Journal of Computer and Communications, 2026, 14(3), 27-47
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

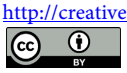
Cybersecurity and Forensic Audio Analysis: Deepfake Detection Based on MFCC, Audio-Text Disconsistency, and Prosodic Features

Nursel Yalçın[1], Kübra Zaptiye[2]

[1]Department of Computer and Instructional Technologies Education, Gazi Faculty of Education, Gazi University, Ankara, Türkiye
[2]Department of Forensic Informatics, Institute of Informatics, Gazi University, Ankara, Türkiye
Email: nyalcin@gazi.edu.tr, zaptiyekubra@gmail.com

How to cite this paper: Yalçın, N. and Zaptiye, K. (2026) Cybersecurity and Forensic Audio Analysis: Deepfake Detection Based on MFCC, Audio-Text Disconsistency, and Prosodic Features. *Journal of Computer and Communications*, **14**, 27-47.
https://doi.org/10.4236/jcc.2026.143003

Received: January 24, 2026
Accepted: March 8, 2026
Published: March 11, 2026



Abstract

Advances in AI-based voice production and conversion technologies have made it possible to create deepfake voices that closely resemble real human speech, raising new security challenges in forensic voice analysis and cybersecurity. Traditional forensic audio analysis methods rely primarily on acoustic characteristics, which can be limited by the increasing realism of deepfake audio. This study proposes an approach for forensic audio analysis that considers the temporal structure of speech, prosodic features and inconsistencies between audio and textual content to enhance the detection of deepfake audio. Accordingly, a dataset was created containing a total of 60 audio recordings, consisting of 30 real and 30 artificial Turkish voice recordings, each 7 - 10 seconds long. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each audio recording; text and word time tags were obtained using an automated speech recognition method. Based on these time tamps, speech rate, pause durations and temporal alignment features were calculated. In addition, prosodic features such as pitch and amplitude were incorporated into the model. All obtained features were classified in two stages using the Random Forest classification algorithm. The model was analyzed in two stages: firstly, without prosodic features, and secondly, with the addition of prosodic features. Model performance was then evaluated using 5-fold cross-validation. The results indicate that incorporating prosodic features leads to higher deepfake Detection accuracy compared to models that rely solely on non-prosodic features. These results aim to demonstrate that temporal and prosodic inconsistencies in forensic audio analysis provide supportive and complementary elements for deepfake audio detection.

Keywords

Deepfake Voice Analysis, Forensic Voice Investigation, Speech-to-Text Discrepancy, Speech-to-Text, Speech Recognition, Artificial Voice, Digital Forgery

1. Introduction

Deepfake technology is a technology that enables the creation of realistic imitations of real people using video and audio recordings. The word deepfake, translated into our language as "deep imitation," emerged in 2017 by combining the words "deep learning" and "fake." Deepfake technology imitates video and audio recordings of real people using machine learning algorithms. While performing these imitations, it uses "Generative Adversarial Networks" (GANs), which can generate new data based on existing datasets [1].

In recent years, rapid advances in AI-based voice synthesis and conversion Technologies have significantly improved the realism of deepfake audio. Deepfakes, particularly those created using voice transformation models, can mimic a person's natural voice characteristics, emphasis, intonation, and emotion by using just a few seconds of audio recordings. These developments are paving the way for the emergence of crime types such as identity theft, digital fraud, content creation for threats and blackmail, and evidence manipulation in the fields of digital forensics and forensic voice analysis. Cybersecurity and cybercrimes are among the areas most affected by the misuse of artificial intelligence. Developments in artificial intelligence and machine learning are being used to exploit artificial intelligence as a criminal element. Security vulnerabilities and cyberattacks, and unauthorized access can constitute AI cybercrimes. This is a new generation digital risk area that threatens information security [2]. The malicious use of deepfake voice technologies directly undermines the reliability of evidence in Forensic processes. Therefore, it is crucial to detect deepfake voices not only technical methods but also thought approaches that can be interpreted and applied in forensic practice.

The main objective of this project is to develop a deepfake voice detection approach based on voice-text discrepancies and prosodic features within forensic audio analysis. In this context, a dataset consisting of real and artificial Turkish voices was created and the voice recordings were converted into text using automatic speech recognition (ASR), with word time stamps were obtained. The aim was to distinguish and classify real and deepfake voices by evaluating acoustic (MFCC), temporal (speech rate, pause structures), and prosodic (pitch, amplitude, breath) features together. Accordingly, a dataset of 60 voices in total (30 real and 30 deepfake) in Turkish was compiled. An analysis was developed using acoustic, temporal, and prosodic features together to numerically demonstrate subtle differences that the human ear cannot easily perceive. The usability of word durations and pauses in deepfake detection using automatic speech recognition

(ASR) was demonstrated. The effect of prosodic features on the detectability of deepfake voices and the cross-validation results with test sets were presented comparatively. The contribution of ASR-based temporal features and prosodic features to deepfake audio detection was experimentally analyzed using the Random Forest classification method.

2. Conceptual Framework

The misuse of artificial intelligence in the creation of video and audio recordings in deepfake technology is of considerable importance in terms of national security, society and privacy. The production and dissemination of fabricated information using video and audio recordings, visual and auditory materials generated with deepfake technology not only leads to ethical violations but also poses a threat that can have serious consequences at the individual and societal levels by manipulating public perception [3].

Forensic phonetics is a scientific field that examines, evaluates, and reports audio data obtained in criminal cases to the relevant authorities. Audio analysis software must have forensic capabilities to be evaluated from a forensic perspective. Forensic phonetics experts primarily examine several areas, including voice identification, forgery and manipulation detection, biometric analysis, phonetic and linguistic analysis, noise analysis, and filtering analysis [4]. The transcription of speech into written text along with the content and technical analysis of the recording are fundamental steps in the examination process [5]. The rise of manipulation techniques such as deepfake has increased the importance of methods for identifying crime related audio recordings related to crimes in forensic audio analysis.

Another important aspect in the analysis process is speech signals. Speech signals are represented through acoustic components such as amplitude, frequency, and duration. These features provide valuable information about the rhythm, accent structure, and perceptual quality of speech [6]. In this context, Mel-Frequency Cepstral Coefficients (MFCCs), which model human auditory perception, are features that summarize the spectral structure of the speech signal and are widely used in speech processing applications [7]. Instead of directly using the raw audio signal, the compact representation of the short-time spectral features of the signal makes the MFCC method effective in applications such as speech recognition and speaker recognition [8]. In the literature, it is stated that the first 12 - 13 MFCC coefficients are sufficient and reliable for representing the spectral envelope of the speech signal, while higher coefficients may contain noise and unnecessary details [9].

