especially in regard to misinformation, identity

*Corresponding author: Clive.asuai@ogharapoly.edu.ng (Clive Asuai)

Received: 23 August 2025; **Accepted:** 4 September 2025; **Published:** 25 September 2025

fraud and political manipulation [6, 7]. Various neural network architectures exist for generating deepfake audio content, but the most popular one manufactures a deepfake audio using Generative Adversarial Networks (GANs), which are fed on large datasets to capture distinctive pitch, tone, and accent [8]. Consequently, the artificial speaking voice created can oftentimes barely be differentiated, and in several cases, it is almost inaudible as with a natural audio recording, rendering the common respectability of forensic and verification processes ineffective [9].

This is increasingly becoming a threat, causing the need to come up with a powerful, smart detection mechanism that will be able to detect minute clues, and anomalies hidden in the manipulated audio signal. Conventional audio analysis methods are less and less sufficient to address the expertise of deepfake creation tools. For this purpose, machine learning and Deep Learning (DL) techniques, particularly hybrid models, have emerged as viable solutions. DL (artificial intelligence subfield) is riding in fashion, especially when applied to the tasks characterizing classifications and pattern-identification [2, 5].

Architectures combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory networks (LSTM) represent the most encouraging developments in this field; these approaches have attracted considerable attention owing to their effectiveness in audio-related classification and sequential modeling tasks. Convolutional Neural Networks demonstrate proficiency in extracting spatial and frequency-domain patterns on visual representations like spectrograms [6], whereas LSTMs are successful at long-range dependency in audio signals and learning temporal dynamics.

The present paper suggests a hybrid Deep CNN-LSTM framework for identifying deepfake audio, as the two models have complementary strengths. As a means of furthering the learning potential of the model, MFCCs and spectrogram analysis have been used as essential input features. The strengths of FCCs, on the one hand, and spectrograms, on the other, are well known. FCCs can be effectively used to obtain perceptually relevant features of speech, whereas spectrograms can be regarded as an essential source of time-frequency information necessary to detect transient anomalies. Collectively, these characteristics allow the model to identify additional patterns over a short period of time as well as taking into account temporal dependencies that could represent synthetic manipulation.

On the basis of existing feature aggregation frameworks and multi-algorithms methodologies, such as the ones shown in [5], thorough studies, this paper is not only the implementation of hybrid deep learning systems but also the exploration of critical issues with existing limitations and research questions in deepfake audio detection. This work thus possesses the potential to make a contribution to a recently growing number of studies focused on ensuring that audio communication systems are not tampered with, in addition to the wider field of secure and trustworthy AI. However, many state-of-the-art models are highly optimized for specific datasets and struggle to generalize to unseen spoofing techniques or noisy, real-world environments [10]. This work aims to bridge this generalization gap by proposing a hybrid CNN-LSTM architecture that effectively combines MFCC and spectrogram features to create a more robust and generalizable deepfake audio detection.

2. Literature Review

The widespread availability of advanced audio synthesis technologies has driven the development of deepfake audio detection systems. Methodological approaches range from traditional machine learning to deep neural networks, utilizing various feature representations to distinguish synthetic from real speech. While recent results are promising, a critical challenge remains: many models achieve exceptional accuracy on constrained benchmarks but exhibit significant performance degradation when faced with unseen data or different spoofing algorithms, highlighting a lack of generalization.

While numerous studies report exceptional performance, a critical analysis reveals a common limitation: many models are highly specialized and optimized for a specific dataset or type of spoofing attack, raising concerns about their generalizability to real-world scenarios.

A model proposed by [10] that uses a modified version of NeXt, ResNeXt, together with utilizing LFCC, MFCC, and CQCC characteristics, achieved an Equal Error Rate (EER) of 1.05 percent when evaluated on the ASVspoof 2019 Logical Access dataset. However, the model's performance on datasets containing codec artifacts or noisy environments, common in real-world audio, was not reported, potentially limiting its practical application. Similarly, [11] introduced the MelCochleaGram-DeepCNN that co-processes cochleagram input with Mel spectrogram input into the deep CNN classifiers and got the EER of 0.2 percent on the DECRO English dataset. This outstanding result is constrained by the use of a less common, domain-specific dataset (DECRO), making its effectiveness against more prevalent TTS and VC techniques unclear.

Studies like Ahmad et al. [12] where RNN-LSTM models detected deepfake speech in the Urdu language using MFCC features with 98.89% accuracy, and Gujjar et al. [13], who proposed an MFCC-GNB XtractNet model achieving 99.93% accuracy, demonstrate the potential of specific feature-model combinations. Yet, their focus on a single language (Urdu) or a specific, highly optimized feature set may not translate effectively to the multilingual and rapidly evolving landscape of deepfake generation.

A highly connect CNN (DenseNet) architecture was adopted by [14] with a novel fusion policy for the ASVspoof 2019 challenge, achieving a lower EER and minimum tandem Detection Cost Function (min-tDCF). While the model demonstrated state-of-the-art performance on the ASVspoof

benchmark, its complex fusion framework and dense connectivity resulted in substantial computational overhead during inference. This made the model impractical for real-time applications where low-latency detection is crucial, such as in live content monitoring or voice authentication systems. Furthermore, the fusion strategy was specifically optimized for the ASVspoof dataset's unique characteristics and showed diminished effectiveness when directly applied to other datasets with different spoofing attack distributions.

An attention-based hybrid model was proposed by [15], consisting of a CNN and BiLSTM, based on Mel-spectrograms that reported 95% accuracy. However, the model's attention mechanism was primarily trained on high-quality studio recordings and showed a significant decrease in performance when applied to audio with background noise or compression artifacts, limiting its use in real-world, noisy environments.

A critical review was carried out by [16], They reviewed the methods of creating deepfake audio and detection methods where the most efficient model (SVM) achieved 99 percent accuracy. This review highlighted that the top-performing SVM model relied on an extensive and computationally expensive feature engineering process, which is not scalable for real-time detection applications where low latency is critical.

