Sigma J Eng Nat Sci, Vol. 44, No. 5, pp. 3403–3419, October, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2166

Research Article

A novel hybrid model integrating Mel-Frequency Cepstral Coefficients and convolutional neural networks for audio deepfake detection in Arabic and English

Abdulqader M. ALMARS[1,*], Alaqilah ALQASİR[2], Lama ALMİSNED[2], Roaa ALHARBİ[2], Elsayed ATLAM[3,4]

[1]*Department of Computer Science, College of Science, Northern Border University, Arar, 91431, Saudi Arabia*
[2]*Department of Information System, College of Computer Science and Engineering, Taibah University, Yanbu, 966144, Saudi Arabia*
[3]*Department of Computer Science, College of Computer Science and Engineering, Taibah University, Yanbu, 966144, Saudi Arabia*
[4]*Department of Computer Science, Tanta University, Tanta, 31527, Egypt*

ARTICLE INFO

Article history
Received: 18 January 2025
Revised: 22 March 2025
Accepted: 14 July 2025

Keywords:
Artificial Intelligence; Audio Deepfakes; Convolutional Neural Networks; Deep Learning; Mel-Frequency Cepstral Coefficients; Speech Recognition; Speech Synthesis

ABSTRACT

Audio deepfakes have become a serious concern because to their unethical use in spreading modified religious decrees, marketing non-existent companies, disseminating false political pronouncements, and defrauding people in financial schemes. The potential harm in personal and social contexts highlights need for an effective detection solution. Although various studies have suggested AI algorithms to detect manipulated content, most focus solely on English recordings and disregard other languages like Arabic. This study introduces HMC, a hybrid model that integrates Mel-Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Networks (CNN) to detect audio deepfakes in Arabic and English. The ability of MFCC approaches to record important speech spectrum data provides a substantial advantage. It is designed to mimic human auditory perception, reflect the sound's short-term power spectrum. The CNN network receives the MFCC features as input to recognize the presence of deepfake audio by distinguishing it from genuine speech. The proposed HMC demonstrates superior performance compared with state-of-the-art techniques, such as Random Forest, XGBoost, LSTM, and RNN, on key assessment measures. On the English datase, the model achieved precision, recall, F-score, and accuracy values of 0.9864, 0.9863, 0.9863, and 0.9863, respectively. On the Arabic dataset, the HMC achieved 0.9937 precision, 0.9936 recall, 0.9936 F-score, and 0.9936 accuracy. These results show the effectiveness of HMC in identifying audio manipulation and limiting the spread of deepfake content. This in turn can help establish confidence in digital communications and protect individuals and organizations from related harm.

Cite this article as: Almars AM, Alqasir A, Almisned L, Alharbi R, Atlam E. A novel hybrid model integrating Mel-Frequency Cepstral Coefficients and convolutional neural networks for audio deepfake detection in Arabic and English. Sigma J Eng Nat Sci 2026;44(5):3403–3419.

***Corresponding author.**
*E-mail address: aalmars@nbu.edu.sa

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey
© Author. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

INTRODUCTION

In In recent years, Audio Deepfakes have emerged as a prominent research topic due to the capability of AI-generated tools to create realistic audio. Originally, this technology was created for a wide range of applications to enhance human existence [1]. For instance, it can be used to create realistic voice overs for movies, TV shows, video games, and educational materials, and to assist people with disabilities communicate (due to throat disease or other medical conditions), akin to the case of Stephen Hawking [2]. However, their prevalence has made it increasingly difficult to distinguish these deepfakes from authentic content.

Nowadays, audio deepfakes has become widely used for unethical purposes, and there are many examples that may be seen in our daily lives, including those of many heads of state, politicians, religious scholars, Chief Executive Officers (CEOs), high-level executives, celebrities, companies, and target individuals [3]. Some of these examples happened: In 2023, fake audio clips of a Saudi religious scholar named Sheikh Abdul Aziz bin Baz, who passed away in 1999, manipulated audio snippets gained traction on various social media platforms, therefore insinuating a potential dissemination of misleading religious fatwas in the times to come. At the same time, the promotion of a non-existent product all by the sports journalist Mustafa Agha work in the Middle East Broadcasting Center (MBC) [4, 5], and A fake audio clip from a 17-minute phone call where Trump criticizes Democrats and is accused of attempting to overturn the 2020 election [6, 7]. Also in 2023, Jennifer DeStefano, a mother living in Arizona, heard over the phone a voice saying "Help me, Mom, please help me". The mother believed that the voice belonged to her 15-year-old daughter, who was out skiing [8]. Figure 1 shows a sample of using deepfake technology for the voices of celebrities in social platforms.

In 2020, in the United Arab Emirates (UAE), a CEO was tricked using a fake voice call into agreeing to make a transfer of $35 million [9]. In 2018, a deepfake video featuring former United State (US) President Barack Obama criticizing Donald Trump was circulated with actor Jordan Peele's voice [10-12]. The usage of audio deepfakes in smart homes puts users at serious risk since they enable hackers to get around authentication mechanisms, access personal data without authorisation, and take control of appliances [13]. Consequently, detecting fake audio is a challenging task, but it is necessary to prevent the spread of misinformation and safeguard individuals' reputations [14-16]. Deepfake technology has extended beyond visual media to audio, posing major security and privacy risks. In particular, voice deepfakes are a concern for smart home users who rely on voice recognition to control their environment. For example, attackers can use this technology to create fake voices that mimic authorized users and bypass voice authentication systems. This gives them access to private information or allow them to control smart home devices by unlocking doors or disabling security systems [17].

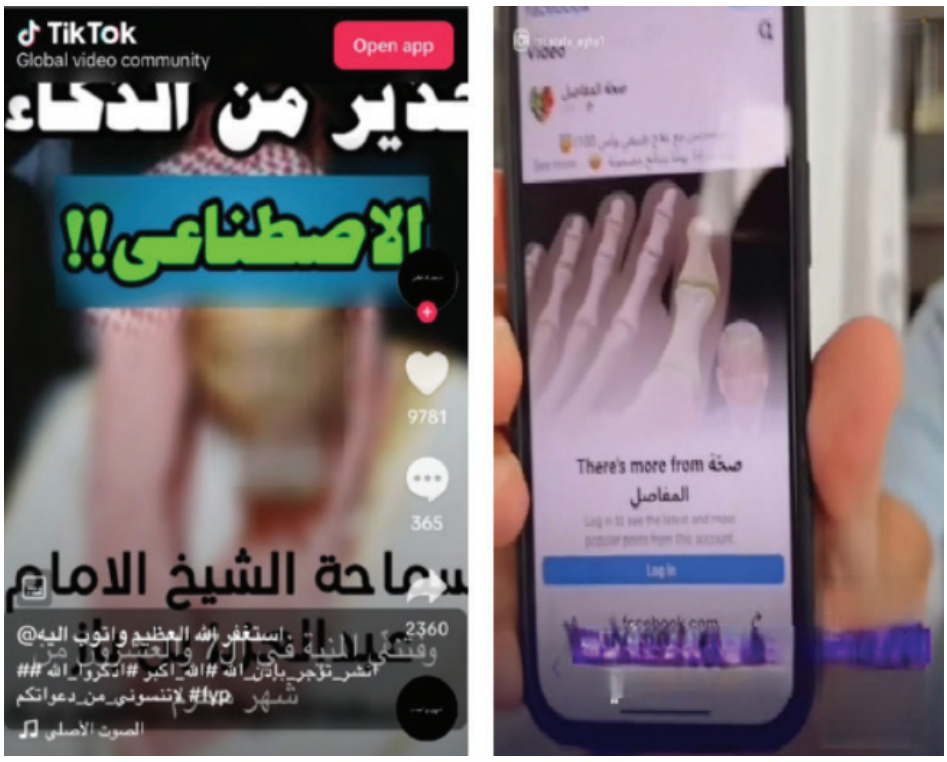

Figure 1. A sample of using deepfake technology for the voices of celebrities.

Several tecnolgoies have been employed in the past decade to improve the detection of audio deepfakes [3, 18–22]. The authors compared learnt and traditional methods for detecting audio spoofing attacks in speaker verification systems using GMM-UBM models [23–25]. Moreover, machine learning methods (e.g., CNN, LSTM, XGBoost, and ABC-CapsNet) have been combined with LFCC and MFCC features to detect fake voices by analyzing spectrograms and speech patterns [26–30]. Although several studies have addressed these issues, three main challenges remain in audio deepfake detection: (1) most models focus on specific languages, such as English, and may not perform well across languages with different speech sounds and acoustic features; (2) few models are designed specifically for Arabic recordings due to the language's complexity and the variety of dialects and slang; and (3) there is a lack of publicly available Arabic datasets for this type of task. Considering the growing risks of deepfake method and its possible misuse for crime and fraud, this study aims to address the existing gaps in the literature. The primary goal of this research is to develop and evaulte a hybrid model that integrates MFCCs and CNNs for the identification of altered voices in Arabic and English. The model uses audio features and deep learning to improve detection performance. In this paper, we introduce the Hybrid MFCC-CNN (HMC) model for fake voice detection. The HMC is capable of differentiating real audio from fabricated audio by analyzing vocal characteristics from English and Arabic datasets. The proposed approach first receives the audio as input, and then extracts features from the voice recording using MFCC technique. Finally, the audio is then classified using CNN into two distinct categories: fake (1) or real (2). The model was trained and tested on a dataset containing real and fake audio recordings in both Arabic and English. Using recordings from both languages helps evaluate the model's ability to detect fake voices. The research results showed that the HMC performed well compared with existing models. This paper summarizes our contributions as follows:

- We developed an innovative approach, called HMC, that combines CNN with MFCC to detect fake recordings in both English and Arabic, significantly improving detection accuracy.
- We collected a new Arabic dataset consisting of 10,422 instances specifically for deepfake voice detection.
- Experiments were conducted on two datasets to verify the HMC. The proposed model show higher precision, recall, and overall accuracy for deepfake audio detection compared to other methods. It achieves an overall accuracy of 98.6% on the English dataset and 99.4% on the Arabic dataset
- The results show that the proposed method can help identify fake content, protect individuals and businesses from potential risks, and improve online safety.