However, prosodic features such as emphasis, intonation, speech rate, and pauses show high variability in natural human speech, whereas synthetic and deepfake speech often exhibit more regular and mechanical patterns [10]. Automatic Speech Recognition (ASR) systems enable the quantitative examination of the relationship between speech signals and text by converting speech into written form and providing word-level temporal alignment information [11]. Using these time

alignment outputs, metrics such as speech rate, pause patterns and temporal variability can be calculated. These features provide distinctive indicators for deepfake voice detection. All features obtained in the study were classified using the Random Forest algorithm, an ensemble learning method compatible with multidimensional and heterogeneous data structures. Random Forest is based on constructing multiple decision trees and combining the classification results independently produced by these trees through majority voting. The algorithm reduces the correlation between trees and increases the generalization ability of the model by constructing each decision tree on randomly selected subsets of training data and features [12].

3. Literature Review

This section presents the use of MFCC for deepfake voice detection, machine learning, and the role of deep-learning models in literature. In the study conducted in 2023, Shanmugam, Ismail, Magalingam, Hashim Nik, & Singh aimed to compile the role of acoustic feature measurements (prosodic/spectral) and MFCC features in detecting depression in speech and to evaluate the effectiveness of this combination, especially to reveal the gap in the Malay language. In this regard, they conducted a comprehensive literature review of studies conducted between 2000 and 2023. The reviewed studies accent that spectral features are frequently used; MFCC is one of the most mentioned features, and language proximity can be effective in conveying emotion features. In conclusion, they emphasize that acoustic measurements and MFCC can be effective in predicting depression, but more studies with local data and evaluation of feature combinations are needed to clarify the results in the Malay language [13].

In the study conducted in 2022 study, Hamza *et al.* aimed to extend on deepfake voice detection in Fake-or-Real datasets by conducting detailed experiments on Fake-or-Real datasets and sub-datasets using machine learning and deep learning-based approaches, and to evaluate the performance of these models. The dataset used was Fake-or-Real (FoR): 195,000+ real and synthetic speech samples and sub-datasets such as "for-rerec, for-2sec, for-norm, for-original". For feature extraction, MFCC, spectral, Zero Crossing Rate (ZCR), energy, and Short-Term Fourier Transform (STFT) were used. Principal component analysis (PCA) was used to explain the variance in the data, and 65 features were used. For classification, Support Vector Machines (SVM), Random Forest, Decision Tree, Multilayer Perceptron (MLP), Gradient Boosting, and XGBoost machine learning models were used. According to the machine learning model classification results, the SVM model using various feature extraction methods achieved an accurate rate of 73.46%. In addition, STFT, Mel-Spectrograms, MFCC and Constant Q-Transform (CQT) feature extraction methods were used with the 19-layer VGG model (VGG19) and an accuracy of 89.79% was obtained. Compared with previous research, the 16-layer VGG model (VGG16) achieved the highest performance with an accuracy of 93% [14].

In the study conducted in 2025, Bohara and Bairwa aimed to present a lightweight and interpretable machine learning framework designed to detect deepfake audio recordings. By converting audio samples into mel-spectrograms and employing ensemble classifiers such as Random Forest, Gradient Boosting, and XGBoost, the system aimed to identify subtle artifacts in synthesized speech. The study utilized the DEEP-VOICE dataset, consisting of real human voice recordings from eight speakers and their corresponding synthetic versions generated through Retrievable Voice Conversion (RVC). For feature extraction, mel-spectrograms, MFCC, Zero Crossing Rate (ZCR), and Root Mean Square Error (RMSE) were used. For modeling, Random Forest, Gradient Boosting, XGBoost, and 5-fold cross-validation were applied. Among the evaluated models, XGBoost achieved the highest accuracy at 99.32%, followed by Random Forest at 98.72% and Gradient Boosting at 97.11%. In conclusion, Mel-spectrogram and MFCC features demonstrated high performance in deepfake voice detection, and improved accuracy and efficiency were obtained with XGBoost. However, the limited number of speakers in the dataset may restrict generalization capability, and future studies are recommended to include multilingual, more diverse datasets and noisy environments [15].

In the study conducted by Iqbal, Abbasi, Javed, Jalil & Al-Karaki in 2022, it aimed to develop a detection system to distinguish between deepfake and real audio recordings. Experiments were conducted on the Fake or Real (FoR) dataset. For feature extraction, MFCC, spectral roll off, centroid, contrast, bandwidth, Zero Crossing Rate (ZCR), Principal Component Analysis (PCA) were used, and for classification, machine learning methods such as Support Vector Machine, Decision Tree, Logistic Regression, Naive Bayes, and XGBoost were used. The aim was to demonstrate the effectiveness of feature engineering and machine learning in deepfake audio detection. As a result of the experiment, SVM gave the best results with 98.83% and XGBoost with 93.50% [16].

In the study conducted by Ali Sharafat *et al.* in 2025, the aim was to design a deep learning-based framework using the Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) model to perform deepfake voice detection. Within the scope of the study, CNN-LSTM-based deep learning and Mel-Spectrogram, data preprocessing and mutual matrix were used on a dataset created by combining the ASVspoof2021_LA_eval (Logic Access) and Deep-Voice datasets. As a result, deepfake voice detection performed with a hybrid structure based on CNN and LSTM achieved 97% accuracy [17].

In the study conducted by Kumar *et al.* in 2024, the aim was to strengthen the distinction between fake and real voices using MFCC-based feature extraction and machine learning with deep learning models for deepfake voice detection. The study utilized the "Fake-or-Genuine/Genuine-or-Fake" dataset and its sub-datasets "for-rerec, for-2sec, for-norm, for-original". The dataset was classified using traditional machine learning methods such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), Decision Tree, Extra

Tree, Gaussian Naive Bayes (GNB), AdaBoost, Gradient Boosting, XGBoost, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and deep learning methods such as the 16-layer VGG model (VGG16), Long Short-Term Memory (LSTM), and CNN. CNN was found to have 1000 accuracy/precision/recall/F1 in all subsets. Other machine learning models obtained lower results compared to CNN [18].

Table 1 provides a summary of the studies reviewed in literature.

Table 1. Literature review table.

