scientific reports

Check for updates

OPEN

Novel transfer learning based acoustic feature engineering for scene fake audio detection

Ahmad Sami Al-Shamayleh[1], Hafsa Riasat[2], Ala Saleh Alluhaidan[3]✉, Ali Raza[4]✉, Sahar A. El-Rahman[5] & Diaa Salama AbdElminaam[6,7]

Audio forensics plays a major role in the investigation and analysis of audio recordings for legal and security purposes. The advent of audio fake attacks using speech combined with scene-manipulated audio represents a sophisticated challenge in fake audio detection. Fake audio detection, a critical technology in modern digital security, addresses the growing threat of manipulated audio content across various applications, including media, legal evidence, and cybersecurity. This research proposes a novel transfer learning approach for fake audio detection. We utilized a benchmark dataset, SceneFake, that contains 12,668 audio signal files for both real and fake scenes. We propose a novel transfer learning method, which initially extracts mel-frequency cepstral coefficients (MFCC) and then class prediction probability value features. The newly generated transfer features set by the proposed MfC-RF (MFCC-Random Forest) are utilized for further experiments. Results expressed that using the MfC-RF features random forest method outperformed existing state-of-the-art methods with a high-performance measure accuracy of 0.98. We have tuned hyperparameters of applied machine learning approaches, and cross-validation is applied to validate performance results. In addition, the complexity of the computation is measured. The proposed research aims to enhance the accuracy measure, and efficiency of identifying manipulated audio content, thereby contributing to the integrity and reliability of digital communications.

Audio is a field within forensic science focused on the collection, examination, and assessment of audio recordings, which can serve as evidence in legal proceedings[1]. This field encompasses a variety of techniques, such as enhancing audio quality, identifying the source of a recording, detecting edits or alterations, and authenticating the integrity of an audio file. An audio fake attack[2] using speech with scene-manipulated audio represents a sophisticated challenge in the realm of fake audio detection. In such an attack, adversaries generate fake audio samples by manipulating scenes or contexts within the audio data, thereby creating a convincing yet deceptive narrative. This type of attack can involve altering the background sounds, speaker identity, or the speech content itself to mislead listeners and detection systems[3].

Fake audio detection, a critical technology in modern digital security, addresses the growing threat posed by manipulated audio content across various applications[4]. In the context of acoustic scenes, which encompass diverse acoustic environments, the ability to detect fake audio is paramount. Acoustic scenes play a vital role in intelligent wearable devices, context-aware services, and robotics navigation systems, all of which rely on accurate acoustic scene classification and recognition to interpret user situations. Misuse of these systems with manipulated audio can lead to significant harm[5], such as misinformation, privacy breaches, and incorrect navigational guidance. For instance, an intelligent wearable device that relies on audio cues to provide real-time assistance could be misled by fake audio, compromising user safety. Similarly, context-aware services that offer personalized experiences based on acoustic scenes could deliver inappropriate responses if they process fake audio inputs. In robotics, navigation systems that depend on authentic acoustic signals for situational awareness could be directed incorrectly, posing operational risks[6].

[1]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19328, Jordan. [2]Department of Computer Science/SST, University of Management and Technology, Lahore 54770, Pakistan. [3]Department of Information Systems, College of Computer and Information Science, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia. [4]Department of Software Engineering, University Of Lahore, Lahore 54000, Pakistan. [5]Computer Systems Program-Electrical Engineering Department, Faculty of Engineering-Shoubra, Benha University, Cairo, Egypt. [6]Faculty of Computers, Misr International University, Cairo, Egypt. [7]Jadara Research Center, Jadara University, Irbid 21110, Jordan. ✉email: asalluhaidan@pnu.edu.sa; ali.raza.scholarly@gmail.com

Speech signal analysis involves examining the characteristics and properties of spoken language captured in audio signals[7]. This analysis can include a variety of techniques to process and interpret acoustic features, such as frequency, amplitude, and temporal patterns, which are essential for understanding and processing human speech. Advanced methods in speech signal analysis often employ machine learning and artificial intelligence to enhance accuracy and efficiency[8,9]. One crucial application of speech signal analysis is fake audio detection, which aims to identify and differentiate between authentic and manipulated or synthetic audio recordings. Machine learning has significantly advanced the detection of fake audio, leveraging features such as MFCCs[10]. MFCCs are a widely used method for representing the short-term power spectrum of an audio signal. They effectively capture the key features of sound, making them ideal for differentiating between authentic and fabricated audio. This approach enhances the robustness and reliability of audio authentication systems, contributing to combating misinformation and ensuring the integrity of audio content.