Related Work

The widespread use of deepfake technology has increased the need for effective methods to detect fake audio. This highlights the need to address the growing threat of deepfakes, which can reduce trust in audio content and spread misinformation.

To identify gaps in the existing literature, this study begins by using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach. The PRISMA flow diagram illustrates the study selection process across the different stages of the review (Fig. 2). The diagram shows how many studies were found, included, and excluded, as well as the reasons for exclusion. In this section, we review previous studies on deepfake audio detection and discusses the strengths and weaknesses of different methods. The review also provides a basis for introducing our proposed deep learning model for detecting deepfake audio in both Arabic and English recordings.

Several studies have explored approaches to detecting fake audio [25, 31–37]. For example, Balamurali et al. [23] conducted a comprehensive test of traditional and learned features to improve the robustness of audio spoofing attack detection. The Gaussian Mixture Model–Universal Background Model (GMM-UBM) were used to assessed audio authenticity using genuine speech data, and a fusion model that integrated the outputs of both branches to enhance detection performance. The proposed approach achieved an equal error rate (EER) of 10.8%. The Hvoice dataset was introduced to support the analysis of genuine and fake voice recordings using histograms [38]. Fake samples were generated using imitation and Deep Voice method. This helps detect key characteristics that distinguish authentic audio from manipulated one.

Additionally, the WaveFake dataset was introduced for audio deepfake detection [39]. The dataset uses English recordings from the LJSpeech and JSUT datasets. Different models, including GMM, MelGAN, MelGAN (L), FB-MelGAN, MB-MelGAN, HiFi-GAN, PWG, and WaveGlow, were invistaged to generate fake recordings. The study also used audio features such as LFCC, Mel-Frequency Cepstral Coefficients (MFCCs), delta features, and Mel spectrograms. Furthermore, explainable deep learning techniques were also utilized for detecting deepfake voices [28 40]. For instance, Lim et al. [41] intoduced a method for deepfake voices using three simple CNN-based models: CNN, CNN-LSTM, and CNN LSTMpermuted. Several XAI methods were utilized to analyze spectrograms and linguistic features of voice signals and to understand the suggested model's output. They use the LJSpeech and ASVspoof 2019 Logical Access datasets for training and evaluation. The study highlights that deepfake voices exhibit a flatter energy distribution and higher accent-related variances compared to human voices. The authors show that audio features and XAI methods can help interpret and detect deepfake voices.

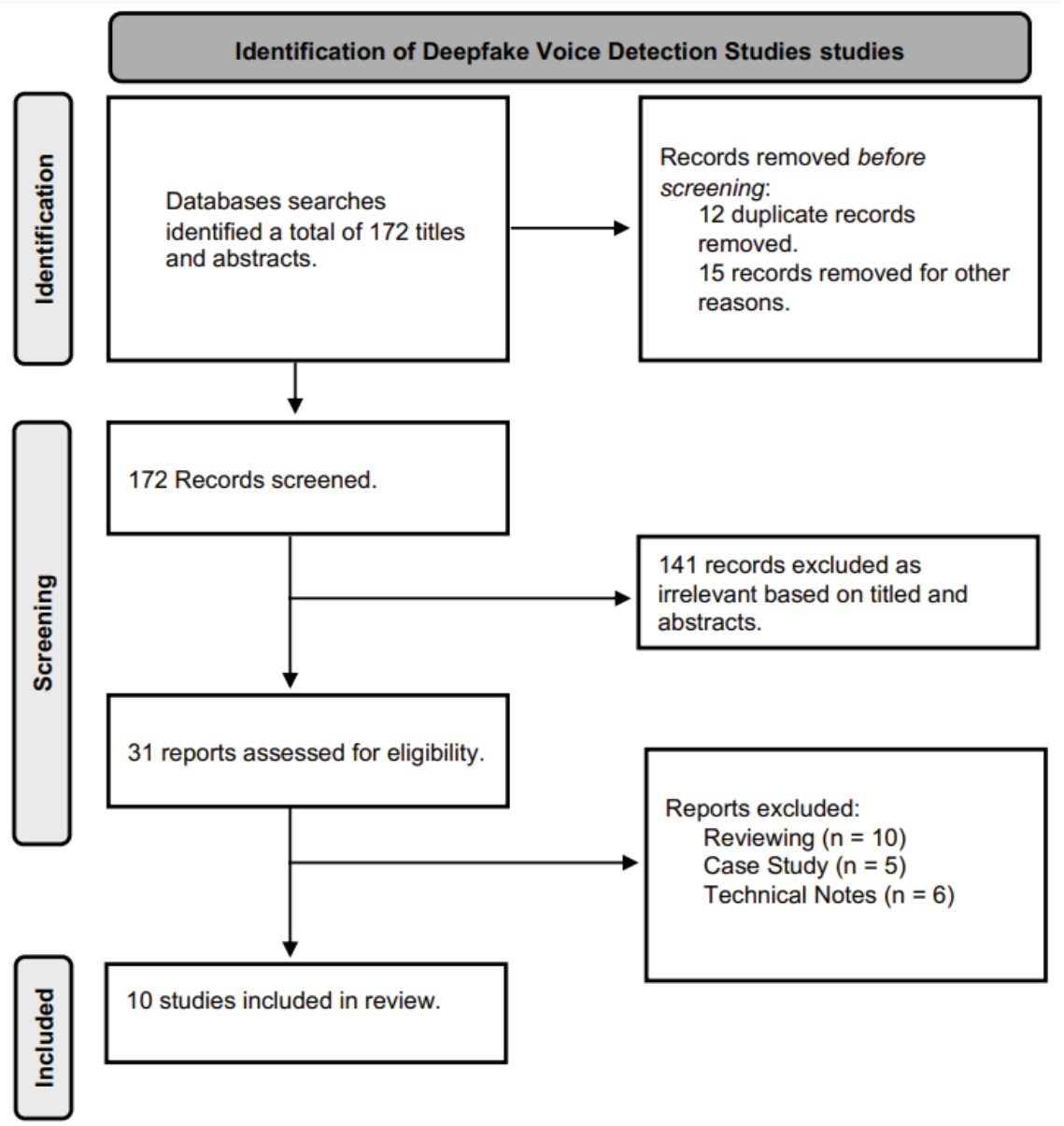

Figure 2. PRISMA flowchart for identification of deepfake voice detection studies.

Khalid et al. [42] presented the FakeAVCeleb dataset which contains audio and video recordings for multimodal deepfake analysis. The authors utilized MFCC features for deepfake detection and employed several machine learning methods, including Capsule, HeadPose, Visual Artifacts (VA), Xception, Meso4, F3Net, Face X-ray, LipForensics, Multimodal, and CDCN, to analyze and evaluate the dataset. Building on this work, Jordan J. and Ahmad Lotfi investigated the real-time detection of AI-generated speech [43]. The research introduces the DEEP-VOICE dataset to train and test several machine learning models, including Extreme Gradient Boosting (XGBoost), Random Forest (RF), Quadratic and Linear Discriminant Analysis, Ridge Regression, Gaussian and Bernoulli Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVMs), Stochastic Gradient Descent (SGD), and Gaussian Process (GP). The final results revelaed that the XGBoost model achieved an average classification accuracy of 99.3% for real-time identification of AI-generated speech with a processing time of approximately 0.004 milliseconds per second of speech. Almutairi et al. [44] used self-supervised deep learning approaches to detect fake audio of Arabic speakers. The authors applied the the proposed method on waveform recordings of Quran reciters, including both authentic and synthetic audio. The findings revealed that it was possible to distinguish fake audio of Quran reciters with a high accuracy of 97% and a low equal error rate of 0.027%.

In recent years, the authors [28] proposed a new methodology called ABC-CapsNet for detecting audio deepfakes. To generate Mel spectrograms, audio samples were preprocessed by resampling to 16 kHz, reducing noise, normalizing amplitude, and removing silent regions. VGG-18 was used to extract features from the spectrogram. These features were enhanced using an attention mechanism and further analyzed by two cascaded capsule networks. The proposed method achieved accuracy rates ranging from 95.5% to 98.1% and Equal Error Rates (EERs) between 0.06% and 1.36% on the ASVspoof 2019 dataset. On the Fake or Real (FoR) dataset, it achieved an overall accuracy of 99% and an EER of 0.04%, with EER values varying across different dataset versions. Moreover, Xie et al. [19] introduced the FakeSound dataset, which was proposed for training and comprehensive evaluation of the deepfake general audio detection model. The FakeSound dataset was automatically generated using a manipulation pipeline that included grounding, masking, and inpainting stages. The dataset was tested on a hybrid model consisting of ResNet, LSTM, and CNN, achieving an accuracy score of 1.0 in the Test-Easy set, 0.85 in the Test-Hard set, and 0.7 in the Test-Zeroshot set.