Author and Year	Object	Method	Data Set	Features Used	Main Findings
[13], 2023	To review the role of acoustic feature measurements (prosodic/spectral) and MFCC features in detecting depression in speech and to evaluate the effectiveness of this combination.	Literature review.	Literature review between 2000 and 2023.	Prosodic/spectral and MFCC	Acoustic measurements and MFCC may be effective in predicting depression, but clarification of the results in the Malay language and further studies with local data, as well as evaluation of feature combinations, are needed.
[14], 2022	To conduct detailed experiments on fake-or-real datasets and sub-datasets using machine learning and deep learning-based approaches for deepfake voice detection, and to evaluate the performance of deep learning and machine learning models.	Support Vector Machines (SVM), Random Forest, Decision Tree, Multilayer Perceptron (MLP), Gradient Boosting, XGBoost machine learning models.	Fake-or-Real	MFCC	The SVM model achieved an accuracy of 73.46%, while the STFT, 19-layer VGG model (VGG19) achieved an accuracy of 89.79%. Compared to previous research, the 16-layer VGG model (VGG16) achieved an accuracy of 93%.
[15], 2025	To provide a lightweight and interpretable machine learning framework designed to detect deepfake audio recordings.	Random Forest, Gradient Boosting, XGBoost, 5-fold cross-validation.	DEEP-VOICE	Mel spectrogram and MFCC, Zero Crossing Rate (ZCR), Root Mean Square Error (RMSE).	XGBoost % 99, 32, Random Forest % 98, 72 and Gradient Boosting % 97, 11.
[16], 2022	Developing a detection system to distinguish between deepfake and real audio recordings.	Support Vector Machine, Decision Tree, Logistic Regression, Naive Bayes, XGBoost.	Fake or Real (FoR)	MFCC, spectral roll off, centroid, contrast, bandwidth, zero-pass rate (ZCR), Principal Component Analysis (PCA), 65 features.	SVM and XGBoost yielded the best results with 98.83% and 93.50% respectively.
[17], 2025	To design a deep learning-based framework using the Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) model for deepfake audio detection.	CNN-LSTM based deep learning	ASVspoof 2021 is a dataset created by combining the _LA_eval (Logic Access) and Deep-Voice datasets.	Mel-Spectrogram, data preprocessing and reciprocal matrix.	Deepfake audio detection performed using a hybrid structure based on CNN and LSTM achieved 97% accuracy.

Continued

[18], 2024	To enhance the distinction between fake and real audio using MFCC-based feature extraction and machine learning with deep learning models for deepfake audio detection.	SVM, KNN, MLP, DT, Extra Tree, GNB, AdaBoost, Gradient Boosting, XGBoost, LDA, QDA, and the 16-layer VGG model (VGG16), Long Short-Term Memory (LSTM), CNN.	Fake-or-Genuine/ Genuine-or-Fak and its sub-datasets for-rerec, for-2sec, for-norm, for-original.	MFCC, noise, stretch, shift, pitch	CNN was found to have 1000 accuracy, precision, recall, F1 across all subsets. Other machine learning models achieved lower results compared to CNN.

When the studies in the literature shown in Table 1 are examined, it is seen that MFCC and Mel-Spectrogram-based features are effective in distinguishing spectral differences between fake (deepfake) and real voices.

4. Methods and Findings

This section covers the general dataset acquisition process, preprocessing stages of audio recordings, extraction of acoustic, temporal, and prosodic features, and the machine learning method used in forensic audio analysis for voice-text discrepancy and prosodic-based deepfake detection. Experimental studies were conducted using the Python programming language. The librosa library was used for writing voice disorders, pre-running, and extraction of Mel-Frequency Cepstral Coefficients (MFCCs). Automatic speech recognition and word-level time alignment were performed using the Whisper ASR model. Praat software was used for the analysis and validation of prosodic features. Deepfake audio samples were obtained via air through ElevenLabs, a voice conversion-based system. The obtained acoustic, temporal, and prosodic features were evaluated using Random Forest illumination with the scikit-learn library.

4.1. Dataset Construction

The dataset was prepared in the Turkish language and consists of a total of 60 audio recordings, including 30 real and 30 deepfake samples. Each audio recording was segmented to a duration of 7 - 10 seconds.

Real audio recordings were obtained by manually extracting segments from publicly available interviews, speech videos, and podcast content using Movavi Video Editor. Deepfake audio recordings were generated using ElevenLabs, an AI-based system employing voice conversion techniques, based on the voices of the same individuals. This approach was preferred to more realistically reflect voice cloning-based deepfake scenarios. All audio recordings were converted to the ".wav" format for analysis.

In accordance with ethical principles and personal data protection requirements, all speaker identities were anonymized using a masking approach. File

naming conventions were anonymized as shown in Table 2, following formats such as "gercekunlu_kisi01" and "deepfakeunlu_kisi01".

Table 2. File naming convention.

Class	Sample File Name	Description
Real	gercekunlu_kisi01.wav	Publicly available real speech segment
Fake	deepfakeunlu_kisi01.wav	Deepfake generated using voice conversion

4.2. Acoustic Feature Extraction (MFCC)

In this study, Mel-Frequency Cepstral Coefficients (MFCCs) were employed to represent the acoustic characteristics of the audio recordings. For each recording, the first 13 MFCC coefficients were computed. Since each audio sample has a duration of 7 - 10 seconds, the MFCC values vary over time throughout the recording. To obtain a compact representation, 13 mean and 13 standard deviation values were extracted for each MFCC-based feature. These features were subsequently used as input to the Random Forest machine learning-based classification model.

4.3. Automatic Speech Recognition (ASR) and Temporal Alignment Approach

To objectively examine temporal and prosodic inconsistencies between real and deepfake speech, each audio recording was transcribed into text using the small version of the Whisper Automatic Speech Recognition (ASR) model. The Whisper model was selected due to its support for multiple languages, including Turkish, and its ability to generate word-level time stamps.

Using the word-level temporal alignment information obtained from the ASR outputs, several features representing the temporal structure and fluency of speech were computed. These features include:

- Total number of words;
- Total speech duration;
- Speech rate (words per second);
- Average word duration;
- Standard deviation of word durations;
- Average and maximum pause durations;
- Number of long pauses.

The extracted features enable quantitative modeling of speech dynamics and facilitate a comparative analysis of structural differences between real and deepfake speech.

4.4. Prosodic Features

To investigate acoustic and prosodic differences between real and deepfake audio recordings, a two-stage analysis approach was adopted. In the first stage, audio signals were processed in the Python environment, and acoustic and prosodic fea-

tures were numerically extracted for each recording. This stage aimed to perform statistical analyses based on objective and comparable metrics.

In the second stage, the same audio recordings were analyzed using the Praat software to enhance the interpretability of the numerical findings and to visually validate the results. In this context, real and deepfake speech samples were comparatively examined through waveform, intensity (amplitude), and fundamental frequency (pitch) contours. Visual inspections enabled clearer observation of irregularities in prosodic structure and deviations from natural speech patterns.

