Journal of Academic Research (Applied Sciences), Vol 29, Issu 2, 2025 "103-110"

# Audio Deep Fake Detection Using an Ensemble Learning Approach

**Dr.Mohammed M Elsheh[1], Mona M Bouaisha[2]**

1- Department of Computer Science, Libyan Academy- Misrata, Libya
2- Department of Computer Science, Faculty of Tourism and Hospitality- Misrata, Libya

m.elshah@lam.edu.ly , mona.aboaisha@it.lam.edu.ly

**Article information**

**Key words**

*Audio Deepfake detection, Libyan dialect, Deep Learning, MFCC, Stacking Ensemble.*

Received 18 11 2025,
Accepted 03 12 2025,
Available online 03 12 2025

**Abstract**

***Short abstract.*** The rapidly evolving tools and techniques based on artificial intelligence pose a serious and growing concern, especially regarding the spread of deepfake technologies and their impact on the credibility and authenticity of audio files. To address this growing challenge and effectively counter these concerns, A hybrid model was specifically designed and developed to detect deepfake voices. This proposed model relies on the latest deep learning techniques, particularly leveraging the power of one-dimensional convolutional neural networks (CNN), in addition to integrating them with a multilayer perceptron (MLP) network and an optimized XGBoost classifier operating within an integrated stacking ensemble. The proposed approach in this work is based on extracting Mel-frequency cepstral coefficients (MFCCs) from audio files which used in training and testing processes. Due to the lack of a suitable and specialized dataset of Arabic audio regarding Libyan dialect publicly available for this specific purpose, an integrated dataset was created and compiled, containing a variety of audio clips, including genuine audio clips, as well as fake ones generated using various artificial intelligence techniques. When testing the model and evaluating its performance, the model demonstrated high accuracy across different datasets, achieving 96.67% on (three seconds clips), 97.50% on (five seconds clips), and 97.00% on (seven seconds clips), confirming its exceptional ability to detect subtle and hidden anomalies.

## I. Introduction

Audio Deepfake (AD) refers to the generation of synthetic audio in which authentic human vocal characteristics are algorithmically replaced with artificial counterparts, thereby introducing substantial risks of misinformation, identity spoofing, and digital deception. Although misinformation has been a persistent phenomenon on the internet, the proliferation of AI-generated deepfakes has markedly intensified the socio-technical threat landscape, rendering the potential consequences significantly more alarming and demanding urgent scholarly and regulatory



mona.aboaisha@it.lam.edu.ly

attention [1]. This technological evolution has made it increasingly challenging to distinguish between authentic human speech and AI-generated voice replicas, posing critical implications for security protocols in sensitive sectors such as banking, authentication systems, and personal property access [2]. These legal and ethical implications underscore the real-world risks posed by deepfakes, which have moved beyond theoretical concerns to concrete criminal applications.

Recent developments in generative artificial intelligence, particularly in the field of speech generation, have introduced new challenges and dimensions. The advancement of text-to-speech (TTS) models, such as WaveNet [3], Tacotron [4], and emerging diffusion-based architectures, has significantly improved the quality of synthetic speech, making it very close to natural human speech. Moreover, while these technological breakthroughs benefit many applications, they also enable the generation of highly realistic and convincing fake voices, raising concerns about potential misuse, such as spreading harmful content like hate speech [5].

Various machine learning and deep learning models have been proposed for detecting fake voices, such as CNNs and RNNs, which analyze different audio features like spectral features and MFCCs. These methods have shown promising results in distinguishing real voices from fake ones. Most datasets use the English language, along with other languages such as French, Portuguese, and Spanish, but they are criticized for the absence of natural voices similar to human voices [6].

## II. Related Works

During recent years, many research studies related to detecting real and fake voices have been published, among them are:

Shaaban et al. [7] proposed a hybrid model based on CNN-BiLSTM with an attention mechanism for detecting deepfake audio. The model used Mel-Spectrogram plots and features such as MFCCs, spectral centroid, and zero-crossing rate, achieving 95% accuracy on the SceneFake and Fake-or-Real datasets (which include real and fake audio samples), outperforming models such as Gradient Boosting and XGBoost. To improve interpretability, an XAI layer using SHAP was added. Al Ajmi et al. [8] introduced a deep neural network for blind detection of mimicked voices, achieving 94.2% accuracy. The model, trained on 1,127 English and Arabic speech clips using 26 temporal and spectral features, including MFCCs, surpassed human accuracy of 85% in identifying mimicked voices, presenting a robust tool for forensic audio analysis in uncertain conditions. Almutairi et al. [9] presented Arabic-AD, a self-supervised deep learning framework for Arabic audio deepfake detection. Utilizing the HuBERT pre-trained model, Arabic-AD detects both imitated and synthetically generated voices. The study created the first single-speaker synthetic Modern Standard Arabic (MSA) dataset and a multi-speaker non-native dataset to evaluate the influence of accents. Arabic-AD outperformed existing methods with 97% detection accuracy and 0.027% Equal Error Rate, while also reducing preprocessing needs.