Furthermore, a robust deepfake audio detection system that combines spectrogram analysis, deep learning, transfer learning, and audio embeddings was proposed in [45]. Frist, the audio features were extracted using three methods: Short-Time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform (WT). Then, the authors then evaluated three approaches using the extracted spectrograms: deep learning models, including CNN, C-RNN, and RNN; transfer learning models, such as ResNet-18, MobileNet-V3, and EfficientNet-B0; and pretrained models, including Whisper, Seamless, SpeechBrain, and pyannote. Finally, the system was evaluated on the ASVspoof 2019 dataset and scored competitive performance. The best result achieved was an error rate of 0.03 with an accuracy of 0.994. Feng et al. [46] suggested a multi-frequency channel attention mechanism (MFCA) and 2D discrete cosine transform (DCT) for audio deepfake detection. The results of experiments reveled that the model could effectively distinguish between fake and real audio on the Fake or Real (FoR) dataset. Moreover, Chakravarty and Mohit Dua extracted audio features using ResNet50 and Linear Discriminant Analysis (LDA) [47]. They tested the proposed system on the ASVspoof 2021 dataset using several ML classifiers, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB). The research Findings showed that the system achieved an accuracy of 99.7% and an Equal Error Rate (EER) of just 0.4% in detecting fake audio.

Although several models have been proposed for deepfake detection, only a limited number of studies have specifically focused on detecting Arabic audio deepfakes [26, 29, 30, 48, 49]. In addition, detection models should not be limited to a single language but should support multiple languages to improve their adaptability and scalability. The Detection of fake voices across different languages is challenging because of differences in linguistic structures and phonetic characteristics. Hence, cross-linguistic models need to consider these variations to successfully differentiate fake audio from real one in Arabic and English. Recent papers have focused on developing models that can detect fake audio in different languages while maintaining reliable detection performance. These research otuputs highlight the need for ML models that can effectively recognize fake audio in different languages [50, 51]. However, Arabic presents unique challenges for audio deepfake detection because of its grammatical complexity and the wide variety of dialects and slang. To overcome this issue, this paper presents a hybrid model of MFCCs and CNNs for the detection of audio deepfakes. The related work is summarized in Table 1.

Table 1. Overview of related research on fake audio classification

Paper	Year	Classifier	Audio Feature Method	Speech Language	Dataset	Accuracy
[23]	2019	GMM-UBM and Fusion model	Several techniques, such as MFCCs, CQCCs, LPCCs, IMFCCs, LFCCs, and others	English	ASVspoof	Reached 10.8 EERs
[38]	2020	Imitation method and the Deep Voice method	Histograms and Audio features based on wavelet coefficients of voice signals	Not mentioned explicitly	HVoice	Not mentioned explicitly
[39]	2021	GMM, MelGAN, MelGAN (L), FB-MelGAN, MB-MelGAN, HiFi-GAN, PWG, and WaveGlow	LFCC, Mel-Frequency Cepstral Coefficients (MFCCs), Delta features, and Mel spectrograms	English and Japanese	LJSPEECH and JSUT	Various classifiers achieved different EERs scores.

Table 1. Overview of related research on fake audio classification (countinued)

Paper	Year	Classifier	Audio Feature Method	Speech Language	Dataset	Accuracy
[41]	2022	CNN, CNN-LSTM, and CNN LSTM-permuted	Short-term Fourier transform	English	LJSpeech and ASVspoof	Various classifiers achieved different EERs scores.
[42]	2022	Capsule, HeadPose, Visual Artifacts (VA), Xception, Meso4, F3Net, Face X-ray, LipForensics, Multimodal, and CDCN	MFCCs	English	FakeAVCeleb	Various classifiers achieved different accuracy scores.
[43]	2023	XGBoost, RF, Quadratic and Linear Discriminant Analyses, Ridge Regression, Gaussian and Bernoulli Naive Bayes, KNN, SVMs, SGD, and GP	MFCCs	English and Japanese	LJSPEECH and JSUT	99.3%
[44]	2023	CNN and LSTM	MFCC and WaveForms	Arabic	MSA	Achieved high detection accuracy (97%) and a low equal error rate (0.027%)
[27]	2024	ABC-CapsNet	VGG18	English	ASVspoof	Achieved accuracies ranging from 95.5% to 98.1% with EERs ranging from 0.06% to 1.36%
[19]	2024	ResNet, LSTM ABD CNN	Not mentioned explicitly	Not mentioned explicitly	FakeSound	Reached an accuracy score of 1.0 in the Test-Easy set, 0.85 in the Test-Hard set, and 0.7 in the Test-Zeroshot set
[45]	2024	Deep learning models, Transfer learning, and Audio embedding	Short-Time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform (WT).	English	ASVspoof	The best result achieved was an error rate of 0.03 with an accuracy of 0.994.
[46]	2024	MFCMNet	MFCCs	English	Fake or Real dataset	88.2%
[47]	2024	SVM, RF, KNN, and NB	ResNet50 and Linear Discriminant Analysis (LDA)	English	ASVspoof	Reached an EER of only 0.4% and an accuracy of 99.7%

MATERIALS AND METHODS

The main objective of this research is to develop an audio deepfake detection model using both English and Arabic datasets. To achieve this, the study intoduce a hybrid approach called HMC to distinguish between real and fake audio recordings. The proposed model first utlized MFCCs to extract important vocal features from the audio. These features are then processed by a CNN to classify each input as either real or fake. Overall, the proposed system consists of four main stages: dataset description, audio preprocessing and feature extraction, deepfake classification, and final prediction. An overview of the proposed system architecture is shown in Figure 3. The following subsections explain each stage in detail.

Datasets Description

In this paper, we performed experiments using two real-world datasets. In the following section, we describe the datasets in detail.

- **English dataset:** In this study, we utilized the widely recognized 'In-the-Wild' Dataset [52], which includes a CSV file containing real human speech samples alongside their DeepFake counterparts. The dataset

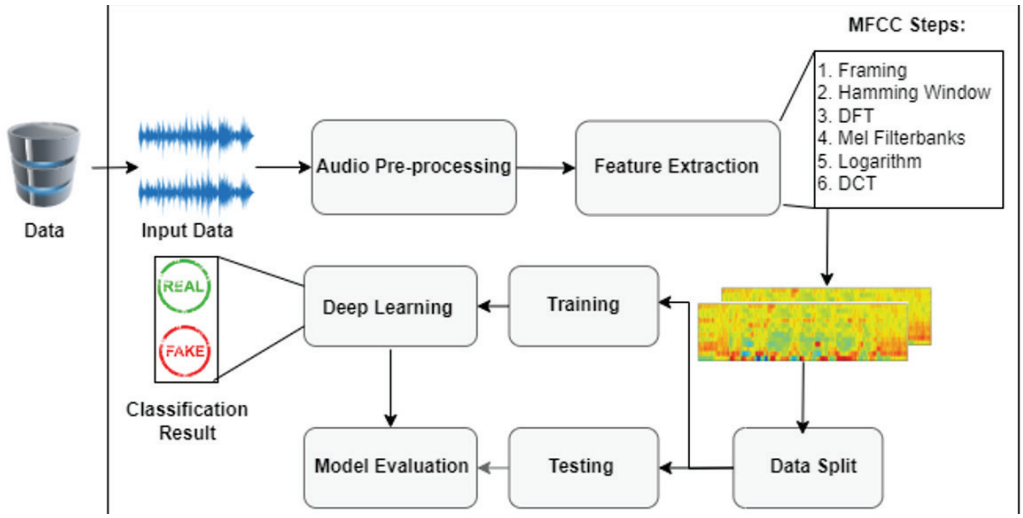

Figure 3. The proposed HMC architecture for fake voice detection.

comprises voice recordings of politicians and public figures. In total, the corpus divided into 19,005 authentic recordings and 12,774 synthetic recordings.

- **Arabic dataset:** In order to enhance the identification of Arabic audio deepfakes, we collected additional audio data from accessible web resources. This dataset was gatherd from YouTube channels, TikTok, and other platforms. The primary motivation for creating this dataset is the lack of publicly available Arabic datasets that distinguish between fake and real audio content shared on social media. The datasets contains a total of 10,422 instances (8,557 real and 1,865 fake). We recruited four experts to collect and annotate the dataset. The dataset annotation workflow comprised a systematic series of steps: discovering relevant audio samples across various digital platforms, examining accompanying user commentary, verifying content authenticity, and classifying each sample as either real or fake.

To mitigate class imbalance, we employed various data augmentation methods to enhance of synthetic audio samples. These techniques encompass:

-**AddGaussianNoise:** This technique incorporates stochastic noise into the audio. The noise amplitude is consistently recorded within the range of 0.01 to 0.02 amplitudes.

– **TimeStretch:** This function stretches the audio duration at random rates. Audio speed is modified by a uniform time-stretching factor between 0.8x and 1.2x.