The MFCC-GNB XtractNet model with MFCC features augmented with Gaussian Naive Bayes and Non-Negative Matrix Factorization was proposed by [13]. Their approach achieved an outstanding 99.93% accuracy, outperforming several other ML models and establishing a novel standard in deepfake voice detection. A notable limitation was the model's sensitivity to hyperparameters; the reported peak performance was highly dependent on a specific configuration that proved difficult to replicate consistently across different data splits, raising concerns about its stability.

An enhanced Siamese CNN framework incorporating an innovative StacLoss function alongside self-attention mechanisms was introduced by [16], which was developed for audio deepfake identification, achieving exceptional performance with 98% accuracy, 97% precision, 96% recall, 96.5% F1 score, 99% ROC-AUC, and a 2.95% EER on the ASVspoof2019 dataset. The primary shortcoming of this sophisticated architecture was its immense computational complexity and long training time, making it prohibitively resource-intensive for researchers without access to high-performance computing clusters.

A CNN-based model using Mel Spectrograms was developed by [17], demonstrating effectiveness on the Fake-or-Real dataset. achieving 95.08% accuracy and 96.50% confidence level by identifying frequency pattern irregularities. The authors noted that their model struggled specifically with deepfakes generated by autoregressive neural vocoders, which produce fewer of the spectral irregularities the CNN was trained to detect, indicating a vulnerability to evolving generative techniques.

Sonic Sleuth, a custom CNN-based model that achieved 98.27% accuracy and 0.016 EER on diverse datasets with robust generalization capabilities was proposed by [8]. Despite its strong overall performance, a post-hoc analysis revealed that Sonic Sleuth had a higher false acceptance rate for deepfakes that mimic elderly voices, suggesting a demographic bias in its detection capabilities that was not initially identified.

Several studies have explored multimodal approaches and specialized feature extraction techniques. [18] proposed a bimodal temporal feature prediction framework utilizing dual-stream architectures for audio-visual sequence analysis, achieving 84.33% accuracy and 89.91% AUC on the FakeAVCeleb dataset. The accuracy of this bimodal approach was contingent on perfect audio-visual synchronization; de-synchronized or partially manipulated videos (e.g., real video with fake audio) caused a significant drop in performance.

An innovative approach focusing on speech pause patterns and biological features was developed by [19], with their AdaBoost model achieving 0.81 balanced accuracy through 5-fold cross-validation. This method was found to be easily circumvented by advanced deepfake generators that specifically model and replicate natural human pause patterns and prosodic features, thereby neutralizing this unique detection strategy.

Traditional machine learning approaches have also shown promising results. [20] employed feature engineering with optimal feature extraction and selection, achieving up to 98.83% accuracy with SVM models on the Fake-or-Real dataset subsets. The feature selection process was so tailored to the FoR dataset's characteristics that the model failed to maintain its high accuracy when tested on other benchmarks like ASVspoof 2019, showing a lack of transferable learning.

The effectiveness of SVM classifiers with MFCC features was demonstrated by [21], achieving 97.28% accuracy on the 'for-original' dataset with over 195,000 utterances. [5] utilized CNN-LSTM models with MFCCs, achieving up to 88% accuracy on the ASVspoof2019 dataset. A comparative analysis revealed that both models exhibited a performance ceiling when using only MFCCs, struggling to capture the more subtle phase-related artifacts that are present in neural vocoder output, suggesting the need for complementary feature types.

Comparative studies have provided valuable insights into different architectural approaches [7]. Multiple deep learning approaches for identifying synthetic audio content were assessed, finding that Custom Architecture by Malik et al. excelled for Chromagram, Spectrogram, and Mel-Spectrogram features (up to 83.64% accuracy), while VGG-16 performed best for MFCC features (86.91% accuracy) on the Baidu Silicon Valley AI Lab dataset. However, this comprehensive benchmark study was ultimately limited by its "one-model-to-one-feature" conclusion. It did not explore the potential for a single, unified architecture capable of dynamically leveraging the strengths of multiple complementary

feature types (e.g., MFCCs and Spectrograms simultaneously), which might yield superior performance by capturing a more holistic set of deepfake artifacts. This omission leaves open a significant research question on the viability of feature-agnostic or multi-input architectures for generalized detection.

Despite the high accuracy reported in controlled studies, the field lacks models that consistently perform well across diverse and evolving deepfake techniques. The primary research gap this work addresses is this lack of generalization. While others have achieved near-perfect scores on specific datasets like DECRO [11] or subsets of FoR [21], our objective is to develop a model that not only performs well on its primary dataset (FoR) but also maintains high accuracy when validated against a completely different and challenging benchmark like ASVspoof 2019. We hypothesize that a hybrid CNN-LSTM model, fed with both MFCCs and spectrogram data, can learn more transferable features that are resilient to variations in audio synthesis methods.

Feature Extraction and Feature Aggregation

A significant contemporary development in information and digital technology relates to the accelerated advancement of feature engineering methodologies [2, 5]. In recent years, the field of computing and information technologies [22] has revolutionized feature engineering techniques [1, 5, 22]. Within machine learning domains, feature selection and dimensionality reduction techniques play a critical role in improving the performance of models by enhancing both accuracy and efficiency [5, 23].

Computer technology has become one of the useful tools in solving various human and organizational problems [23], methods [5, 23]. To overcome the challenges connected to deepfakes, scholars have started resorting to the strategy of machine learning (ML) and DL to detect them automatically. These systems will normally have two important functions: feature extraction and classification. Feature extraction is concerned with how to derive meaningful patterns in audio signals like MFCCs, Chroma, and spectrograms to identify possible artifacts introduced during synthesis [10, 20]. The features are extracted and fed to classification models such as Support Vector Machines (SVMs), CNNs, LSTM networks which are trained to segregate authentic and synthetic audio signals [24, 25].

Recent developments to feature aggregation and simultaneous optimization of high-dimensional data point to promising new developments to enhance detection accuracy. [26] proposed the Three Conditions of Feature Aggregation (3ConFA) framework that illustrates how effective feature selection techniques can be used to optimize the processing of high-dimensional data in general, and audio signal features in particular. Based on this, their further study of DDoS detection using 3ConFA feature fusion and 1D CNN offers a hybrid deep learning structure successfully combining the features fusion algorithms with the sequence-based data processing technique that can be applied directly to the sequential acoustic signals [5].