In a study conducted by Hamza et al. [10], machine learning models were proposed based on the "Fake or Real" (FoR) database, which includes more than 195,000 samples of human and synthetic speech from sources such as Deep Voice 3 and Google Wavenet TTS. The study used preprocessing, normalization, and feature extraction techniques such as MFCC and other acoustic features. The results showed the superiority of the SVM model, which achieved an accuracy of 97.57% for FoR-2sec and 98.83% for FoR-rerec. Other models also performed well, such as MLP and Random Forest, while the Gradient Boosting model outperformed others on the FoR-norm dataset with an accuracy of 92.63%, confirming the effectiveness of SVM in detecting deepfake audio across different dataset formats.

Baleisteros et al. [11] proposed the Deep4SNet classification model, which encoded the audio dataset and classified between real and fake audio using a two-dimensional CNN model (spectrogram). This model achieved an accuracy of 98.5% in distinguishing natural fake audio. Rimau and Tzerbos [12] proposed a deep learning approach for detecting AI-generated speech. The authors collected the FoR dataset, which contains approximately 198,000 audio samples from the latest deep learning-based speech synthesizers, in addition to real speech, and improved the performance of various deep neural networks (DNNs) for voice detection. The study demonstrates the effectiveness of DNNs, achieving an accuracy of 92.00% in detecting unseen synthetic speech, significantly outperforming human performance (65.7%).

## III. Research Design

The framework of this research is demonstrated in Fig. 1. It classifies audio recordings as real or AI-generated using a pipeline that includes data collection, preprocessing, feature extraction (MFCCs and STFT), and data augmentation. To train three base learners (MLP, 1D CNN, and optimized XGBoost), build a stacking suite, and evaluate the models against benchmarks. Details of each phase are shown as follows.

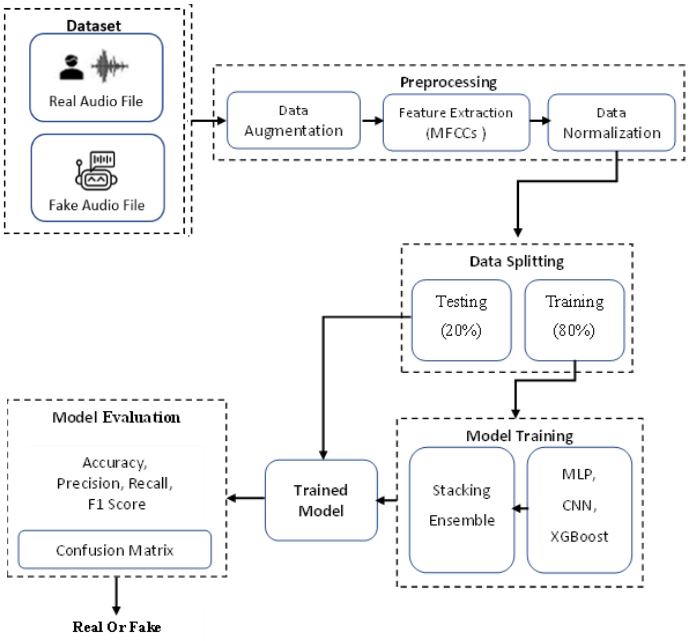


Figure 1. Framework of the proposed Audio Deepfake detection.

## *A. Data Collection:*

To develop a robust system for detecting deepfake audio from the Libyan Arabic dialect, we collected audio recordings from publicly available podcasts and interviews on YouTube featuring well-known and political figures. Initially, all the recordings are varied in length. So, they were preprocessed by dividing the recordings into segments of different durations. An RVC model was trained using genuine audio clips to create deepfake copies. Three datasets were created, each containing audio clips with durations of three seconds, five seconds, and seven seconds, respectively. Each dataset contains 600 audio clips (300 real and 300 fake).