– **PitchShift:** This method shifts the pitch of the audio by a random number of semitones. The pitch shift is uniformly sampled

Audio Pre-Processing and Feature Extraction

This stage involves converting raw data (signals) into meaningful representations,resulting in spectrograms that serve as input to the proposed model. The proposed model employs MFCCs to achieve this [53] . MFCCs represent the spectral envelope of a sound signal and are widely used in audio techniques for speech recognition tasks. They convert the audio signal from a linear frequency scale to the Mel scale, which closely mirrors the perceived traits of human hearing. The conversion formula from frequency to the Mel scale is:

$$m(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

In order to obtain MFCC spectrograms, the following steps are followed: (1) A default sampling rate is used to sample the audio signals; (2) The audio is then segmented into short, overlapping frames; (3) A Hamming window is employed for each frame to limit spectral leakage; (4) Each individual frame is converted from the time domain into the frequency domain through the Discrete Fourier Transform (DFT); (5) Mel filter banks are then used to extract energy from different frequency ranges; (6) The logarithm of the filter bank energies is calculated to simulate the human auditory system; and (7) a Discrete Cosine Transform (DCT) is utilized to decorrelate and minimize the dimensionality of the filter bank coefficients.

In this study, feature extraction was performed using Librosa (https://librosa.org/doc/main/generated/librosa.feature.mfcc.html), a widely adopted audio processing library in Python. First, the raw audio file paths were retrieved to supply the analysis pipeline. The audio signals were then loaded via librosa.load fucntion. The fucntion produces a tuple containing the audio time series and the sampling rate. To capture the spectral characteristics, we applied librosa.feature.mfcc to extract the MFCC. . In the experiment, we set the sample rate to 22,050 Hz, and the number of MFCC coefficients for each frame was set to 40. After the extraction of MFCCs, we normalized the audio file using mean normalization. After feature extraction,

mean normalization was applied to center the MFCC features around zero. This step centers the data around zero by computing the average value of each coefficient across all time frames and subtracting it from each dimension. This helps reduce differences in feature values and provides a more consistent representation of the audio data. The mean normalization equation is given as follows

$$x' = \frac{x - u}{\max(x) - \min(x)} \quad (2)$$

It's important to mention that we adopted a simple hold-out validation approach to evaluate the proposed model. Hence, the original dataset is divided as follows: 70% for training, 20% for testing, and 10% for validation. Finally, we stored the feature vectors for all audio files and passed them to the next stage for classification. Figures 4 and 5 display samples of real and fake audio, respectively, represented as MFCC spectrograms. These figures illustrate the variation in audio features over time.

Convolutional Neural Networks (CNN)

CNNs have consistently demonstrated robust performance across diverse machine learning domains. In this study, we propose a CNN-based approach to distinguish between real and genuine audio recordings in both Arabic and English. In this study, MFCCs were chosen as the principal feature representation because of their ability to capture compact spectral characteristics while mimicking human auditory perception. CNNs were chosen for their capability to learn discriminative spectrotemporal patterns from two-dimensional MFCC spectrograms while retaining a computationally efficient architecture. Combining MFCC feature extraction with CNN classification provides a practical approach to audio deepfake detection across the two languages.

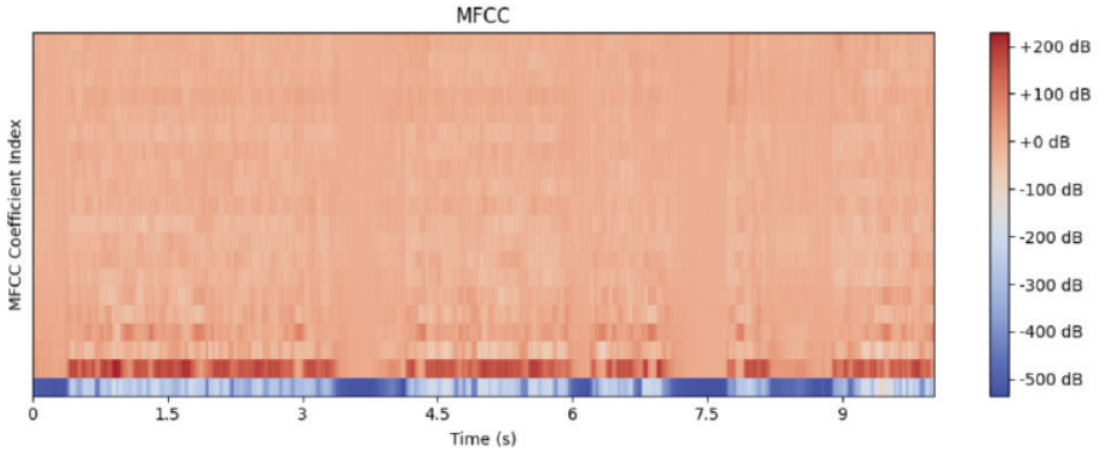

Figure 4. Sample of real audio represented as MFCC spectrogram.

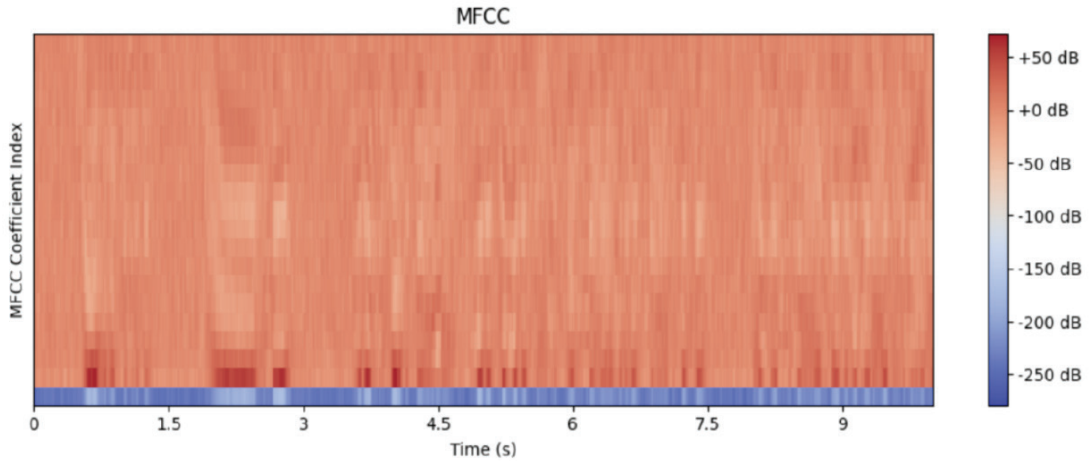

Figure 5. Sample of fake audio represented as MFCC spectrogram.

Algorithm 1 Pseudo-code for Audio Classification Using HMC

Input: Raw audio data
Output: Classification result (REAL or FAKE)
Step 1: Audio Pre-processing and Feature Extraction
    Load audio using librosa library
    Perform audio pre-processing on the input data
    Extract features from the audio data using librosa. feature.mfcc() function
Step 2: CNN Model
    Train a 1D CNN model on the extracted MFCC features
    Model Architecture:
        1. Conv1D layer (filters=32, kernel size=3, activation=ReLU)
        2. AveragePooling1D layer (pool size=2)
        3. Conv1D layer (filters=64, kernel size=3, activation=ReLU)
        4. AveragePooling1D layer (pool size=2)
        5. Conv1D layer (filters=128, kernel size=3, activation=ReLU)
        6. AveragePooling1D layer (pool size=2)
        7. Flatten layer
        8. Dense layer (units=32, activation=ReLU)
        9. Dropout layer (rate=0.5)
        10. Dense output layer (units=number of classes, activation=Softmax)
    Substeps:
        a. Split data into training, validation, and testing sets
        b. Train the model using binary cross-entropy loss function and Adam optimizer
        c. Evaluate the model on test data
Step 3: Classification Result
    Output the classification result (REAL or FAKE)
Step 4: Model Evaluation
    Evaluate model performance based on accuracy and other relevant metrics

In the suggested model, the CNN takes the MFCC features extracted in the previous stage as input, learns important patterns from them, and classifies each audio recording as either real or fake. The network contains three convolutional layers. Each one of convolutional layer is followed by a pooling layer. The goal of convolutional layers is to find relevant features from the input. While, the goal of pooling layers is to reduce the spatial dimensions and retain the most important information. The outputs are then passed to a fully connected layer to produces the final classification. Specifically, the initial convolutional layer employs a 3×3 kernel that scans the input matrix to capture local spectrotemporal patterns and extract salient features. This operation can be expressed using the following equation:

$$F = X^{*} W + b \tag{3}$$

In this case, the input matrix is denoted by X, the kernal matrix by W, the bias term by b, and the output feature map by F . Following the convolution operation, we apply the Rectified Linear Unit (ReLU) activation function to add non-linearity into the model and it is computer as follows:

$$f = max\ (0,x) \tag{4}$$

Second, the average pooling function is used to minimize the spatial dimensions and identify the mean of each feature map using the following equations:

$$P = avgpool(f(x)) \tag{5}$$

where P is the pooled feature map, $f(x)$ is the activation map after applying the ReLU function on the previous operation. The final output for the classification tasks is obtained by adding the completely connected layers at the end.