4.5. Classification and Modeling

To automatically distinguish between real and deepfake speech, the Random Forest classification algorithm was employed. Two models were constructed to evaluate the impact of prosodic features. In the first model, a total of 34 features were used, comprising MFCC-based acoustic features and ASR- and temporal alignment-based features. In the second model, fundamental frequency (pitch), amplitude (energy), and breathing-related prosodic features were additionally included, resulting in a total of 49 numerical features. The distribution of features is presented in Table 3.

Table 3. Numerical distribution of MFCC, ASR, and prosody-based features.

Feature Group	Number of Features
MFCC (mean + standard deviation)	26
ASR + temporal alignment	8
Pitch	2
Energy (RMS)	4
Breathing/Silence	9

The attributes given in the numerical distributions in Table 3. consider not the meaning content of the speech, but rather the word durations on the audio signals and the discrepancies in the normal flow of speech in the texts obtained from the ASR outputs; that is, how the text extracted from the audio recordings is realized on the sound. With these attributes, it is quantitatively evaluated whether the sound is produced in sync with the text and with timing like natural human speech.

During the modeling process, a 5-fold cross-validation approach was employed to objectively evaluate classification performance. In this method, the dataset was divided into five equal subsets. Each subset was used once as the test set, while the remaining four subsets formed the training set. In this way, every sample was included in both training and testing phases, and the model's performance was averaged across different data partitions. Finally, to evaluate the model performance, a fixed data split was applied, dividing the dataset into 75% training and 25% test data. This fixed test split evaluation was used as an additional analysis to the 5-fold cross-validation results to verify the discriminative power of the trained model.

5. Results

This section presents the quantitative and qualitative findings obtained through the applied methodologies. The results related to the acoustic, temporal, and prosodic features extracted from real and deepfake audio recordings, as well as evaluations of classification performance, are discussed under this heading.

For each audio recording, the mean and standard deviation of the 13 MFCC coefficients were calculated separately, resulting in a total of 26 MFCC-based feature values per sample. Table 4. presents the MFCC values obtained for gercekunlu_kisi01 and deepfakeunlu_kisi01.

Table 4. MFCC mean and standard deviation values obtained for "gercekunlu_kisi01" and "deepfakeunlu_kisi01".

MFCC Feature	Real mean/std	Deepfake mean/std
mfcc_mean_01	−454.3235778808594	−230.4464111328125
mfcc_mean_02	67.21533966064453	76.89713287353516
mfcc_mean_03	−15.089624404907227	−3.1268277168273926
mfcc_mean_04	33.00016784667969	25.889347076416016
mfcc_mean_05	−2.685513496398926	−1.889790654182434
mfcc_mean_06	−25.871206283569336	−4.986875534057617
mfcc_mean_07	−13.536458015441895	−15.22636604309082
mfcc_mean_08	−8.50728702545166	−11.570257186889648
mfcc_mean_09	1.3628044128417969	−16.674142837524414
mfcc_mean_10	6.6070637702941895	−8.242124557495117
mfcc_mean_11	−15.504878044128418	−13.095629692077637
mfcc_mean_12	−5.026834964752197	−6.054250717163086
mfcc_mean_13	11.52348518371582	−8.076401710510254
mfcc_std_01	117.60819244384766	152.226806640625
mfcc_std_02	43.27265930175781	62.51508712768555
mfcc_std_03	30.987220764160156	31.45989418029785
mfcc_std_04	32.24073028564453	28.25926971435547
mfcc_std_05	23.689128875732422	27.441259384155273
mfcc_std_06	17.865034103393555	14.266642570495605
mfcc_std_07	16.285785675048828	18.385417938232422
mfcc_std_08	15.265581130981445	13.844430923461914
mfcc_std_09	13.020662307739258	15.286628723144531
mfcc_std_10	12.588805198669434	12.764235496520996
mfcc_std_11	11.32986068725586	11.176764488220215
mfcc_std_12	9.968082427978516	10.783785820007324
mfcc_std_13	10.07063102722168	11.100276947021484

An examination of the MFCC mean and standard deviation values calculated for the real and deepfake audio samples in Table 4. reveals distinct spectral differences between the two speech types. In particular, the MFCC standard deviation values in deepfake speech tend to exhibit higher or more irregular distributions. These findings indicate that although deepfake voice generation systems can imitate the spectral structure of speech, they fail to fully capture the temporal spectral stability and natural variability inherent in human speech.

Using the word-level time stamps obtained from ASR outputs, features related to speech rate, word durations, and pause durations were calculated. Figure 1 and Figure 2 present the ASR and temporal alignment results for real audio recordings.

```
Çıkan metin:
anlamı paylaşmakla ilgili konular üzerine çalışıyorum. Fakat dahada önemlisi ya
bu giderek bu iletişim ne korkunç bir şey oldu ya

Kelime zamanları:
anlamı 0.0 0.86
paylaşmakla 0.86 1.8
ilgili 1.8 2.16
konular 2.16 2.8
üzerine 2.8 3.0
çalışıyorum. 3.0 3.8
Fakat 4.26 4.56
daha 4.56 4.82
da 4.82 5.0
önemlisi 5.0 5.66
ya 5.66 6.84
bu 6.84 7.08
giderek 7.08 7.68
bu 7.68 8.22
iletişimle 8.22 8.96
korkunç 8.96 9.44
bir 9.44 9.56
şey 9.56 9.7
oldu 9.7 9.94
ya 9.94 10.2
```

Figure 1. Results of transcribing the "gercekunlu_kisi01" audio recording into text using the Automatic Speech Recognition (ASR) model and the corresponding word-level time alignments.

An examination of the obtained results shows that while the overall meaning of the speech is preserved during transcription, spelling and punctuation errors are present, such as in the transcription of the phrase "daha da." This observation indicates that ASR systems may not always produce fully error-free outputs, particularly for agglutinative languages such as Turkish. However, since the primary objective of this study is not word-level semantic accuracy but rather the analysis of temporal and prosodic structures, these transcription errors do not negatively affect the analysis process.

Figure 2 presents the results of the alignment-based features extracted from the real audio sample "gercekunlu_kisi01". A total of 20 words were detected in the recording, and the total speech duration was measured as 10.18 seconds. Accordingly, the speech rate was calculated to be approximately 1.97 words per second. The average word duration was found to be 0.448 seconds, while the standard

deviation of word durations was 0.246 seconds.