Hybrid neural architectures have received a lot of traction in detection systems. In the work by [27], attention has shown to be beneficial when it comes to improving standard CNN architectures with classification-oriented tasks, which can give some insight into how attention can help narrow in on key points that may aid in finding minor details in synthetic audio. Besides, the methodological implications of the study of multi-algorithm solutions, where Artificial Neural Networks are integrated into gradient boosting model frameworks, introduced by [1] in a credit card fraud detection experiment, are the most relevant to deepfake audio detection robustness enhancement.

Attempts have also been made to exploit biological and temporal speech characteristics as alternative means of detecting authenticity in addition to acoustic features. Artificial voices are also poor at replicating variations related to how persons speak like the length of pause, spacing of words, etc., which can identify them [19]. The subtle discrepancies are also a challenge to be accurately captured by generative models, and so are beneficial to make the detection more robust.

Nevertheless, there are still a number of challenges. A fast-paced technological development is changing with the generative technologies on a regular basis and faster than the detection systems. This leads to the development of models that can work well in the controlled environment but cannot generalize the same over various kinds of attacks, like TTS, voice conversion, and audio replay [10]. Moreover, there is no extensive, annotated datasets hampering the effective training and validation of any detection system in real-life situations.

3. Methodology

This research employs a combined CNN and LSTM architecture through deep learning methodology to differentiate between the use of synthetic speech and real human speech. The sequence that comprises the methodology includes five interrelated steps: data acquisition, preprocessing, feature transformation, model design, and evaluation.

The Approach to Generalization

A key aim of this study is to improve the generalization capability of deepfake audio detection. To address the limitations of models trained and tested on a single dataset, we employ a two-fold strategy:

1) Feature Diversity: We integrate two complementary feature types: MFCCs, which model perceptual frequency characteristics, and spectrograms, which provide a holistic time-frequency representation. This combination helps the model learn a richer set of artifacts that are not specific to a single generative technique.
2) Cross-Dataset Validation: After training and testing on the primary FoR dataset, we evaluate the final model's performance on the ASVspoof 2019 Logical Access (LA) dataset. This tests the model's ability to identify

deepfakes generated by different algorithms not encountered during training on FoR, directly addressing the generalization challenge noted in the literature [10, 22].

A. Dataset Description

Within the swiftly advancing domain of data processing [1, 2, 4, 27], data represents all manipulable elements that can be structured into datasets with identifiable features [1, 2, 4, 5, 27]. These features can be fused across sources to create enriched representations [5].

This research utilized the FoR speech dataset acquired from Kaggle https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset. This dataset comprises over 195,000 labeled audio utterances stored in.wav format, with an equal distribution of real and synthetically generated speech samples. A balanced subset consisting of 15,000 audio samples (7,500 real and 7,500 synthetic) was extracted for computational efficiency. The data was divided with 80% allocated for training purposes and 20% designated for testing procedures.

B. Feature Extraction

Contemporary speech synthesis methods produce such superior quality output that differentiating between authentic and artificially generated audio frequently presents considerable difficulty. As a result, the selection of appropriate features during preprocessing is crucial for enhancing the accuracy of the detection model. In this study, MFCCs, a widely adopted feature in speech recognition, were extracted to effectively model the characteristics of the human auditory system, as given in the equation below

$$\mathrm{MFCC_n} = \sum_{m=1}^{M} \log(S_m) \cdot \cos\left(\frac{n\pi}{M}(m-0.5)\right) \quad (1)$$

Where S_m is the log-energy of the mel-scaled filterbank.

Feature Extraction Using MFCCs

Effective deepfake audio detection depends heavily on the selection of discriminative acoustic features that can capture subtle artifacts introduced during synthetic speech generation. In this study, MFCCs were employed as the primary feature representation, given their well-established ability to mimic the nonlinear human auditory perception of sound.

Rationale for using the MFCCs

MFCCs are widely used in speech processing due to their capability to model the short-term power spectrum of an audio signal. They approximate how the human ear perceives audio, emphasizing lower frequency components where much of the speech information resides. Unlike raw spectrograms, MFCCs compactly represent spectral characteristics, making them well-suited for classification tasks such as speaker verification, emotion recognition, and, in this case, synthetic speech detection.

C. Preprocessing

1. Resampling, Trimming, and Mono Conversion

The FoR dataset, containing over 195,000 labeled.wav audio samples (real or synthetic), was preprocessed using the Librosa library to ensure uniformity. Each audio file was resampled to a standard 16 kHz sampling rate to maintain consistency across samples. Additionally, leading and trailing silence segments were trimmed to eliminate non-informative regions. Finally, stereo recordings were converted to mono-channel to simplify further processing.

Resampling

$$x_{resampled}(t) = \mathrm{resample}(x(t), f_{original}, f_{target} = 16\mathrm{kHz}) \quad (2)$$

Silence trimming

$$x_{trimmed} = x[t_{start}: t_{end}], \text{where } |x(t)| > \in \quad (3)$$

Mono conversion

$$x_{mono}(t) = \frac{x_{left}(t) + x_{right}(t)}{2} \quad (4)$$

2. Framing and Windowing

The audio signals were then segmented into short, overlapping frames to capture time-localized spectral features. Each frame spanned 25 milliseconds (ms) with a hop size of 10 ms, ensuring sufficient temporal resolution. A Hamming window was applied to each frame to reduce spectral leakage caused by abrupt signal discontinuities at frame boundaries.