Real-Fake Classification

In this study, the proposed model applied binary cross-entropy loss function to distinguish between two classes, such as fake and real audio. Furthermore, we use the Adam optimizer, which is widely recognised for its effectiveness and capacity for flexible learning rate, to train the model. The learning rate is set fixed at 0.1. The model was trained using a batch size of 30 over 20 epochs to ensure sufficient learning effectively and early stopping method was applied to test the suggested model's performance on the dataset and avoid overfitting. The output of this stage is illustrated in the equation below:

$$\text{Output} = \begin{cases} \text{Real} & if\ p \leq 0.50 \\ \text{Fake} & 0\text{therwise} \end{cases} \tag{6}$$

The output is determined using Softmax activation function as follows: If the probability p is less than or equal to 0.50, the output is classified as "Real". otherwise, the output is classified as "Fake". The complete pseudo-code for audio classification using HMC is presented in Algorithm 1.

RESULTS AND DISCUSSION

In this section, we first discuss the dataset used in this study, followed by the evaluation methods employed to test the research findings, and finally, the results and their implications.

Evaluation Metrics and State-of-the-Art Methods

The evaluation model's performance involved the use of several widely adopted performance metrics, including accuracy, precision, recall, and F1-Score. The overall accuracy of a model's predictions is the ratio of correctly classified instances to the total number of instances, calculated by the equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Precision estimates the model's accuracy in identifying positive cases using equation below:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Recall measures the model's capacity to accurately identify positive cases and is defined in Equation (9):

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

The F1-Score, as expressed in Equation (10), provides an overall evaluation of the model's performance by acting as a balanced ratio between Precision and Recall. In our evaluation, we employed several state-of-the-art classifiers and well-known feature extraction methods to assess their performance in comparison to the proposed model. The classifiers utilized include Random Forest, XGBoost, LSTM, and RNN. Each classifier was tested with various feature extraction techniques, namely Mel- Spectrogram, CQT (Constant-Q Transform), and MFCCs (Mel Frequency Cepstral Coefficients) [20, 37, 41, 43, 47, 54–56]. In this research, we implemented and evaluated the proposed model using Google Colab Pro, which offers high-performance capabilities for AI, data analytics, and high-performance computing. It provides 89 GB of high-bandwidth memory, 40 GB of RAM, and 235 GB of SSD storage. In our study, the model hyperparameters (such as kernel sizes, number of filters, layer depths, and activation functions) were chosen based on a combination of best practices reported in related literature and preliminary experiments that focused on validation performance. Our objective was to achieve a balance between detection accuracy and model simplicity.

$$F1 - Score = \frac{2(Precision * Recall)}{(Precision + Recall)} \quad (10)$$

Results

The expirmental results of the proposed HMC model in comparison with state-ofthe-art models are presented in Tables 2 and 3. Each model is evaluated based on four key performance metrics: precision, recall, F-score, and accuracy, as outlined in Section 4.1

Based on the results from the English dataset (Table 2), the performance of the proposed HMC model was compared with four baseline classifiers, namely Random Forest, XGBoost, LSTM, and RNN, using Mel-spectrogram, CQT, and MFCCs. According to the findings, we present the following insightful analysis. First of all, the results show that Random Forest achieves its best performance with Mel-spectrogram features, with precision, recall, and F1-score all reaching 0.9500. Although it also does quite well with MFCCs, the CQT features cause a little decline in performance. Second, XGBoost also performs best with Mel-spectrogram features, while its performance improves slightly with CQT features. Third, in contrast, LSTM performs less effectively with Mel-spectrogram and CQT features but improves when using MFCCs compared with both Random Forest and XGBoost. Fourth, the RNN classifier shows a notable performance disparity between Mel-Spectrogram and CQT features. However, it achieves

Table 2. Performance comparison of classifiers using different features on the English dataset

Classifier	Feature Extraction Methods	Precision	Recall	F-Score
Random Forest	Mel-Spectrogram	0.9506	0.9500	0.9500
	CQT (Constant-Q Transform)	0.9021	0.9018	0.9017
	MFCCs	0.9499	0.9490	0.9489
XGBoost	Mel-Spectrogram	0.9816	0.9815	0.9816
	CQT (Constant-Q Transform)	0.9357	0.9356	0.9356
	MFCCs	0.9466	0.9458	0.9458
LSTM	Mel-Spectrogram	0.9312	0.9308	0.9308
	CQT (Constant-Q Transform)	0.9088	0.9086	0.9086
	MFCCs	0.9701	0.9702	0.9702
RNN	Mel-Spectrogram	0.8619	0.8547	0.8541
	CQT (Constant-Q Transform)	0.8525	0.8491	0.8493
	MFCCs	0.9608	0.9607	0.9607
HMC	Mel-Spectrogram	0.9872	0.9866	0.9869
	CQT (Constant-Q Transform)	0.9686	0.9668	0.9667
	MFCCs	0.9864	0.9863	0.9863

Table 3. Performance comparison of classifiers using different features on the Arabic dataset

Classifier	Feature Extraction Methods	Precision	Recall	F-Score
Random Forest	Mel-Spectrogram	0.9163	0.9080	0.9083
	CQT (Constant-Q Transform)	0.8966	0.8966	0.8966
	MFCCs	0.9629	0.9615	0.9614
XGBoost	Mel-Spectrogram	0.9334	0.9332	0.9332
	CQT (Constant-Q Transform)	0.9075	0.9074	0.9072
	MFCCs	0.9706	0.9699	0.9699
LSTM	Mel-Spectrogram	0.8085	0.8085	0.8086
	CQT (Constant-Q Transform)	0.8886	0.8874	0.8876
	MFCCs	0.9679	0.9678	0.9678
RNN	Mel-Spectrogram	0.8801	0.8767	0.8765
	CQT (Constant-Q Transform)	0.8431	0.8390	0.8378
	MFCCs	0.9818	0.9817	0.9817
HMC	Mel-Spectrogram	0.9373	0.9306	0.9294
	CQT (Constant-Q Transform)	0.9720	0.9718	0.9716
	MFCCs	0.9937	0.9936	0.9936

outstanding results with MFCCs recording a precision of 0.9608, recall of 0.9607, and an F-score of 0.9607. Fifth, the proposed HMC model outperforms all baseline classifiers in terms of precision, recall, and F1-score across the three feature extraction methods.

For the Arabic dataset, the research results presented in Table 3 illustrate that Random Forest and XGBoost achieve comparable performance when using CQT and MFCC features. The LSTM and RNN models exhibit the weakest performance across all evaluation metrics compared to Random Forest and XGBoost. However, their performance shows a slight improvement when using MFCCs. LSTM achieves a precision of 0.9679, recall of 0.9678, and F1-score of 0.9678, while RNN achieves a precision of 0.9818, recall of 0.9817, and F1-score of 0.9817. Despite the strong performance of these baseline models, HMC achieves the best overall results, with a precision of 0.9937, recall of 0.9936, F1-score of 0.9936, and accuracy of 0.9936. These findings highlight the effectiveness of the proposed model for detecting audio deepfakes in both English and Arabic. Figures 6

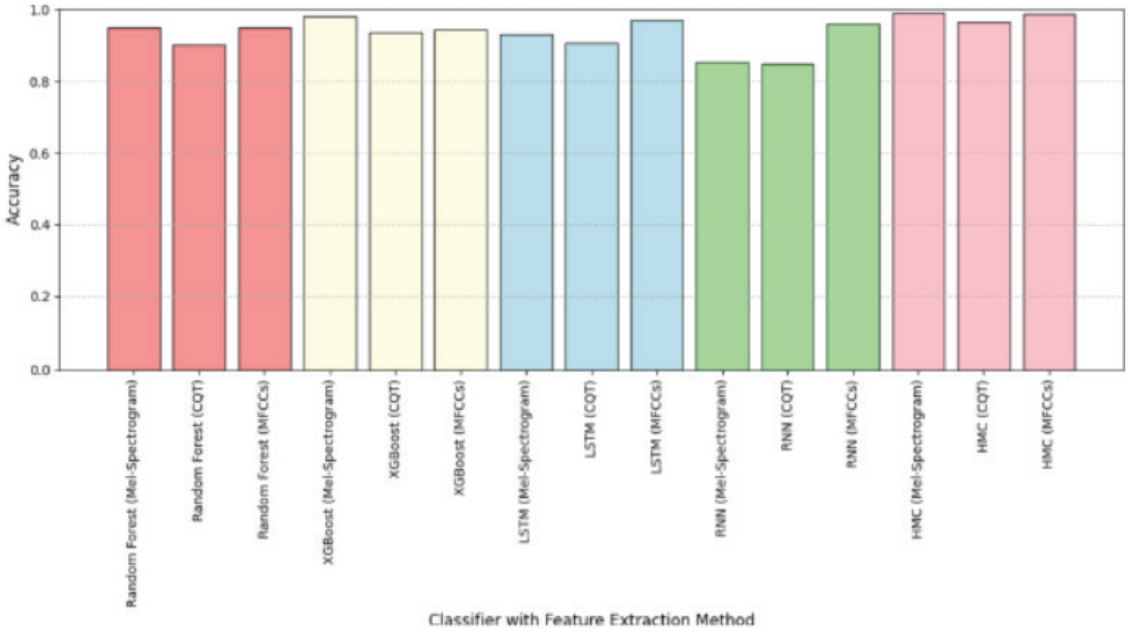

Figure 6. The accuracies of the proposed model compared to other classifiers on the English dataset.