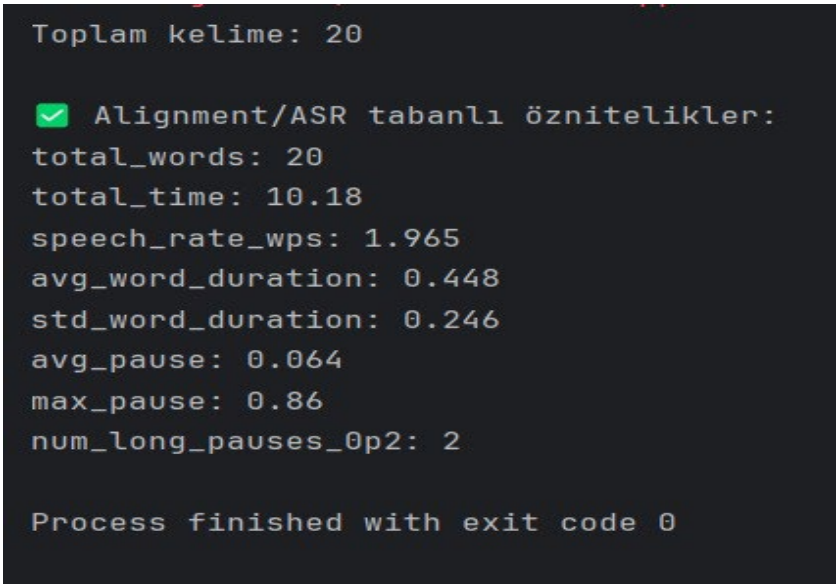


Figure 2. Alignment-based feature results obtained from the "gercekunlu_kisi01" audio recording.

An analysis of pause-related features indicates that the average pause duration was 0.064 seconds, the longest pause lasted 0.86 seconds, and the number of pauses longer than 0.2 seconds was identified as two. These findings demonstrate the presence of natural pauses in real speech, arising from breathing, cognitive processing, and emphasis.

ASR transcription and temporal alignment results of the deepfake audio recordings are presented in **Figure 3** and **Figure 4**.

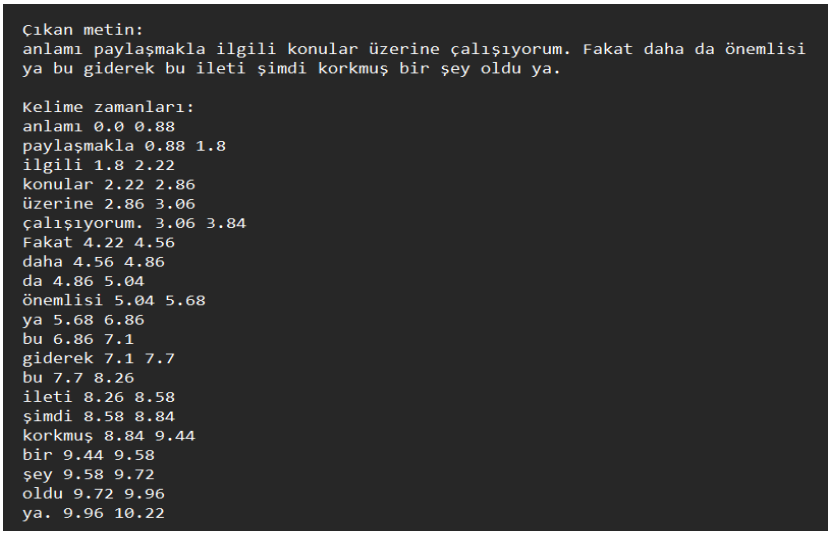


Figure 3. Results of transcribing the "deepfakeunlu_kisi01" audio recording into text using the Automatic Speech Recognition (ASR) model and the corresponding word-level time alignments.

Figure 3 illustrates the transcription output and word-level alignment durations obtained by converting the deepfake audio recording "deepfakeunlu_kisi01" into text using the ASR model. An examination of the results reveals noticeable errors in word selection, spelling, and punctuation. For instance, the word "iletişim" was transcribed as "ileti şimdi," and "korkunç" was transcribed as "korkmuş." These evident errors indicate that although deepfake speech generation aims to replicate the flow of natural human speech, it does not fully achieve naturalness. Such discrepancies reflect limitations in the synthetic speech production process, particularly in capturing nuanced articulatory and prosodic patterns present in genuine human speech

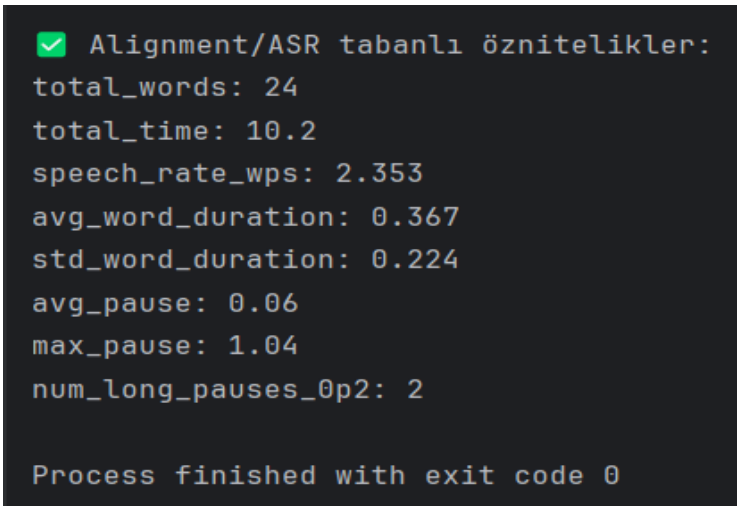


Figure 4. Alignment-based feature results obtained from the "deepfakeunlu_kisi01" audio recording.

Figure 4 presents the results of the alignment-based features extracted from the "deepfakeunlu_kisi01" deepfake audio recording. A total of 24 words were detected in the deepfake sample, and the speech duration was measured as 10.2 seconds. Accordingly, the speech rate increased to 2.35 words per second. The average word duration was calculated as 0.367 seconds, with a standard deviation of 0.224 seconds.

An analysis of pause-related features indicates that the average pause duration was 0.06 seconds, the longest pause lasted 1.04 seconds, and the number of pauses longer than 0.2 seconds was two. However, the higher speech rate and the more homogeneous distribution of word durations suggest that the artificial generation process fails to fully capture the natural variability of human speech.

A comparative analysis of real and deepfake audio recordings revealed that deepfake speech exhibits higher speech rates and shorter word durations with lower variance. In contrast, real speech recordings demonstrate greater temporal variability in terms of word durations and pause structures. These findings indicate that natural phenomena such as breathing, cognitive planning, and prosodic diversity inherent in human speech are not adequately represented in deepfake

audio. Features derived from speech-text alignment, such as speech rate, average word duration, and the standard deviation of word durations, are shown to provide distinctive and interpretable indicators for distinguishing between real and synthetic speech. These results highlight the significance and explanatory power of speech-text alignment-based features in deepfake voice detection.