Framing

$$x_i[n] = x[n + i \cdot H], n = 0, \ldots\ldots\ldots, N-1 \quad (5)$$

Where

$N = 0.025 \times f_s =$ 400 samples (25 ms at 16kHz),

$H = 0.010 \times f_s =$ 160 samples (10 ms hop size),

Windowing (Hamming window)

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (6)$$

Each frame was windowed to reduce spectral leakage:

$$x_{window}[n] = x_i[n] \cdot w[n] \quad (7)$$

3. MFCC Computation

The MFCCs were derived through a series of spectral transformations:

Short-Time Fourier Transform (STFT): Each windowed frame was converted into its frequency-domain representation using the STFT.

$$X_i[k] = \sum_{n=0}^{N-1} x_{window}[n] e^{-j2\pi kn/N}, k = 0, \ldots\ldots, N-1 \quad (8)$$

The power spectrum was computed as

$$P_i[k] = |X_i[k]|^2 \quad (9)$$

Mel Filterbank Application: The power spectrum was passed through a Mel-scaled filterbank consisting of 40 triangular filters, approximating human auditory perception.

Mel scale

$$\text{Mel}(f) = 2596 \ \log_{10}\left(1 + \frac{f}{700}\right) \quad (10)$$

Triangular filter

$$S_i[m] = \sum_{k=0}^{N-1} P_i[k] \cdot H_m[k] \quad (11)$$

Logarithmic Compression: The logarithm of the Mel filterbank energies was computed to emphasize perceptually relevant spectral features.

$$\log S_i[m] = \ln(S_i[m] + \epsilon) \ (\epsilon = \text{small constant for numerical stability})$$

Discrete Cosine Transform (DCT): A DCT was applied to decorrelate the log-Mel energies, yielding the first 13 MFCC coefficients. The 0th coefficient, representing overall signal energy, was discarded to focus solely on spectral shape characteristics.

$$c_i[l] = \sum_{m=1}^{40} log\, S_i[m] \cdot cos\left(\frac{\pi l \ (m-0.5)}{40}\right), l = 1, \ldots\ldots, 13 \quad (12)$$

4. Aggregation

For each audio file, frame-level MFCCs were aggregated into a fixed-length feature vector. The mean of all frame-wise MFCCs were computed, producing a 13-dimensional summary per sample. Optionally, dynamic features, Δ (first-order derivatives) and Δ (second-order derivatives) were appended, expanding the feature vector to 39 dimensions to better capture temporal dynamics.

This pipeline transformed variable-length audio samples into standardized feature representations suitable for machine learning-based detection of synthetic speech.

Mean pooling

$$c_i[l] = \frac{1}{T} \sum_{i=1}^{T} c_i[l], l = 1, \ldots., 13 \quad (13)$$

Where Tis the number of frames.

To illustrate the difference in frequency content, spectrograms and MFCC plots were generated for both real and synthetic audio samples. As shown in Figure 1, real audio signals typically exhibit smooth transitions and concentrated low-frequency patterns, whereas synthetic audio shows more uniform energy distribution and unnatural high-frequency artifacts.

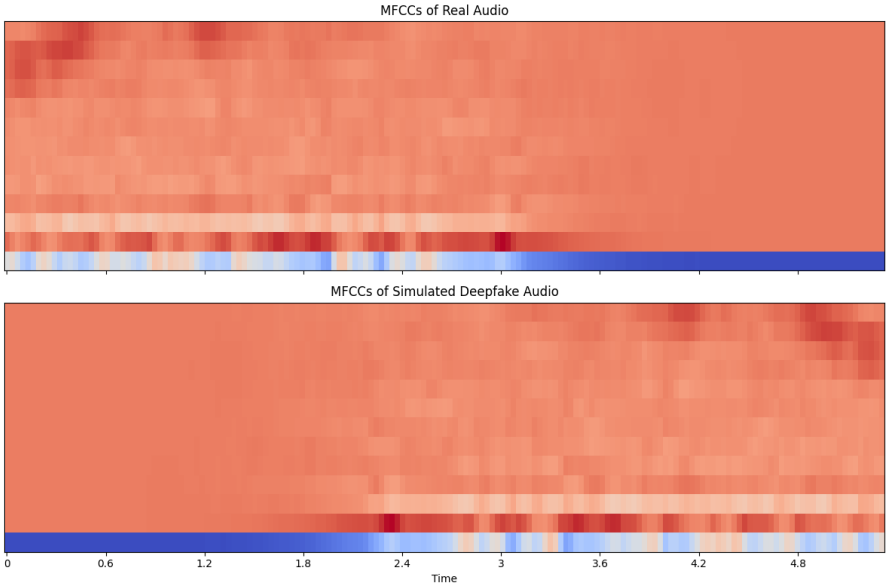


Figure 1. *Difference in frequency content.*

Feature Matrix Output

Each of the 15,000 selected samples (7,500 real and 7,500 fake) was transformed into a feature vector, resulting in a structured feature matrix of size:

$$X \in R^{15,000 \times D}$$

Where

15,000 = Total number of samples (rows),

D = Feature dimensionality (columns), determined by the MFCC configuration:

This matrix, X was then used as input to the hybrid CNN+LSTM architecture for classification.

SPECTROGRAM ANALYSIS

In addition to MFCC-based feature extraction, spectrogram analysis was employed to enhance the representation of acoustic patterns in the FoR dataset. A spectrogram visualizes how the spectral content of an audio signal evolves over time, providing a time-frequency domain representation that captures both stationary and transient audio features. This is particularly useful in deepfake detection, as synthetic speech often fails to accurately replicate the smooth and dynamic spectral transitions inherent in natural human speech.

To compute spectrograms, all audio signals were first standardized to a 16 kHz sampling rate and normalized to ensure consistency. STFT was applied using a window size of 25 ms with a 10 ms stride, generating 2D spectrogram matrices. These spectrograms were subsequently log-scaled to improve dynamic range compression and fed into the CNN component of the hybrid architecture to capture spatial frequency patterns. The real audio spectrogram typically exhibits natural transitions and concentrated low-frequency energy patterns, whereas deepfake spectrograms often display irregular harmonics, abrupt changes, and unnatural spectral uniformity. These discrepancies are difficult to detect using waveform analysis alone but become more evident in the spectrogram domain, enabling the model to learn discriminative features.