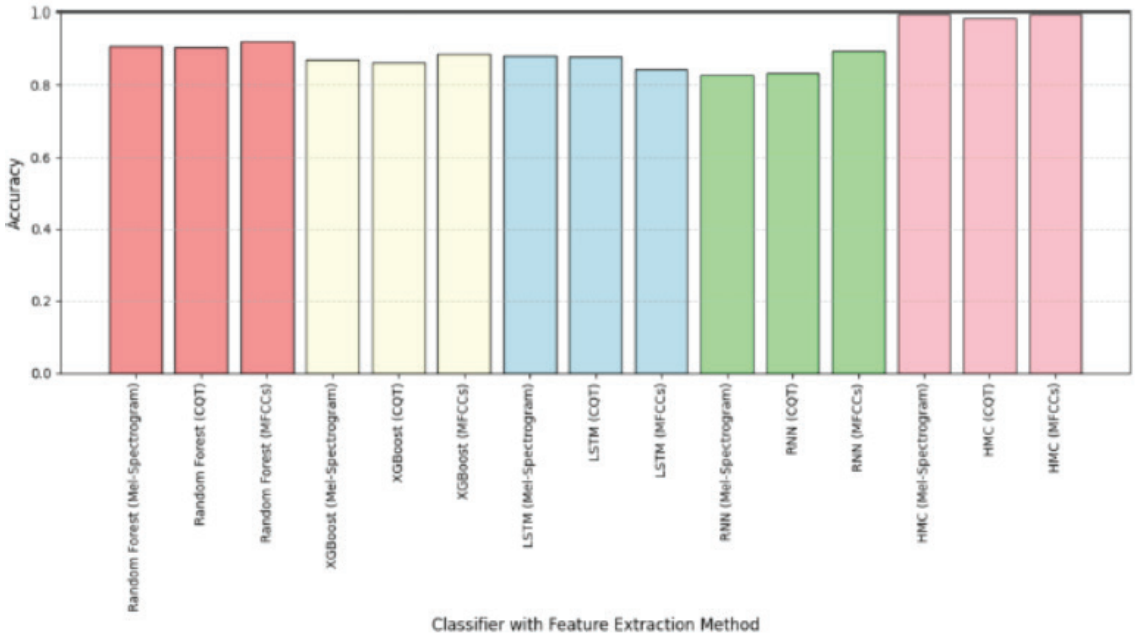

Figure 7. The accuracies of the proposed model compared to other classifiers on the Arabic dataset.

and 7 compare the accuracy of the proposed HMC model with other state-of-the-art models. The results show that HMC achieved the highest accuracy on both datasets. It reached 0.986 for English and 0.994 for Arabic.

The accuracies of HMC model and other state-of-the-art models are comparatively shown in Figures 6 and 7. The results show that the proposed model achieved the highest accuracy scores of 0.986 for the English dataset and 0.994 for the Arabic dataset.

Figures 8 and 9 depict the overall training and validation performance of the proposed model in comparison to state-of-the-art models on the collected Arabic and English audio datasets. Different epoch settings were tested, with 20 epochs used as the default setting. For both datasets, the model achieved high training and validation accuracy, along with low training and validation loss. This indicats stable learning and good generalization. Figures 10 and 11 display the confusion matrices for the HMC model. The results demonstrate that the model excels in distinguishing between genuine and fake audio samples in both datasets. However, the results reveal that misclassifications occasionally arise when handling short or poor audio clips, recordings containing heavy ambient noise, or samples subject to intensive data augmentation and artificial alteration.

Discussion

Detecting audio deepfakes across multiple languages has become increasingly important because of the widespread sharing of multilingual content on social media platforms. Hence, this study introduces HMC model that combines MFCCs with a CNN to detect fake audio recordings in Arabic and English languages. The model demonstrates strong performance across the Arabic and English datasets. MFCCs are applied to extract meaningful speech features and captur important spectral characteristics of audio signals. The CNN is subsequently employed to learn and classify subtle artifacts associated with audio manipulation. By the integration of these methodologies, the proposed framework establishes an effective solution for multilingual deepfake detection.

The experimental results show that the proposed HMC method is more effective than the state-of-art models, achieving an accuracy of 98.6% on the English dataset and 99.4% on the Arabic dataset. These results emphasize the high efficacy of the model especially for English and Arabic where synthetic speech research is comparatively mature. The main advantages of the proposed architecture can be summarized in three main points as follows:

1. The proposed HMC model combining MFCCs and CNN constitutes an important advance to address the challenges of Arabic and English audio deepfake detection. The HMC model can identify deepfakes with high accuracy by leveraging the well-known benefits of MFCCs in feature extraction and CNNs in classification.
2. The model's success is attributed to its ability to capture fundamental characteristics of human speech in a way that is consistent with human auditory system. Moreover, CNNs learn spatial hierarchies from spectrogram-like inputs, which allows them to detect subtle audio manipulations and local patterns.

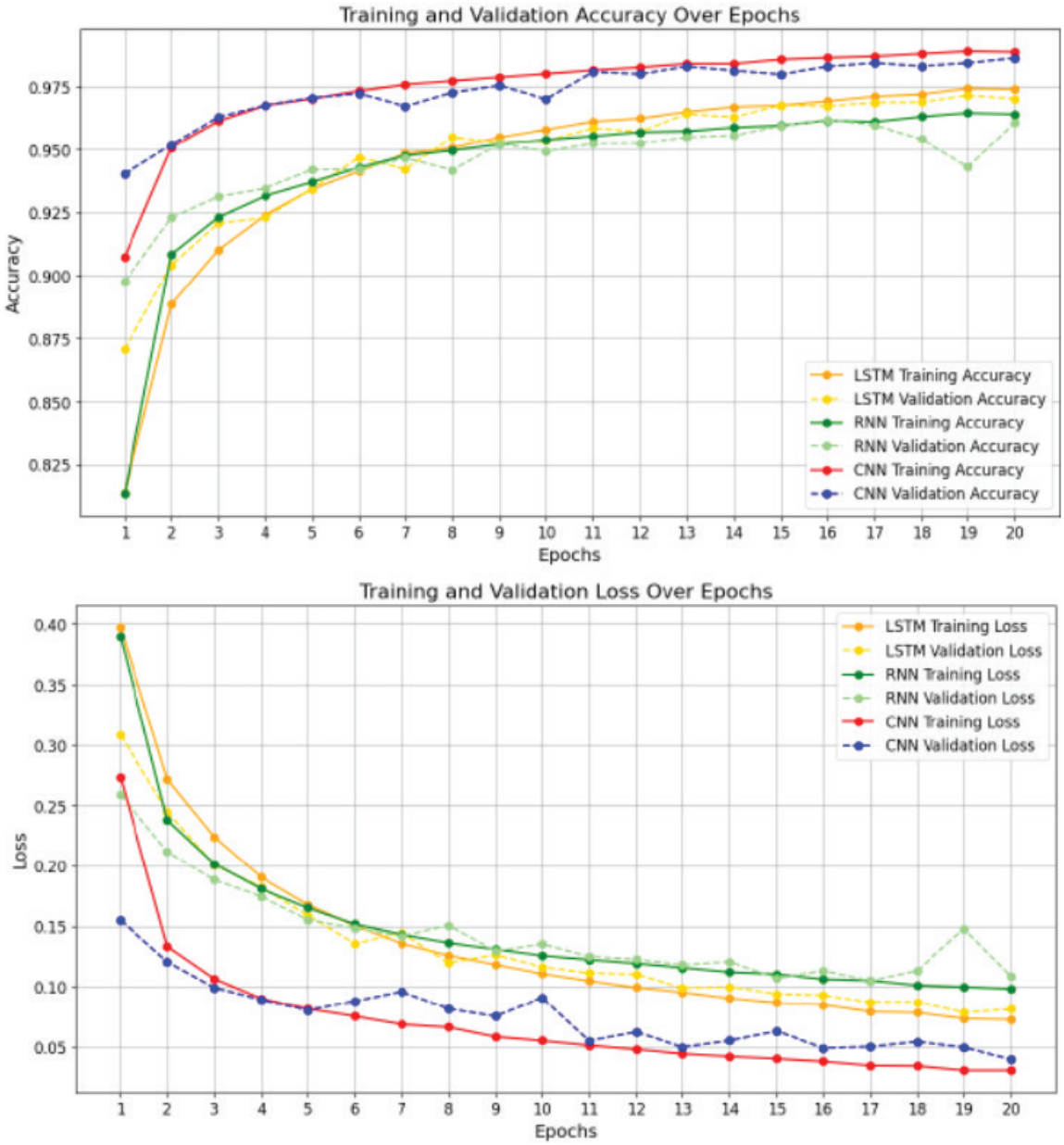

Figure 8. Performance anlaysis of the HMC and other models on the English audio dataset.

3. The model's lightweight design makes it scalable and potentially suitable for deployment in real-world systems, including mobile or embedded platforms.