This research presents an innovative transfer learning approach for preventing audio fake attacks. The acoustic signal dataset is utilized for model building, and MFCC features are extracted. Several advanced machine and deep neural networks are evaluated during the experimental comparisons. The proposed research aims to enhance the accuracy, measure, and efficiency of identifying manipulated audio content, thereby contributing to the integrity and reliability of digital communications.

The conceptual alignment between the proposed transfer feature approach and the parallel absolute-relative features[11] focuses on leveraging additional information beyond standard feature extraction to improve classification. The parallel absolute-relative features construct a "relative feature" by computing relationships between utterances, emphasizing where a feature stands in relation to others rather than the absolute feature itself. This is beneficial in phonotactic language recognition, where relational comparisons enhance classification. In contrast, our transfer feature approach focuses on enhancing feature representation through transfer learning. Instead of relying on inter-feature relationships, we initially extract MFCC features, then class prediction probabilities, and finally transform these into a new feature set using MFCC-Random Forest (MfC-RF). This aims to capture both spectral characteristics and probabilistic class distribution.

The significant research contributions are followed as:

- We propose a novel transfer learning method, MfC-RF, which generates Class prediction probability features from MFCC signal features. Results show that the proposed approach helps to achieve high performance.
- We have built one deep neural network and four machine learning methods in comparison. We have tuned hyperparameters of each approach, and cross-validation is applied to validate performance results. In addition, the complexity of the computation is measured, and a state-of-the-art comparison is performed.

The remaining manuscript is formatted as "Literature review" section performed a comparative analysis with state-of-the-art approaches. Section "Proposed methodology" described the novel proposed methodology. Section "Results and discussions" evaluates the performance scores of applied methods. The research findings are summed up in "Conclusion and future work" section.

Literature review

Automatic speaker verification frameworks are vulnerable to spoofing attacks, especially those involving replay and Deep-Fake audio. To enhance discrimination, a study[12] uses the ASVspoof 2021 dataset to evaluate spoofing detection methods. The proposed framework combines deep learning and Mel-spectrogram features, employs a self-attention mechanism, and uses ResNet for final classification. The hybrid feature approach improves spoofing detection by 74.60% and 60.05%, respectively. Future research should explore more complex neural network architectures and feature fusion for improved system security. The hybrid feature framework with a self-attention mechanism significantly enhances spoofing detection, offering a promising direction for improving ASV system security against sophisticated spoofing attacks.

The AVA-CL model, a multi-modal approach to deepfake detection, uses audio-visual inconsistencies to accurately identify and distinguish fake content[13]. The model uses feature fusion and contrastive learning to match audio and visual features, capturing intrinsic correlations and inconsistencies. It outperforms many state-of-the-art methods but faces limitations like facial flickering in fake videos. Future work will focus on forgery localization and improving interpretability to address existing limitations and improve detection accuracy. The AVA-CL model represents a promising multi-modality approach to deepfake detection, leveraging audio-visual inconsistencies to accurately identify and distinguish fake content.

This article[14] presents a method for identifying deepfake audio, aiming at synthetic speech data generated by Text-to-Speech (TTS) algorithms. The Vocal Emotion Analysis (VEA) Network was trained on a dataset containing emotional expressions, and a supervised classifier was used to distinguish between real and synthetic speech tracks. The system demonstrated high effectiveness in detecting deepfake audio, confirming emotional content as a strong discriminative feature. Further research is needed to integrate additional semantic features and improve performance in diverse auditory environments. This research offers a novel and effective approach to deepfake audio detection, advancing the field by highlighting the significance of emotional content in distinguishing synthetic speech, thereby enhancing the integrity of digital communications.

The study[15] evaluates the effectiveness of physical and perceptual acoustic features in detecting Deepfake audio, a threat to daily life. Using datasets from the ASVSpoof Challenge and ADD Challenge, the researchers used statistical analysis to examine the distribution differences of 16 features between real and synthesized audio. The results showed that PLP and CQCC features significantly improved detection, but most features performed poorly in Track 2, indicating the need for further refinement. Future work will involve validating the statistical characteristics analysis across different spoofing datasets and integrating distinctive feature characteristics into

detection classifier structures. This study highlights the potential of perceptual features in enhancing Deepfake audio detection and sets the stage for future advancements in the field.

This research[16] focuses on detecting and interpreting vocoder fingerprints in fake audio using datasets from eight state-of-the-art vocoders. Using t-SNE visualization, the study identified LFCC features as effective for vocoder fingerprint detection, with ResNet being the best-performing model. Future research should address limitations and enhance detection methods and interpretations to establish benchmarks and encourage innovative methods in fake audio detection.