## *B. Data preprocessing:*

To ensure the quality of the audio samples and properly train the proposed model, the data underwent a preprocessing procedure. The audio clips were carefully cleaned to remove silence and unnecessary parts, and their length was standardized to achieve consistency. To enhance data diversity and model robustness, techniques such as adding light background noise, slight changes in pitch and speech speed, and reversing some audio signals were used. These measures improved the model's ability to generalize when handling new audio samples.

## *C. Feature Extraction*

After preprocessing, acoustic features were extracted from the denoised and enhanced audio signals. MFCCs were calculated to represent the spectral and perceptual characteristics of speech, with their first and second derivatives computed to improve temporal and dynamic representation. Additionally, features derived from the short-time Fourier transform (STFT) were integrated to capture variations in the frequency range. These combined features formed a comprehensive representation for each audio sample, enabling the model to efficiently distinguish between real and synthetic speech.

## *D. Data Balancing*

To avoid data imbalance, RandomOverSampler is used to duplicate minority class samples until the classes are equal, balancing the distribution of classes between real and synthetic samples and ensuring fair model training. This prevents bias from class imbalance and improves model performance.

## *E. Model Architectures*

The used hybrid stacking system is consistsing of three base models and a final meta-model:

- **A Multi-Layer Perceptron (MLP):** This deep convolutional design helps distinguish between original and modified audio samples by learning hierarchical feature mappings. The model captures complex and nonlinear relationships between extracted features through multiple fully connected layers that progressively improve the internal representation of the input data.

- **E-Dimensional Convolutional Neural Networks (1D-CNNs):** The CNN model focuses on identifying local temporal and frequency-related patterns within audio data. Its layered structure allows for the automatic detection of key features and spectral variations that may indicate artificial synthesis, making it particularly effective in analyzing short-term audio features and temporal dependencies**.**
- **The optimized XGBoost model**: It is based on gradient-boosted decision trees to improve classification accuracy between real and fake voices by correcting previous errors and focusing on the most distinctive features. GridSearchCV was used to select the best parameters for the model, including the number of trees, depth, and learning rate.
- **Ensemble Model:** To leverage the strengths of different models, a stacking ensemble approach was used that combines the predictions of MLP, CNN, and XGBoost models through a meta-classifier. This integration improves accuracy and stability in detecting fake audio samples.

The models were trained on a balanced dataset of audio clips in authentic and fake Libyan dialects. classified into three time intervals: three seconds, five seconds, and seven seconds. A thorough evaluation of the model was conducted to determine its effectiveness in detecting deepfake audio across different versions of the dataset.

### *F. Evaluations Metrics*

Evaluation metrics for classification include accuracy, recall, F-measure, precision, and confusion matrix. Accuracy indicates correct class classifications, recall measures actual class retrieval, and F-measure provides a consolidated assessment of precision and recall.

## IV. Results and Discussion

To evaluate the performance of the proposed model using the enhanced stacking strategy, which integrates an MLP, CNN, and the XGBoost algorithm for deepfake detection, several performance metrics were analyzed. The used dataset in this evaluation consisted of genuine and fake Arabic audio samples, preprocessed into distinctive representations and divided into training and testing sets. These features served as inputs to the model during the training phase, enabling the system to learn discriminative patterns between original and manipulated audio signals. The evaluation was conducted on an unseen test dataset, and the model was tested on datasets of varying lengths: audio clips of three, five, and seven seconds. Table I shows the model's performance based on audio file duration.

TABLE I. COMPARISON OF MODEL PERFORMANCE ACCORDING TO AUDIO CLIP DURATION

| Dataset | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| **3-second** | 95.75% | 97.66% | 96.69% | **96.67%** |
| **5-second** | 97.50% | 97.50% | 97.00% | **97.50%** |
| **7-second** | 96.07% | 98.00% | 97.02% | **97.00%** |

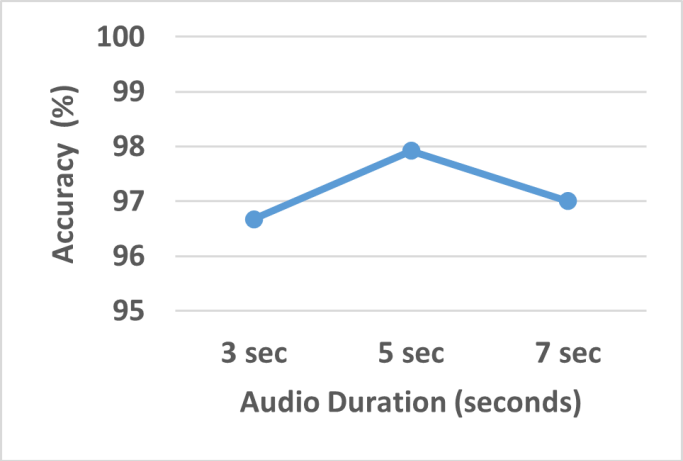


Figure 2. Model performance according to audio duration (three seconds, five seconds, and seven seconds).