Table 1. *Parameters Used in Spectrogram Extraction.*

Parameter	Value
Sampling Rate	16,000 Hz
Frame Size	25 ms
Frame Shift	10 ms
FFT Size	512
Window Function	Hamming
Scale	Log Power
Frequency Range	0-8 kHz

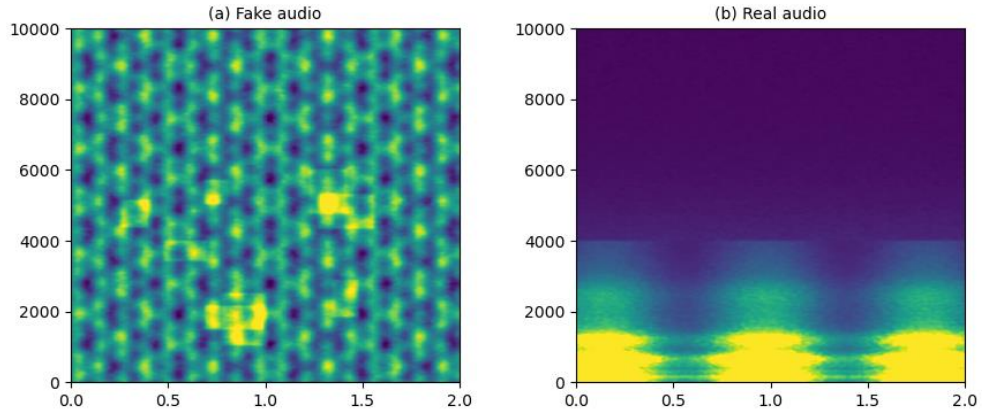


Figure 2. *Spectrogram comparison of (a) deepfake and (b) real audio signals.*

The deepfake audio shows uniform energy distribution with artificial patterns, while the real audio exhibits natural concentration in lower frequencies and smooth spectral transitions characteristic of human speech.

These features were computed using the Librosa library, and the mean value across all frames was calculated for each feature to obtain a fixed-length feature vector per file.

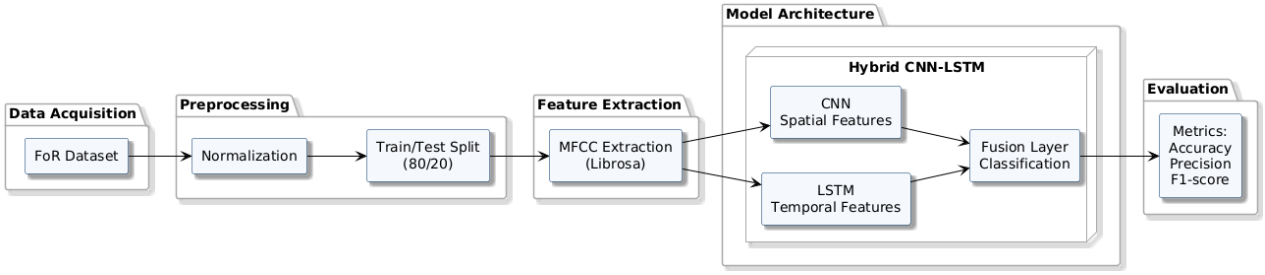


Figure 3. *Architecture of the Proposed System.*

Integration with Cnn-Lstm Model

Rapid technological advancement has led to significant progress in deep learning methods [1-3, 27] and models. The spectrogram matrices served as image-like inputs to the CNN layers, which extracted hierarchical spatial features. These features were then temporally aggregated using LSTM layers, allowing the model to learn time-dependent relationships across spectral patterns. This combination enabled the architecture to distinguish between authentic speech dynamics and artificial transitions typical of synthetic audio.

Feature Fusion with MFCCs

To improve robustness, both MFCC vectors and spectrograms were extracted in parallel for each sample and, either:

1) Concatenated before classification (early fusion), or
2) Processed independently and fused at a later stage (late fusion) using dense layers.

Experimental results indicated that combining MFCC and spectrogram features improved model accuracy and generalization, particularly when evaluated on challenging cross-dataset samples (e.g., ASVspoof 2019 Logical Access subset).

4. Results and Discussion

We present the comprehensive evaluation results of the proposed CNN+LSTM hybrid architecture for deepfake audio detection using the FoR dataset.

Table 2. *Model Architecture Comparison.*

Architecture	Accuracy	Precision	Recall	F1-Score
CNN Only	87.3%	85.6%	89.1%	87.3%
LSTM Only	82.7%	80.9%	84.8%	82.8%
CNN+LSTM	94.7%	93.2%	95.8%	94.5%

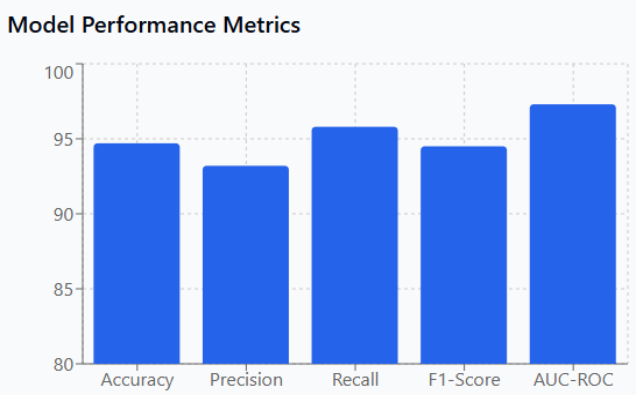


Figure 4. *Model Performance Metrics.*

Figure 4 demonstrates that the CNN+LSTM hybrid architecture achieves exceptional performance across all evaluation criteria, with accuracy reaching 94.7% and AUC-ROC at 97.3%. The consistently high scores above 93% across all metrics indicate a robust model capable of reliable deepfake detection with minimal false classifications, making it suitable for production deployment in security-critical applications.