Despite the promising results of the study, some challenges are needed to be discussed especially in dealing with the complex languages like Arabic that has many dialects. One of the major limitations is the limited availability of datasets to train the model thoroughly in multiple languages. Its generalizability across dialects and acoustic conditions needs further testing. In addition, the model at the moment is not interpretable, which means that it does not provide sufficient insight into how the decisions are made. Future work will involve the integration of explainable AI (XAI) approaches, such as Grad-CAM or LIME, to visualize important regions in CNN-based architectures or feature importance analysis for MFCC inputs. This will increase the model's transparency, helping users to better understand and trust the model's predictions. Finally, careful consideration must be given to the ethical and social implications, including the potential for misuse in surveillance and bias toward some dialects. Future solutions to these problems could involve strategies like generation of more balanced data sets, and utilization of fairness-aware training techniques.

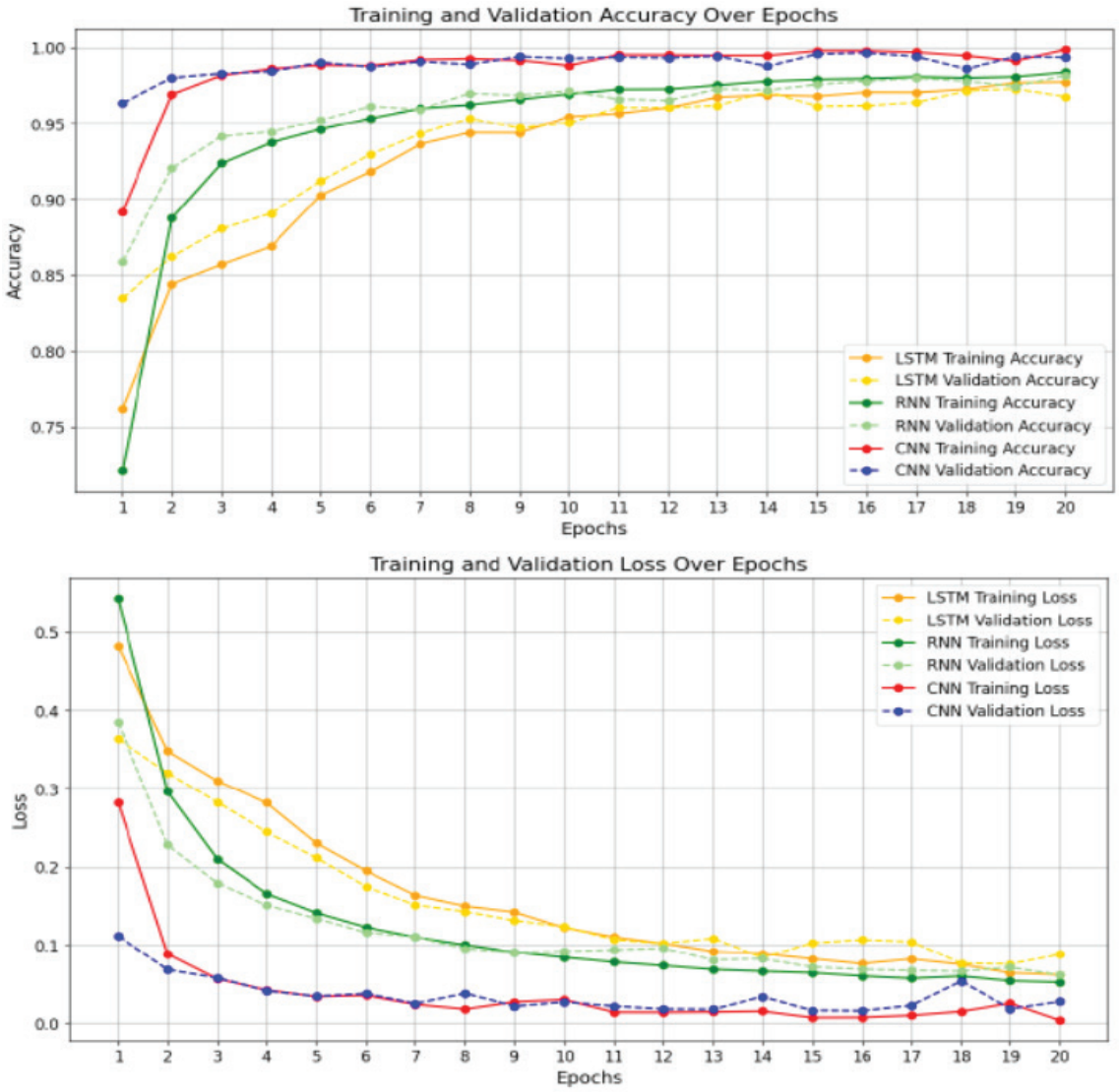

Figure 9. Performance anlaysis of of the HMC and other models on the Arabic audio dataset.

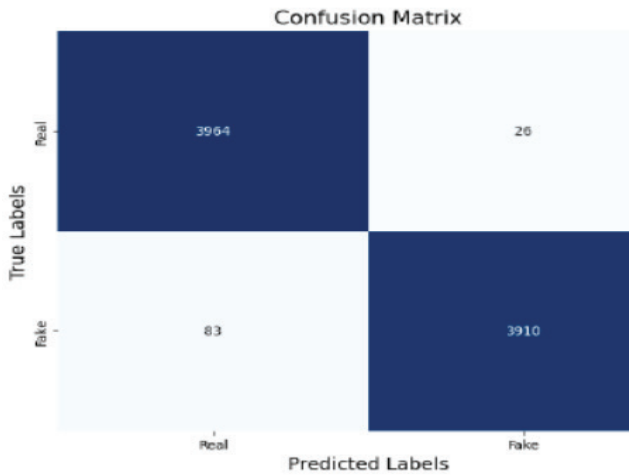

Figure 10. Confusion matrix for the HMC on the English dataset.

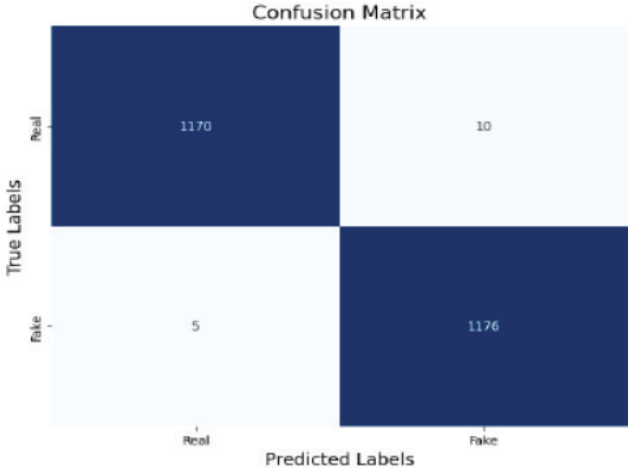

Figure 11. Confusion matrix for the HMC on the Arabic dataset.

CONCLUSION

In this paper, we propose a new hybrid model for detection of deepfake audios in English and Arabic recordings. The model combines MFCCs with CNNs, and represents an important step forward for multilingual audio deepfake detection. The novelty of this approach is that it works in a dual-language, scalable, adaptable and accurate manner to detect fake voices in different linguistic scenarios. The proposed approach aims to revolutionize online trust by tackling deepfakes, enhancing the safety of the digital environment, detecting content manipulation, and shielding individuals and companies from harmful fake content. We evaluate the model on two datasets, English and Arabic, and it achieves better results than existing models in terms of accuracy, precision, recall, and F-score. The proposed model has achieved the accuracy of 98.6% and 99.4% respectively based on the experimental results. However, there are a number of challenges to be overcome. Future work may include:

1. Short term: Use explainable AI (XAI) techniques such as Grad-CAM or LIME for CNN-based visualization and feature importance analysis for MFCC inputs to improve model interpretability. Furthermore, evaluate the proposed model in different languages and different dialects to evaluate the generalizability and robustness of the proposed model in a multilingual context.
2. Mid-term: Utilize adversarial training and evaluate the model against synthetic attacks that deceive deepfake detectors, to evaluate and enhance adversarial robustness.
3. Long-term: expand the model to a larger multimodal platform that can detect fake texts, voices and videos. This may include Transfer Learning and Federated Learning for scalable, real-time and privacy-preserving deployment in real-world scenarios.

AUTHORSHIP CONTRIBUTIONS

A. Alaqilah, L. Almisned, R. Alharbi, A.M. Almars: Conceptualization, Methodology, Software, Investigation. A. Alaqilah, L. Almisned, R. Alharbi: Formal analysis, Writing—original draft preparation. E. Atlam, A.M. Almars: Writing—review and editing, Supervision. A.M. Almars, A. Alaqilah, L. Almisned, R. Alharbi: Visualization. All authors reviewed the manuscript critically for important content.

DATA AVAILABILITY STATEMENT

The data are available from the corresponding author upon request.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Atlam ES, Almaliki M, Elmarhomy G, Almars AM, Elsiddieg AM, ElAgamy R. SLM-DFS: A systematic literature map of deepfake spread on social media. Alex Eng J 2025;111:446–455. [CrossRef]

[2] AI gave Stephen Hawking a voice-and he used it to warn us against AI. Available at: https://qz.com/1231092/ai-gave-stephen-hawking-a-voice-and-he-used-it-to-warn-us-against-ai. Accessed October 24, 2023.