Fig. 2 shows that classification accuracy peaked at five seconds of audio duration, suggesting this length provides sufficient acoustic information. Accuracy slightly decreased beyond this point, indicating potential performance saturation or the influence of irrelevant segments.

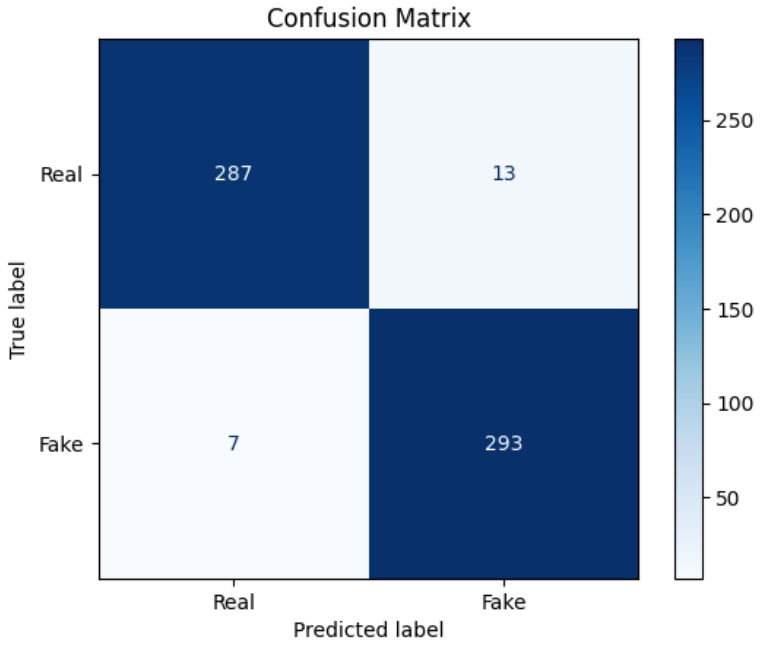


Figure 3. F1-score confusion matrix for the model on the 3-second dataset.

In the first experiment, using three-second audio clips, the model demonstrated excellent performance in classifying real and fake audio clips. According to the confusion matrix Fig. 3, the model successfully classified 287 real clips and 293 fake clips, with only 13 errors in classifying real clips as fake and seven errors in classifying fake clips as real. The model's accuracy reached 96.7%, with a precision of 75.7% and a recall of 97.6%, and an average F1-score of 96.6%. These results indicate the model's ability to extract distinctive audio features even from short clips, reflecting the effectiveness of the audio analysis mechanism in detecting differences between real and fake sounds.

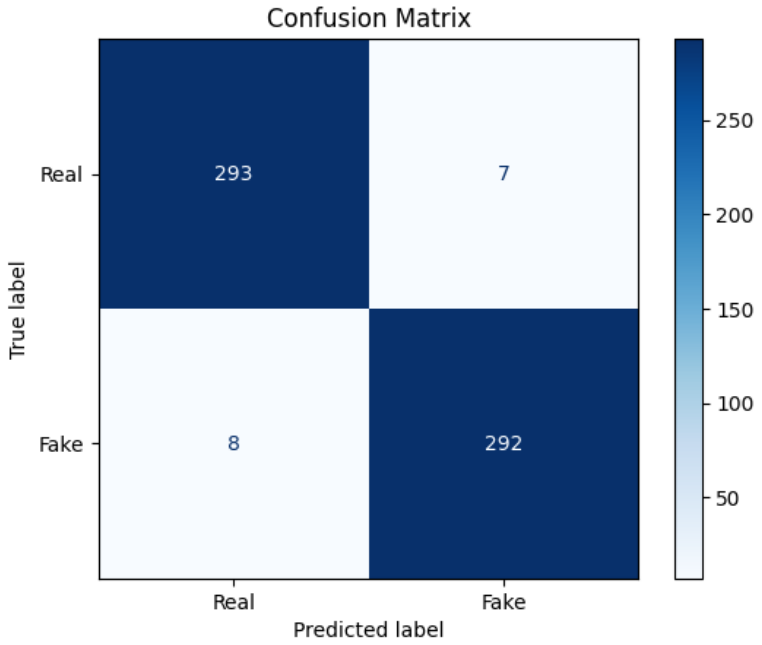


Figure 4. F1-score confusion matrix for the model on the 5-second dataset.