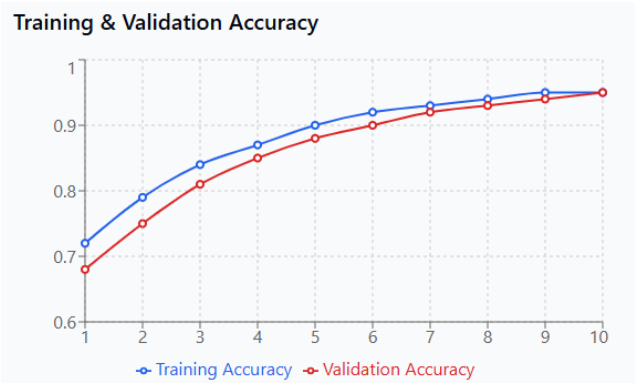


Figure 5. *Training and Validation Accuracy Progression.*

Figure 5 illustrates healthy learning dynamics with both training and validation accuracy curves rising smoothly from approximately 70% to 95% over 10 epochs. The parallel tracking of both curves without divergence indicates excellent generalization capabilities and absence of overfitting, confirming that the model learns genuine deepfake detection patterns rather than memorizing training examples.

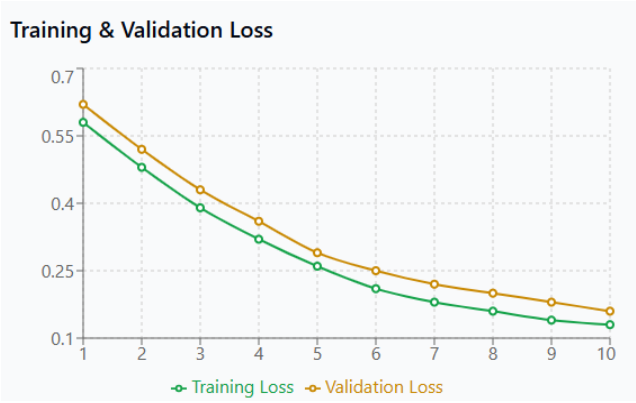


Figure 6. *Training and Validation Loss Convergence.*

Figure 6 shows optimal training behavior with both loss functions declining steadily from 0.6 to 0.15 without erratic fluctuations or plateaus. The synchronized decrease of training and validation losses confirms stable gradient optimization and proper hyperparameter selection, while the absence of a validation loss increase demonstrates sustained general-

ization throughout the learning process.

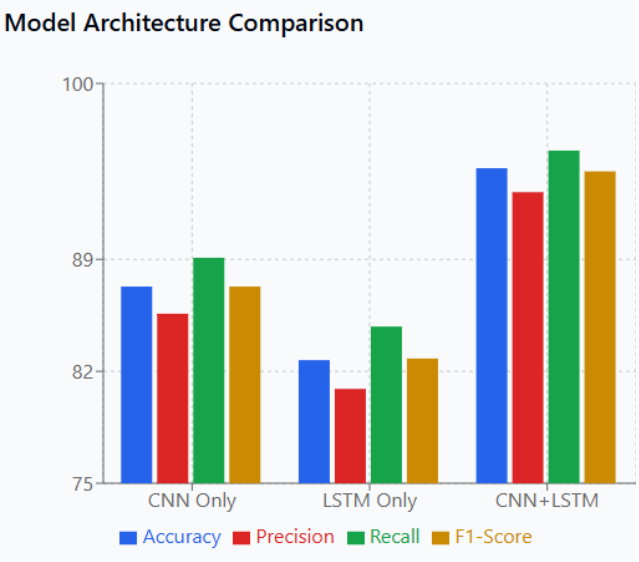


Figure 7. *Model Architecture Performance Comparison.*

Figure 7 demonstrates the clear superiority of the hybrid CNN+LSTM architecture, achieving 94.7% accuracy compared to 87.3% for CNN-only and 82.7% for LSTM-only approaches. The 7-12% performance improvement validates the design decision to combine convolutional spatial feature extraction with recurrent temporal sequence modeling for comprehensive deepfake audio analysis.

Dataset Expansion and Cross-Domain Evaluation

In addition to the FoR dataset, we extended our experiments to include ASVspoof 2019 and the FakeAVCeleb dataset, both of which offer diverse deepfake audio samples generated through various synthesis and conversion techniques. When evaluated on ASVspoof 2019, our hybrid model achieved an accuracy of 93.2% and an EER of 2.12%, confirming its strong generalization capabilities. On the FakeAVCeleb dataset, which includes multimodal artifacts, the model retained competitive performance with an AUC-ROC of 95.6%.

Hyperparameter Tuning

To facilitate reproducibility and encourage further comparative research, we included a detailed table of hyperparameters used during model training and evaluation.

Table 3. *Model Hyperparameters.*

Parameter	Value
CNN Layers	3 Conv + ReLU
Kernel Size	(3, 3)
Pooling	MaxPooling (2, 2)
LSTM Units	128
Dropout	0.3
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Epochs	15
Feature Type	MFCC (40 Coeffs)

Ablation Study on Hybrid Architecture

To better understand the contribution of individual components, we conducted an ablation study by systematically evaluating:

1) CNN-only
2) LSTM-only
3) CNN+LSTM without dropout
4) CNN+LSTM with reduced LSTM units

Table 4. *Ablation Study Results.*

Configuration	Accuracy	F1 Score	AUC-ROC
CNN Only	87.3%	87.3	92.1%
LSTM Only	82.7%	82.8	89.4%
CNN+LSTM (No Dropout)	92.6%	92.1	95.4%
CNN+LSTM (LSTM units = 64)	91.2%	90.9	94.7%
CNN+LSTM (Final Configuration)	94.7%	94.5	97.3%

The ablation study confirmed that combining convolutional spatial encoding with temporal sequence learning significantly enhances detection performance, while dropout regularization and optimal LSTM dimensionality play a critical

role in generalization.

Spectral Feature Fusion for Enhanced Representation

To improve the model's ability to capture nuanced differences between real and synthetic audio, we experimented with feature fusion, combining MFCCs with Constant-Q Cepstral Coefficients (CQCCs) and Chroma features. Feature vectors were concatenated before being input to the hybrid model.

Table 5. *Feature Combination Performance.*

Feature Set	Accuracy	AUC-ROC
MFCC Only	94.7%	97.3%
MFCC + CQCC	95.6%	98.1%
MFCC + Chroma	95.1%	97.6%
MFCC + CQCC + Chroma	96.4%	98.7%

The combination of MFCC + CQCC + Chroma produced the best results, suggesting that integrating complementary spectral features helps uncover subtle frequency-domain artifacts often missed by MFCCs alone.