The findings related to prosodic features specifically intensity (amplitude) and fundamental frequency (pitch) are presented in Figure 5 and Figure 6.

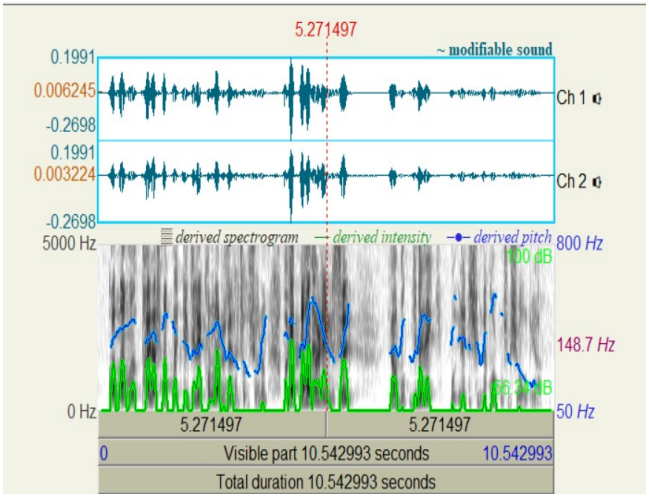


Figure 5. Intensity (green) and fundamental frequency (blue) analyses of the "gercekunlu_kisi01" audio recording.

As shown in Figure 5, the intensity values of the real audio recording exhibit pronounced fluctuations over time. Energy levels increase in emphasized segments and display sharp decreases during breathing and pause intervals. Similarly, the fundamental frequency (pitch) values vary continuously throughout speech, reflecting natural intonation patterns characteristic of human speech.

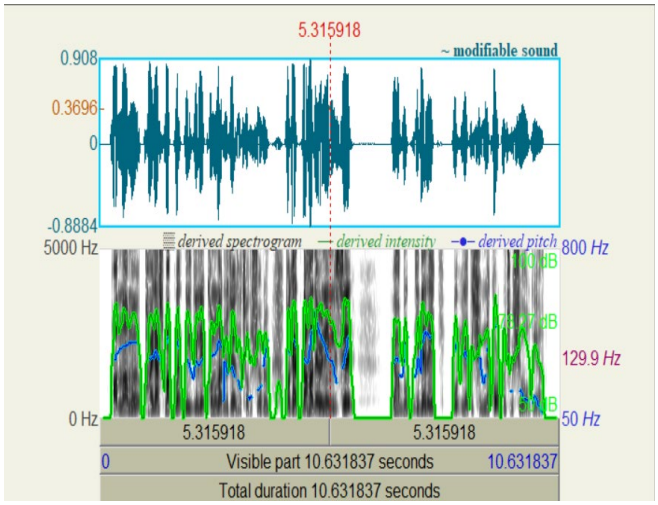


Figure 6. Intensity (green) and fundamental frequency (blue) analyses of the "deepfakeunlu_kisi01" audio recording.

As shown in Figure 6, the intensity values of the deepfake audio recording display a more homogeneous and regular distribution, while the fundamental frequency (pitch) contour exhibits greater stability and limited variability. This finding indicates that artificial speech generation systems are unable to fully model the emphasis, intonation, and energy transitions present in natural human speech.

All extracted features were combined to form a unified feature vector consisting of a total of 49 numerical features for each audio recording.

The results obtained using the Random Forest classification model show that in the first model created for deepfake audio detection, only MFCC-based spectral features and ASR and alignment-based temporal features were evaluated, without considering prosodic features (pitch, amplitude, breath). In the second model, pitch (fundamental frequency), amplitude (energy), and breath-based prosodic features were evaluated along with MFCC-based spectral features and ASR and alignment-based temporal features.

The results obtained using the Random Forest classification model show that in the first model created for deepfake audio detection, only MFCC-based spectral features and ASR and alignment-based temporal features were evaluated, without considering prosodic features (pitch, amplitude, breath). In the second model, pitch (fundamental frequency), amplitude (energy), and breath-based prosodic features were evaluated along with MFCC-based spectral features and ASR and alignment-based temporal features. In the first stage, classification was performed using only MFCC-based spectral features and ASR and alignment-based temporal features. In this context, for each audio recording; A total of 34 numerical features were input to the model: 26 MFCC-based features consisting of the mean and standard deviation values of 13 MFCC coefficients, and 8 temporal features including total word count, speech duration, speech rate, word durations, and pause information obtained from ASR outputs. Figure 7 shows the accuracy values obtained without prosodic features.

```
MODEL-1 (Prosodi Hariç): MFCC + ASR/Hizalama
Dosya: combined_features.csv
Toplam örnek: 60
Toplam sayısal özellik: 34
5-Fold Accuracy sonuçları:
  Fold 1: 0.917
  Fold 2: 0.917
  Fold 3: 1.000
  Fold 4: 0.833
  Fold 5: 0.833
Ortalama Accuracy: 0.900
Standart Sapma:   0.062
```

Figure 7. Accuracy values obtained without prosodic features.

When the results given in Figure 7 are examined, the accuracy values obtained when the Random Forest model trained without prosodic features is examined in 5-fold cross-validation are calculated as 91.7%, 91.7%, 100%, 83.3%, and 83.3%, respectively. It was found that the accuracy values ranged from 83.3% to 100%, the average accuracy was 90.0% (Accuracy = 0.900), and the standard deviation was 0.062.

In the second stage, prosodic features were also included in the classification process in addition to the previous model. For each audio recording, pitch (fundamental frequency), amplitude (energy), and breath-based prosodic features were calculated and added to the model. This increased the total number of numerical features used to 49. Figure 8 shows the accuracy values obtained by including prosodic features.

```
MODEL-2 (Prosodi Dahil): MFCC + ASR/Hizalama + Prosodi
Dosya: combined_features_v2.csv
Toplam örnek: 60
Toplam sayısal özellik: 49
5-Fold Accuracy sonuçları:
  Fold 1: 0.917
  Fold 2: 0.917
  Fold 3: 1.000
  Fold 4: 0.833
  Fold 5: 0.833
Ortalama Accuracy: 0.900
Standart Sapma:   0.062
```

Figure 8. Accuracy values obtained by including prosodic features.

When the results given in Figure 8 are examined, the accuracy values obtained when the Random Forest model trained with prosodic features is examined in 5-fold cross-validation are calculated as 91.7%, 91.7%, 100%, 83.3%, and 83.3%, respectively. It was found that the accuracy values ranged from 83.3% to 100%, the average accuracy was 90.0% (Accuracy = 0.900), and the standard deviation was 0.062.