In the second experiment, the duration of the audio clips was extended to five seconds to evaluate the effect of length on model performance. The results showed a noticeable improvement in classification accuracy, as evident in the confusion matrix in Fig. 4. It reached 97.50%, compared to the previous experiment (three seconds). The

model correctly classified 293 real clips and 292 fake clips, with errors reduced to only 15 cases out of 600 test samples. The Precision and Recall values showed a clear similarity between the two classes, both exceeding 97%, reflecting the model's ability to consistently distinguish between real and fake sounds when a longer duration is available, enabling it to extract more stable and distinctive spectral features. The results indicate that increasing the audio clip duration contributes to enhancing the model's reliability without causing an increase in dispersion or audio noise.

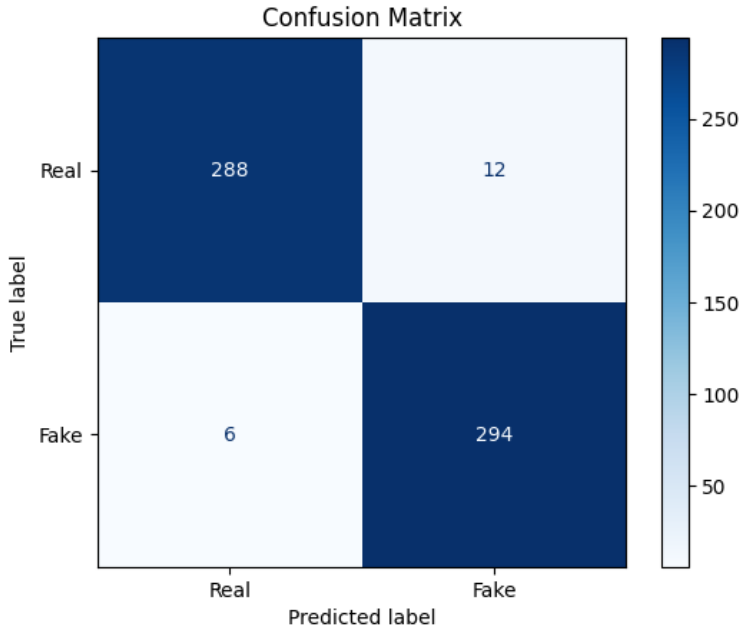


Figure 5. F1-score confusion matrix for the model on the 7-second dataset.

In the third experiment, an audio clips with duration of seven seconds were used. The results showed remarkable stability in performance, as reflected in Fig. 5. The model scored a classification accuracy of 97.0% and Precision, Recall, and F1-score values all approximately 97%. These results indicate that increasing the clip duration to seven seconds did not lead to a significant improvement compared to the five-second duration, reflecting that the model had reached a performance stability level where an optimal balance between clip length and classification accuracy can be achieved. This finding demonstrates that using medium-length clips (around five seconds) is sufficient to extract essential audio features without the need to increase the duration, which helps reduce processing time without affecting performance accuracy.

## V. Conclusion and Future Works

The proposed model, based on the enhanced Hybrid Stacking mechanism, demonstrated high efficiency in classifying audio clips into real and fake categories through testing it on an Arabic dataset collected from the Libyan dialect, containing files of real voices and voices cloned by artificial intelligence. All scenarios achieved accuracy exceeding 96%. It was found that the length of the audio clip affects performance, with the highest accuracy at five seconds, while increasing the duration to seven seconds did not show a significant improvement, indicating the model's stability and saturation in learning. These results suggest that a duration of five seconds represents an optimal balance between the amount of extracted audio information and processing efficiency, enhancing the model's applicability in real-time voice verification systems without the need for long clips or large data. The model proves its capability to achieve accurate and efficient voice classification based on spectrally limited temporal features, opening the way for the development of reliable and practical detection systems in the fields of audio security and digital content.

While these results demonstrate the potential of the ensemble learning model in detecting deepfake audio, there are several areas for possible improvement. Our undergoing work is focusing on using an image-based model for pattern recognition in spectrograms and digital feature models. Integrating digital features and spectrogram images could enhance the model's ability to recognize subtle patterns in sound, which would improve its generalization capability across various real-world scenarios.

References

[1] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," *Frontiers in Big Data,* vol. 5, p. 1001063, 2023.