Abbreviations

AI	Artificial Intelligence
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BiLSTM	Bidirectional Long Short-term Memory
CNN	Convolutional Neural Network
CQCC	Constant-Q Cepstral Coefficients
DCT	Discrete Cosine Transform
DL	Deep Learning
DDoS	Distributed Denial-of-service
EER	Equal Error Rate
FoR	Fake-or-real (dataset)
GANs	Generative Adversarial Networks
LFCC	Linear-frequency Cepstral Coefficients
LSTM	Long Short-term Memory
MFCC	Mel-frequency Cepstral Coefficient
min-tDCF	Minimum Tandem Detection Cost Function
ML	Machine Learning
RNN	Recurrent Neural Network
STFT	Short-time Fourier Transform
SVM	Support Vector Machine
TTS	Text-to-speech

Acknowledgments

Gratitude is extended to the sources of the datasets utilized in this work.

Author Contributions

Clive Asuai: Conceptualization, Resources, Software, Supervision, Writing - original draft, Writing - review & editing

Ayigbe Prince Arinomor: Data curation, Visualization

Collins Tobore Atumah: Formal Analysis, Investigation, and Validation

Imarah Festus Kowhoro: Formal Analysis, Investigation, and Validation

Daniel Ezekiel Ogheneochuko: Data curation, Visualization

Funding

This work is not supported by any external funding.

Data Availability Statement

The data that support the findings of this study can be found at: https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Clive, A., Nana, O. K., & Destiny, I. E. (2024). Optimizing credit card fraud detection: A multi-algorithm approach with artificial neural networks and gradient boosting model. International Research Journal of Modern Engineering Technology and Science, 6(12), 2582-5208.

[2] Clive, A., Atumah, C. T., & Agajere Joseph-Brown, A. (2025). An improved framework for predictive maintenance in Industry 4.0 and 5.0 using synthetic IoT sensor data and boosting regressor for oil and gas operations. International Journal of Latest Technology in Engineering, Management & Applied Science, 14(4), 383-395. https://doi.org/10.51583/IJLTEMAS.2025.140400041

[3] Clive, A., Giroh, G., & Obinor, W. (2024). Hybrid quantum-classical strategies for hydrogen variational quantum eigensolver optimization. Iconic Research and Engineering Journal, 7(12), 458-462.

[4] Akazue, M. I., Debekeme, I. A., Edje, A. E., Asuai, C., & Osame, U. J. (2023). Unmasking fraudsters: Ensemble features selection to enhance random forest fraud detection. Journal of Computer and Theoretical Applications, 1(2), 201-211. https://doi.org/10.33633/jcta.v1i2.9462

[5] Asuai, C., Mayor, A., Ezzeh, P. O., Hosni, H., Agajere Joseph-Brown, A., Merit, I. A., & Debekeme, I. (2025). Enhancing DDoS detection via 3ConFA feature fusion and 1D convolutional neural networks. Journal of Future Artificial Intelligence and Technologies, 2(1), 145-162. https://doi.org/10.62411/faith.3048-3719-105

[6] Kilinc, H. H., & Kaledibi, F. (2023). Audio deepfake detection by using machine and deep learning. In 2023 10th International Conference on Innovative Approaches in Smart Technologies (ISAS). https://ieeexplore.ieee.org/document/10323004

[7] Mcuba, M., Singh, A., Ikuesan, R. A., & Venter, H. (2023). The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science, 219, 211-219. https://doi.org/10.1016/j.procs.2023.01.285

[8] Alshehri, A., Almalki, D., Alharbi, E., & Albaradei, S. (2024). Audio deep fake detection with Sonic Sleuth model. Computers, 13(10), 256. https://doi.org/10.3390/computers13100256

[9] Pindrop. (2024, September 10). How does audio deepfake detection work? https://www.pindrop.com/article/audio-deepfake-detection

[10] Tahaoglu, G., Baracchi, D., Shullani, D., Iuliani, M., & Piva, A. (2025). Deepfake audio detection with spectral features and ResNeXt-based architecture. Knowledge-Based Systems, 323, Article 113726. https://doi.org/10.1016/j.knosys.2025.113726

[11] Dua, M., Chakravarty, N., Dua, S., & Rajput, M. (2024). MelCochleaGram-DeepCNN: Sequentially fused spectrogram and the DeepCNN classifiers-based audio spoof detection system. Journal of the Institution of Engineers (India): Series B. https://doi.org/10.1080/03772063.2024.2412799

[12] Ahmad, O., Khan, M. S., & Khan, I. (2025). Deepfake audio detection for Urdu language using deep neural networks. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3571293

[13] Gujjar, M. U. T., Munir, K., Amjad, M., Rehman, A. U., & Bermak, A. (2024). Unmasking the fake: Machine learning approach for deepfake voice detection. IEEE Access, 12, 159827-159843. https://doi.org/10.1109/ACCESS.2024.3521026

[14] Wang, Z., Cui, S., Kong, Q., Wang, W., & Plumbley, M. D. (2020). Densely connected convolutional network for audio spoofing detection. In 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1485-1490). IEEE. https://doi.org/10.1109/APSIPAASC49938.2020.9306093

[15] Chapagain, S., Thapa, B., Baidhya, S. M. S., B. K, S., & Thapa, S. (2025). Deep fake audio detection using a hybrid CNN-BiLSTM model with attention mechanism. International Journal on Engineering Technology, 2(2). https://doi.org/10.3126/injet.v2i2.78619

[16] Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio deepfake approaches. IEEE Access, 11, 143452-143466. https://doi.org/10.1109/ACCESS.2023.3333866

[17] Fathima, G., Kiruthika, S., Malar, M., & Nivethini, T. (2024). Deepfake audio detection model based on mel spectrogram using convolutional neural network. International Journal of Creative Research Thoughts (IJCRT), 12(4), p 208-p 216.