Table 5 visually demonstrates the cross-validation results, the reciprocal matrix, and the classification performance of the Random Forest model at each level, both with and without prosodic features.

When comparing the 5-fold cross-scan results of the two models created, it was observed that the matrices complemented each other for both models, and the results obtained in appearance were the same. Accuracy values between the folds ranged from 83.3% to 100%. The similar and high levels of sensitivity, recall, and F1-score values for both real and fake results reveal that the model does not show any bias between the classes. These findings demonstrate that, on the current da-

taset, MFCC systematic acoustic features, along with ASR and time alignment features and prosodic features, demonstrate the discriminative power of deepfake audio detection.

Table 5. Visually presents the 5-fold cross-validation results of the two models created.

```
Fold 1
Confusion Matrix [real, fake]:
[[5 1]
 [0 6]]
Accuracy: 0.917

Classification Report:
              precision    recall  f1-score   support

        fake      0.857     1.000     0.923         6
        real      1.000     0.833     0.909         6

    accuracy                          0.917        12
   macro avg      0.929     0.917     0.916        12
weighted avg      0.929     0.917     0.916        12
```

Folding 1. confusion matrix, and classification report results.

```
Fold 2
Confusion Matrix [real, fake]:
[[6 0]
 [1 5]]
Accuracy: 0.917

Classification Report:
              precision    recall  f1-score   support

        fake      1.000     0.833     0.909         6
        real      0.857     1.000     0.923         6

    accuracy                          0.917        12
   macro avg      0.929     0.917     0.916        12
weighted avg      0.929     0.917     0.916        12
```

Folding 2. confusion matrix, and classification report results.

```
Fold 3
Confusion Matrix [real, fake]:
[[6 0]
 [0 6]]
Accuracy: 1.000

Classification Report:
              precision    recall  f1-score   support

        fake      1.000     1.000     1.000         6
        real      1.000     1.000     1.000         6

    accuracy                          1.000        12
   macro avg      1.000     1.000     1.000        12
weighted avg      1.000     1.000     1.000        12
```

Folding 3. confusion matrix, and classification report results.

```
Fold 4
Confusion Matrix [real, fake]:
[[4 2]
 [0 6]]
Accuracy: 0.833

Classification Report:
              precision    recall  f1-score   support

        fake      0.750     1.000     0.857         6
        real      1.000     0.667     0.800         6

    accuracy                          0.833        12
   macro avg      0.875     0.833     0.829        12
weighted avg      0.875     0.833     0.829        12
```

Folding 4. confusion matrix, and classification report results.

```
Fold 5
Confusion Matrix [real, fake]:
[[4 2]
 [0 6]]
Accuracy: 0.833

Classification Report:
              precision    recall  f1-score   support

        fake      0.750     1.000     0.857         6
        real      1.000     0.667     0.800         6

    accuracy                          0.833        12
   macro avg      0.875     0.833     0.829        12
weighted avg      0.875     0.833     0.829        12
```

Folding 5. confusion matrix, and classification report results

Figure 9 shows the results obtained with a fixed data pane, in addition to the cross-scan results, for the purpose of evaluating the model over a fixed data pane.

```
 5-Fold CV sonuçları:
Fold 1: 0.917
Fold 2: 0.917
Fold 3: 1.0
Fold 4: 0.833
Fold 5: 0.75

 Ortalama Accuracy: 0.883
 Standart Sapma: 0.085

Confusion Matrix [real, fake]:
[[7 0]
 [0 8]]

Detaylı sınıflandırma raporu:
              precision    recall  f1-score   support

        fake       1.00      1.00      1.00         8
        real       1.00      1.00      1.00         7

    accuracy                           1.00        15
   macro avg       1.00      1.00      1.00        15
weighted avg       1.00      1.00      1.00        15
```

Figure 9. Classification performance of the trained model using fixed data split on the test set.

Model performance, evaluated on an independent test set, shows that adding prosodic features to the model allows it to correctly classify all real and deepfake audio recordings with 100% accuracy, indicating high discrimination even across independent data sets. However, the limited size of the dataset means the content of the test set can affect the accuracy rate.

The findings indicate that energy, fundamental frequency, and breath-based prosodic features offer distinctive indicators for revealing unnatural speech patterns specific to deepfake voices.

6. Conclusion and Recommendations

This is feasible; an approach combining voice-text inconsistencies with temporal and prosodic features has been proposed and evaluated for the detection of deepfake voices in forensic audio analysis. The dataset, consisting of Turkish real and deepfake voices, encompassing both widespread and borderline features, demonstrates that the combination of acoustic and temporal features is effective in separating deepfake voices. The results obtained using the Random Forest representation model for classifying deepfake and real voices reveal a high degree of selective ordering of the features used in the experiment. This study serves as a preliminary work demonstrating that the proposed feature set possesses discriminatory power in deepfake detection. The sample size used in this study is considered limited in terms of machine learning models and may exhibit overlearning tendencies, especially with small datasets like the Random Forest classification model used in this study. Therefore, it is recommended that future studies evaluate the

generalizability of the method using larger datasets containing longer and more diverse speech samples. Furthermore, comparing different deepfake production techniques, integrating emotion recognition-based features, and examining ASR-based word error rates in more detail stand out as potential research directions that could improve deepfake detection performance. Moreover, adapting the developed approach to forensic reporting processes could strengthen the method's practical application and forensic validity. In conclusion, given the rapid development of AI-based deepfake technologies, developing reliable and explainable detection methods for both audio and multimedia audio-visual content is becoming increasingly critical for the field of digital forensics.

7. Discussion

The deepfake approach proposed in this study shows that acoustic feature analysis alone is insufficient for detecting artificially generated deepfake voices, and that they differ from real human speech not only in terms of spectral structure but also in terms of intonation and speech patterns, word timing, and prosodic features in the voice recording. This finding is consistent with previous studies emphasizing that modern voice cloning techniques and text-to-speech (TTS) systems, even if they produce a spectrally similar voice recording, cannot fully model the behavior and temporal naturalness of the voice in the recording [19] [20].

Previous studies have shown that models trained with MFCC-based feature extraction are effective in distinguishing spectral differences between fake and real voices in deepfake voice detection. As seen in the literature, the discriminative power of deepfake production methods is lower than that of methods based on simple spectrogram analyses. Therefore, the use of transforming the MFCC feature to closely resemble human hearing perception and representing it with numerical data is gaining importance.