[2] D. R. Chandran, "Use of AI voice authentication technology instead of traditional keypads in security devices," *Journal of Computer and Communications,* vol. 10, pp. 11-21, 2022.

[3] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499,* vol. 12, p. 1, 2016.

[4] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135,* 2017.

[5] R. Ranjan, K. Pipariya, M. Vatsa, and R. Singh, "SynHate: Detecting Hate Speech in Synthetic Deepfake Audio," *arXiv preprint arXiv:2506.06772,* 2025.

[6] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering,* vol. 47, pp. 3447-3458, 2022.

[7] O. A .Shaaban and R. Yildirim, "Audio Deepfake Detection Using Deep Learning," *Engineering Reports,* vol. 7, p. e70087, 2025.

[8] S. A. Al Ajmi, K. Hayat, A. M. Al Obaidi, N. Kumar, M. S. Najim AL-Din, and B. Magnier, "Faked speech detection with zero prior knowledge," *Discover Applied Sciences,* vol. 6, p. 288, 2024.

[9] Z. M. Almutairi and H. Elgibreen, "Detecting fake audio of arabic speakers using self-supervised deep learning," *IEEE Access,* vol. 11, pp. 72134-72147, 2023.

[10] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, *et al.*, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access,* vol. 10, pp. 134018-134028, 2022.

[11] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4SNet: deep learning for fake speech classification," *Expert Systems with Applications,* vol. 184, p. 115465, 2021.

[12] R. Reimao and V. Tzerpos, "Synthetic speech detection using neural networks," in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD),* 2021, pp. 97-102.

# اكتشاف التزييف العميق الصوتي باستخدام نهج التعلم التجميعي

**د. محمد م. الشح[1]، منى م. بوعيشة[2]**
1 - قسم علوم الحاسوب، الأكاديمية الليبية – مصراتة، ليبيا
2- قسم علوم الحاسوب، كلية السياحة والضيافة – مصراتة، ليبيا

استلمت الورقة بتاريخ 2025/11/18، وقبلت بتاريخ 2025/12/03، ونشرت بتاريخ 2025/12/03
**الكلمات المفتاحية:** اكتشاف التزييف العميق الصوتي، اللهجة الليبية ، التعلّم العميق، معاملات التردد الميلي (MFCC)، التكديس التجميعي ( Stacking Ensemble).

**الملخص**

إنّ الأدوات والتقنيات القائمة على الذكاء الاصطناعي، والتي تتطور بسرعة، تشكّل مصدر قلق خطير ومتزايد، خاصة فيما يتعلق بانتشار تقنيات التزييف العميق وتأثيرها على مصداقية وموثوقية الملفات الصوتية. ولمعالجة هذا التحدّي المتنامي ومواجهة هذه المخاوف بفاعلية، تم تصميم وتطوير نموذج هجين خصيصًا لاكتشاف الأصوات المزيفة عميقًا. يعتمد النموذج المقترح على أحدث تقنيات التعلّم العميق، وذلك بالاستفادة من قوة الشبكات العصبية الالتفافية أحادية البعد (**CNN**)، بالإضافة إلى دمجها مع شبكة الإدراك متعدد الطبقات (**MLP**) ومصنّف **XGBoost** المحسّن، والذي يعمل ضمن إطار تجميعي متكامل قائم على أسلوب التكديس (**Stacking Ensemble**). يرتكز النهج المقترح في هذا العمل على استخراج معاملات التردد الميلي (**MFCCs**) من الملفات الصوتية، التي استُخدمت في عمليات التدريب والاختبار. ونظرًا لعدم توافر مجموعة بيانات عربية مناسبة ومتخصصة تتعلق باللهجة الليبية لهذا الغرض، تم إنشاء وتجميع مجموعة بيانات متكاملة تحتوي على مجموعة متنوعة من المقاطع الصوتية، بما في ذلك المقاطع الحقيقية، إضافة إلى المقاطع المزيفة التي تم توليدها باستخدام تقنيات مختلفة للذكاء الاصطناعي. وعند اختبار النموذج وتقييم أدائه، أظهر النموذج دقة عالية عبر مجموعات البيانات المختلفة، حيث حقق نسبة 96.67% على (مقاطع مدتها ثلاث ثوانٍ)، و97.50% على (مقاطع مدتها خمس ثوانٍ)، و97.00% على (مقاطع مدتها سبع ثوانٍ)، مما يؤكد قدرته الاستثنائية على اكتشاف الشذوذات الدقيقة والمستترة.