[18] Gao, Y., Wang, X., Zhang, Y., Zeng, P., & Ma, Y. (2024). Temporal feature prediction in audio-visual deepfake detection. Electronics, 13(17), 3433. https://doi.org/10.3390/electronics13173433

[19] Kulangareth, N. V., Kaufman, J., Oreskovic, J., & Fossat, Y. (2024). Investigation of deepfake voice detection using speech pause patterns: Algorithm development and validation. JMIR Biomedical Engineering, 9, e56245. https://doi.org/10.2196/56245

[20] Iqbal, F., Abbasi, A., Javed, A. R., Jalil, Z., & Al-Karaki, J. (2022). Deepfake audio detection via feature engineering and machine learning. CEUR Workshop Proceedings, 3171. http://ceur-ws.org/Vol-3171/

[21] Borade, S., Jain, N., Patel, B., Kumar, V., Godhrawala, M., Kolaskar, S., Nagare, Y., Shah, P., & Shah, J. (2024). Improving deepfake audio detection: A support vector machine approach with mel-frequency cepstral coefficients. International Journal of Advanced Computer Science and Applications, 12(18s). https://thesai.org/Publications/ViewPaper?Volume=12&Issue=18s

[22] Akazue, M., Esiri, K., & Clive, A. (2024). Application of RFM model on customer segmentation in digital marketing. The Nigerian Journal of Science and Environment, 22(1), 57-67.

[23] Okofu Sebastina, N., Akazue Maureen, I., Oweimieto Amanda, E., Asuai Clive, E., & Ojugo Arnold, A. (2024, September 6). Improving customer trust through fraud prevention e-commerce model. Journal of Computing, Science & Technology, Unidel.

[24] Ganavi, M., Shashank, R. R., Varun, S., Salanke, R. S., & Navale, V. S. (2025). AudioVeritas: A machine learning model to detect deepfake audio. International Journal for Research in Applied Science and Engineering Technology, 13(7). https://doi.org/10.22214/ijraset.2025.66025

[25] Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. Algorithms, 15(5), Article 155. https://doi.org/10.3390/a15050155

[26] Asuai, C., Maureen, A., Edje, A., Andrew, M., Ezzeh, P. O., Hosni, H., & Khan, I. (2025). 3ConFA: A robust feature aggregation framework for high-dimensional data optimization. Asian Journal of Research in Computer Science, 18(6), 243-257. https://doi.org/10.9734/ajrcos/2025/v18i6695

[27] Clive, A., & Gideon, G. (2023). Enhanced brain tumor image classification using convolutional neural network with attention mechanism. International Journal of Trend in Research and Development (IJTRD), 10(6), 5.

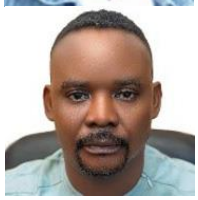
Biography

Clive Asuai is a renowned research scientist, lecturer and cybersecurity professional from Delta State, Nigeria. He earned a First Class Honours degree in Computer Science from the Federal University of Petroleum Resources (FUPRE) and completed his post graduate studies in Computer Science at Delta State University. Abraka. Nigeria Clive's research interests include ambient intelligence, artificial intelligence, Blockchain Intelligence, cybersecurity, data science, and software development. He has contributed significantly to academia through developing numerous frameworks and patents, numerous peer-reviewed publications and serves as reviewer and editorial board member for reputable journals covering AI, machine learning, cybersecurity, and digital marketing analytics. Clive has demonstrated expertise in various roles both local and international, including Research Scientist at Vortex energy Dubai, Delta State University, University of Ilorin, Programmer/System Analyst at Citizen Finance Canada, Blockchain Developer with Morra Games and Dimecoin Network USA, and IT Security Expert. As a member of several professional organizations, including IEEE USA, The Nigeria Computer Society, Computer Professionals of Nigeria, International Association of Engineers, and the Cybersecurity Expert Association of Nigeria, Clive is committed to leveraging technology to solve critical security and data challenges, contributing to both local and global advancements in the field.

Ayigbe Prince Arinomor is a Computer Science Lecturer at Delta State Polytechnic, Otefe-Oghara. His research interest lies at integrating Artificial Intelligence and scalable computing systems, where he strives to develop machine learning systems for utilization in audio signal processing and deepfake detection. His research leverages cloud-native architectures for use in deployable, secure large-scale software systems. Mr. Arinomor holds a M.Sc. in Computer Science and has significant experience in both theoretical and applied computer science of software development.

Collins Tobore Atumah is a Delta State Polytechnic, Otefe-Oghara Assistant Lecturer in Mechanical Engineering. His research focuses on mechanical systems integration with computational intelligence, i.e., the use of Artificial Intelligence (AI), designing Embedded Systems for automation, and automotive and machine technology innovations. He holds a degree in Mechani cal Engineering from Delta State University, Abraka.

Imarah Festus Kowhoro is a Lecturer in the Department of Electrical and Electronics Engineering of Delta State Polytechnic, Otefe-Oghara, specializing in the emerging edge of intelligent systems. His research area is oriented towards synergy-integrated use of Robotics, Artificial Intelligence (AI), and Automation, particularly to develop novel Embedded Systems for adaptive control and machine perception.



Daniel Ezekiel Ogheneochuko is a computer scientist and academic whose research interests span the very mainstream aspects of modern computing: Artificial Intelligence (AI), Cloud Computing, Software Development, Networking, and Cybersecurity. His research is driven by interest in coming up with practical solutions to ICT challenges and leveraging technology for economic development. He is a registered CPN member and TRCN member with a B.Sc. in Computer Science and a PGD in Technical Education.

Research Field

Clive Asuai: Ambience Intelligence**,** Artificial Intelligence, Data Science, Blockchain Technology, Machine Learning, Cybersecurity, IoT.
Ayigbe Prince Arinomor: AI, Cloud Computing, Software Develpoment.
Collins Tobore Atumah: AI, Embedded Systems, Machines and Automobile.
Imarah Festus Kowhoro: Robotics, AI, Automation, Embedded Systems.
Daniel Ezekiel Ogheneochuko: Software Development, Cloud Computing, Networking, AI.