In studies conducted on the FoR dataset, the division of the dataset into subsets due to its length, and the shorter duration and more controlled nature of CNN or classical classifiers, have enabled them to achieve high accuracy. However, the reporting of 100% accuracy in all subsets requires examination for possible data leakage [18].

The limited number of speakers in the dataset used, as in the DEEP-VOICE dataset, may restrict the variety of speech styles. In this case, the model should be evaluated using samples generated by different deepfake production tools, and with different languages and accents [15].

Generally, when considering the studies conducted, MFCC-based features provide a fast and interpretable detection advantage with classical machine learning models. CNN and CNN-LSTM architectures achieve higher accuracy, especially on complex datasets [17].

Overall, the sample size used in this study is considered a small dataset in terms of the performance of machine learning models. Models run on limited, small datasets, as used in this study, risk learning the training data-specific structure and

performing poorly on new data (overfitting). This study demonstrates that relying solely on acoustic feature groups for deepfake audio detection may be insufficient; instead, an approach using spectral, temporal, and prosodic features together provides more reliable results. However, if the applied methods have limited resources and limited datasets, classical machine learning methods, such as Random Forest, can be used for detection, while for complex datasets with many records, better results are expected can be obtained by preferring more powerful models such as CNN-LSTM.

Authors' Contributions

Conceptualization of the study N.Y., development of the methodology K.Z, data collection N.Y., K.Z., data analysis K.Z., software applications, visualization, and writing of the article were performed by the authors. The authors have read and approved the final version of the article.

Artificial Intelligence Statement

At the time of writing this article, artificial intelligence was used only for English grammar corrections.

Copyright Statement

Conflicts of Interest

No conflict of interest or common interest has been declared by authors.

References

[1] Chesney, R. and Citron, D. (2019) Deepfakes and the New Disinformation War. *Foreign Affairs*, **98**, 147-155.

[2] Güngör, M. (2024) Yapay zekânin suç potansiyelinin değerlendirilmesi. *Bilişim Hukuku Dergisi*, **6**, 620-660. https://doi.org/10.55009/bilisimhukukudergisi.1498256

[3] Kıvanç, T. and Çötok Akıncı, N. (2024) Dijital iletişim ve deepfake. In: İşman, A., Öztunç, M. and Akıncı, Çötok N., Eds., *Iletişim Çalişmalari.* Eğitim Yayınevi, 66-74.

[4] Sever, H. (2008) Adli ses incelemeleri ve hukuki boyutu. Adalet.

[5] Ullah, R., Asghar, I., Malik, H., Evans, G., Ahmad, J. and Roberts, D.A. (2024) Enhancing Forensic Audio Transcription with Neural Network-Based Speaker Diarization and Gender Classification. 2024 *International Conference on Engineering and Emerging Technologies* (*ICEET*), Dubai, 27-28 December 2024, 1-6. https://doi.org/10.1109/iceet65156.2024.10913726

[6] Liu, L., Li, W., Morris, S. and Zhuang, M. (2023) Knowledge-Based Features for Speech Analysis and Classification: Pronunciation Diagnoses. *Electronics*, **12**, Article 2055. https://doi.org/10.3390/electronics12092055

[7] Muda, L., Begam, M. and Elamvazuthi, I. (2010) Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW)

Techniques. arXiv:1003.4083.

[8] Rabiner, L. and Schafer, R. (2010) Theory and Applications of Digital Speech Processing. Prentice Hall Press.

[9] Yan, Y., Simons, S.O., van Bemmel, L., Reinders, L.G., Franssen, F.M.E. and Urovi, V. (2025) Optimizing MFCC Parameters for the Automatic Detection of Respiratory Diseases. *Applied Acoustics*, **228**, Article 110299. https://doi.org/10.1016/j.apacoust.2024.110299

[10] Wagner, M. and Watson, D.G. (2010) Experimental and Theoretical Advances in Prosody: A Review. *Language and Cognitive Processes*, **25**, 905-945.

[11] Jurafsky, D. and Martin, J.H. (2025) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd Edition, Prentice Hall PTR, 4-37. https://web.stanford.edu/~jurafsky/slp3

[12] Biau, G., Deevroye, L. and Lugosi, G. (2008) Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research*, **9**, 2015-2033.

[13] Shanmugam, M., Ismail, N.N.N., Magalingam, P., Hashim, N.N.W.N. and Singh, D. (2023) Understanding the Use of Acoustic Measurement and Mel Frequency Cepstral Coefficient (MFCC) Features for the Classification of Depression Speech. In: Al-Sharafi, M.A., Al-Emran, M., Tan, G.WH. and Ooi, KB., Eds., *Studies in Computational Intelligence*, Springer, 345-359. https://doi.org/10.1007/978-3-031-48397-4_17

[14] Hamza, A., Javed, A.R.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z., *et al.* (2022) Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access*, **10**, 134018-134028. https://doi.org/10.1109/access.2022.3231480

[15] Bohara, R. and Bairwa, A.K. (2025) Detecting Deepfake Audio Using Spectrogram-Based Machine Learning Approaches. *IEEE Access*, **13**, 149478-149489. https://doi.org/10.1109/access.2025.3602531

[16] Iqbal, F., Abbasi, A., Javed, A.R., Jalil, Z. and Al-Karaki, J. (2022) Deepfake Audio Detection via Feature Engineering and Machine Learning. In: Drakopoulos, G. and Kafeza, E., Eds., *Proceedings of the CIKM 2022 Workshops. CEUR Workshop Proceedings*, Vol. 3318. https://ceur-ws.org/Vol-3318/paper4.pdf

[17] Ali Sharafat, H.M., Rivzi Muslim, S.M., Tariq, H., Majeed, S., Tariq, A. and Iq-bal, M.M. (2025) AI-Based Deepfake Audio Detection Technique from Real and Fake Audio Dataset. *Journal of Computing & Biomedical Informatics*, **8**.

[18] C, S.K., G, S.N.R., Patnam, V., Deepak, S., M. V, S.P. and G, T.R. (2024) Leveraging Deep Learning Architectures for Deepfake Audio Analysis. *Proceedings of 6th International Conference on Smart Systems and Inventive Technology*, **19**, 266-276. https://doi.org/10.29007/sl8m

[19] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., *et al.* (2019) ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. *Interspeech* 2019, Graz, 15-19 September 2019, 1008-1012. https://doi.org/10.21437/interspeech.2019-2249

[20] Gessinger, I., Raveh, E., Steiner, I. and Möbius, B. (2021) Phonetic Accommodation to Natural and Synthetic Voices: Behavior of Groups and Individuals in Speech Shadowing. *Speech Communication*, **127**, 43-63. https://doi.org/10.1016/j.specom.2020.12.004