[3] Almars AM. Deepfakes detection techniques using deep learning: a survey. J Comput Commun 2021;9:20–35. [CrossRef]

[4] Mustafa Al-Agha warns against fabricating sound through artificial intelligence in Arabic. Available at: https://slaati.com/2023/08/28/p2357403.html. Accessed November 23, 2023.

[5] Ayish MI. Arab television goes commercial: a case study of the Middle East Broadcasting Centre. Gazette (Leiden) 1997;59:473–494. [CrossRef]

[6] Did a right-wing network interview a fake AI Trump? Available at: https://gizmodo.com/real-america-s-voice-ai-trump-interview-1850796479. Accessed November 23, 2023.

[7] Jacobson GC. The presidential and congressional elections of 2020: a national referendum on the Trump presidency. Polit Sci Q 2021;136:11–45. [CrossRef]

[8] AI scam calls: this mom believes fake kidnappers cloned her daughter's voice. CNN. Available at: https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html. Accessed November 23, 2023.

[9] Fraudsters cloned company director's voice in $35 million heist, police find. Forbes. Available at: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=2a5ef1fe7559. Accessed November 23, 2023.

[10] Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news. The Verge. Available at: https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed. Accessed November 23, 2023.

[11] Sheppard KE. The peeliverse: Jordan Peele and his transmedia storyworld [dissertation]. University of British Columbia; 2023.
[12] Wani TM, Amerini I. Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In: International Conference on Image [conference details incomplete].
[13] Securing voice authentication in the deepfake era. Available at: https://cyberint.com/blog/thought-leadership/securing-voice-authentication-in-the-deepfake-era/. Accessed December 9, 2023.
[14] Guo J, Zhao Y, Wang H. Generalized spoof detection and incremental algorithm recognition for voice spoofing. Appl Sci 2023;13:7773. [CrossRef]
[15] Vafeiadis M, Bortree DS, Buckley C, Diddi P, Xiao A. Refuting fake news on social media: nonprofits, crisis response strategies and issue involvement. J Prod Brand Manag 2020;29:209–222. [CrossRef]
[16] Rasel M, Salam MA, Mohammad A. Safeguarding media integrity: cybersecurity strategies for resilient broadcast systems and combatting fake news. Unique Endeavor Bus Soc Sci 2023;2:72–93.
[17] Girgin ABA, Gümüşçekici G. From past to present: spam detection and identifying opinion leaders in social networks. Sigma J Eng Nat Sci 2022;40:441–463.
[18] Rabhi M, Bakiras S, Di Pietro R. Audio-deepfake detection: adversarial attacks and countermeasures. Expert Syst Appl 2024;250:123941. [CrossRef]
[19] Xie Z, Li B, Xu X, Liang Z, Yu K, Wu M. FakeSound: deepfake general audio detection. arXiv [Preprint]. 2024. arXiv:2406.08052. [CrossRef]
[20] Ma H, Yi J, Wang C, Yan X, Tao J, Wang T, et al. CFAD: a Chinese dataset for fake audio detection. Speech Commun 2024;164:103122. [CrossRef]
[21] Yi J, Wang C, Tao J, Zhang CY, Fan C, Tian Z, et al. SceneFake: an initial dataset and benchmarks for scene fake audio detection. Pattern Recognit 2024;152:110468. [CrossRef]
[22] Osman MM, Büyük O. Parabolic filter mel frequency cepstral coefficient and fusion of features for speaker age classification. Sigma J Eng Nat Sci 2020;38:2177–2191.
[23] Balamurali B, Lin KE, Lui S, Chen JM, Herremans D. Toward robust audio spoofing detection: a detailed comparison of traditional and learned features. IEEE Access 2019;7:84229–84241. [CrossRef]
[24] Wang Z, Hansen JH. Towards improving synthetic audio spoofing detection robustness via meta-learning and disentangled training with adversarial examples. IEEE Access 2024. [CrossRef]
[25] Mobtahej P, Zhang X, Hamidi M, Zhang J. Deep learning-based anomaly detection for compressors using audio data. In: 2021 Annual Reliability and Maintainability Symposium (RAMS); 2021. p. 1–7. [CrossRef]
[26] Khalid H, Kim M, Tariq S, Woo SS. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection; 2021. p. 7–15. [CrossRef]
[27] Wani TM, Gulzar R, Amerini I. ABC-CapsNet: attention based cascaded capsule network for audio deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024. p. 2464–2472. [CrossRef]
[28] Khan S. Explainable deepfake detection using deep learning. 2022.
[29] Banuroopa K, Shanmuga Priyaa D. MFCC based hybrid fingerprinting method for audio classification through LSTM. Int J Nonlinear Anal Appl 2021;12:2125–2136.
[30] Lyu S. Deepfake the menace: mitigating the negative impacts of AI-generated content. Organ Cybersecur J Pract Process People 2024. [CrossRef]
[31] Türk O, Schröder M. A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis. In: INTERSPEECH; 2008. p. 2282–2285. [CrossRef]
[32] Hu T, Khishe M, Mohammadi M, Parvizi GR, Karim SHT, Rashid TA. Real-time COVID-19 diagnosis from X-ray images using deep CNN and extreme learning machines stabilized by chimp optimization algorithm. Biomed Signal Process Control 2021;68:102764. [CrossRef]
[33] Wu Z, Xiao X, Chng ES, Li H. Synthetic speech detection using temporal modulation feature. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; 2013. p. 7234–7238. [CrossRef]
[34] Esmaeilpour M, Cardinal P, Koerich AL. A robust approach for securing audio classification against adversarial attacks. IEEE Trans Inf Forensics Secur 2020;15:2147–2159. [CrossRef]
[35] Osman MM, Büyük O, Tangel A. Effect of number and position of frames in speaker age estimation. Sigma J Eng Nat Sci 2023;41:243–255. [CrossRef]
[36] Khishe M, Caraffini F, Kuhn S. Evolving deep learning convolutional neural networks for early COVID-19 detection in chest X-ray images. Mathematics 2021;9:1002. [CrossRef]
[37] Chen F, Yang C, Khishe M. Diagnose Parkinson's disease and cleft lip and palate using deep convolutional neural networks evolved by IP-based chimp optimization algorithm. Biomed Signal Process Control 2022;77:103688. [CrossRef]
[38] Ballesteros DM, Rodriguez Y, Renza D. A dataset of histograms of original and fake voice recordings (H-Voice). Data Brief 2020;29:105331. [CrossRef]
[39] Frank J, Schönherr L. WaveFake: a data set to facilitate audio deepfake detection. arXiv [Preprint]. 2021. arXiv:2111.02813.

[40] Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 8695–8704. [CrossRef]
[41] Lim SY, Chae DK, Lee SC. Detecting deepfake voice using explainable deep learning techniques. Appl Sci 2022;12:3926. [CrossRef]
[42] Khalid H, Tariq S, Kim M, Woo SS. FakeAVCeleb: a novel audio-video multimodal deepfake dataset. arXiv [Preprint]. 2021. arXiv:2108.05080.
[43] Bird JJ, Lotfi A. Real-time detection of AI-generated speech for deepfake voice conversion. arXiv [Preprint]. 2023. arXiv:2308.12734.
[44] Almutairi ZM, Elgibreen H. Detecting fake audio of Arabic speakers using self-supervised deep learning. IEEE Access 2023;11:72134–72147. [CrossRef]
[45] Pham L, Lam P, Nguyen T, Nguyen H, Schindler A. Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models. arXiv [Preprint]. 2024. arXiv:2407.01777. [CrossRef]
[46] Feng Y. Audios don't lie: multi-frequency channel attention mechanism for audio deepfake detection. arXiv [Preprint]. 2024. arXiv:2412.09467.
[47] Chakravarty N, Dua M. A lightweight feature extraction technique for deepfake audio detection. Multimed Tools Appl 2024;83:67443–67467. [CrossRef]
[48] Lataifeh M, Elnagar A, Shahin I, Nassif AB. Arabic audio clips: identification and discrimination of authentic cantillations from imitations. Neurocomputing 2020;418:162–177. [CrossRef]
[49] Almars AM. DeepGenMon: a novel framework for monkeypox classification integrating lightweight attention-based deep learning and a genetic algorithm. Diagnostics (Basel) 2025;15:130. [CrossRef]
[50] Jeon Y, Lee GG. Multi-level attention aggregation for language-agnostic speaker replication. arXiv [Preprint]. 2024. arXiv:2403.04111. [CrossRef]
[51] Panfili L. Cross-linguistic acoustic characteristics of phonation: a machine learning approach [dissertation]. 2018.
[52] Deepfake-Total.com. In-the-wild dataset. Available at: https://deepfake-total.com/in the wild. Accessed July 22, 2024.
[53] Xu M, Duan LY, Cai J, Chia LT, Xu C, Tian Q. HMM-based audio keyword generation. In: Pacific-Rim Conference on Multimedia; 2004. p. 566–574. [CrossRef]
[54] Dattoli G, Lorenzutta S, Cesarano C. Bernstein polynomials and operational methods. J Comput Anal Appl 2006;8.
[55] Cesarano C, Germano B, Ricci PE. Laguerre-type Bessel functions. Integral Transforms Spec Funct 2005;16:315–322. [CrossRef]
[56] Dattoli G, Cesarano C. On a new family of Hermite polynomials associated to parabolic cylinder functions. Appl Math Comput 2003;141:143–149. [CrossRef]