The research[17] focuses on detecting deepfakes by identifying audio-visual inconsistencies using a multi-modal approach. It uses a benchmark called DefakeAVMiT and uses Temporal-Spatial Encoder values (TSE) for feature embedding and Multi-Modal Joint-Decoder values (MMD) for fusing audio-visual information. The proposed method includes AVoiD-DF for joint audio-visual learning, Temporal-Spatial Encoder for inconsistencies, and Multi-Modal Joint-Decoder for multi-modal interaction. Experimental results show the method outperforms existing techniques. Future work will focus on developing a universal approach and improving interpretability for better detection of multi-modal deepfakes. The AVoiD-DF method presents a promising solution for multi-modal deepfake detection, encouraging further advancements and improved interpretability in the field.

The study[18] explores the use of occlusion-based data augmentation techniques to improve voice forgery detection models. Data from the ASVspoof2017 and ASVspoof2019 competitions were used to assess the impact of these techniques on fake audio detection. The LCNN model was trained using these techniques, with Cutmix showing the best results. The study found that the effectiveness of these techniques varies depending on the dataset and conditions, suggesting no single technique is universally superior. Future work should focus on identifying suitable augmentation techniques for different datasets and improving their generalizability.

The study[19] investigates the impact of acoustic scenes and sound events on sound event detection and acoustic scene classification using multitask learning (MTL) techniques. The TUT Acoustic Scenes data of the year 2016/2017 and TUT Sound Events data of the year 2016/2017 datasets are used to evaluate the effectiveness of the proposed methods. Domain adversarial training and fake-label-based methods are used to assess the impact of cross-information on SED and ASC. The results show that while MTL methods incorporate mutual information, single-task-based methods perform better. The research suggests that future work should focus on improving MTL methods to better capture and utilize the implicit mutual benefits between acoustic scenes and sound events.

The study[20] investigates stereo faking audio, a technique where mono audio is converted to stereo to enhance perceived quality. Three stereo audio datasets were used for evaluations. MFCC was analyzed to distinguish between real and fake stereo audio. An identification algorithm based on 80-dimensional MFCC features and a Support Vector Method (SVM) was proposed. The algorithm effectively detected stereo-faking audio, showing potential for further advancements in audio forensics.

The research[21] proposes a dataset called SceneFake, which focuses on scene-fake audio detection. It includes manipulated audio generated by tampering with real utterances using speech enhancement technologies. The study analyzes fake attacks with numerous technologies and signal-to-noise ratios to evaluate detection effectiveness. Benchmark results show that models trained on the ASVspoof of the year 2019 dataset do not reliably detect scene fake utterances. While they perform well on the SceneFake training set values and seen testing set values, their performance on unseen data is poor. The study acknowledges limitations in current work and suggests further research to improve detection methods and explore additional speech enhancement technologies.

The study[22] focuses on improving acoustic event detection (AED) by incorporating scene conditioning through a multitasking network that performs acoustic scene classification (ASC). The proposed method uses predicted scenes from ASC as additional features for AED and introduces a fake-scene-conditioned loss to improve training efficiency. Experimental results show a 23% increase in the F1 score and a 56% decrease in false alarm rate for scenes without events. The study suggests further refinement for broader applicability and efficiency, with future research focusing on addressing any unidentified limitations and exploring additional applications.

The study[23] focuses on detecting fake audio, specifically low-quality and partially fake ones, using mismatched data for training. Unsupervised pretraining models were used to analyze and detect these audios, demonstrating their effectiveness in this domain. The results showed an EER of 32.80% for low-quality data fake audio detection and 4.80% for partially fake audio detection, with the latter ranking first in the competition. However, the study does not address potential limitations across different scenarios or datasets. Future research should explore improving detection methods and examining the generalizability of the approach.

Proposed methodology

In this proposed study, we utilized a benchmark dataset named SceneFake, containing real and fake audio signals. Figure 1 represents the steps of the proposed methodology. During our method's steps, we applied basic preprocessing steps on the signal dataset. The MFCC features are then extracted from the preprocessed dataset. Subsequently, we applied the novel proposed transfer learning feature engineering approach. The newly generated rich-level features are then split into training (80%) and testing (20%) portions. We built several machine and deep neural learning models on the training dataset and evaluated their performance measures using the testing data. Additionally, we tuned the hyperparameters of the applied models. Finally, the hyperparameter-tuned model was utilized to detect fake audio.

Scene fake and real audio signals

This research utilized a benchmark dataset named SceneFake[21] for conducting the research experiments. The dataset contains 12,668 audio signal files for both real and fake scenes. The manipulated fake audio in the

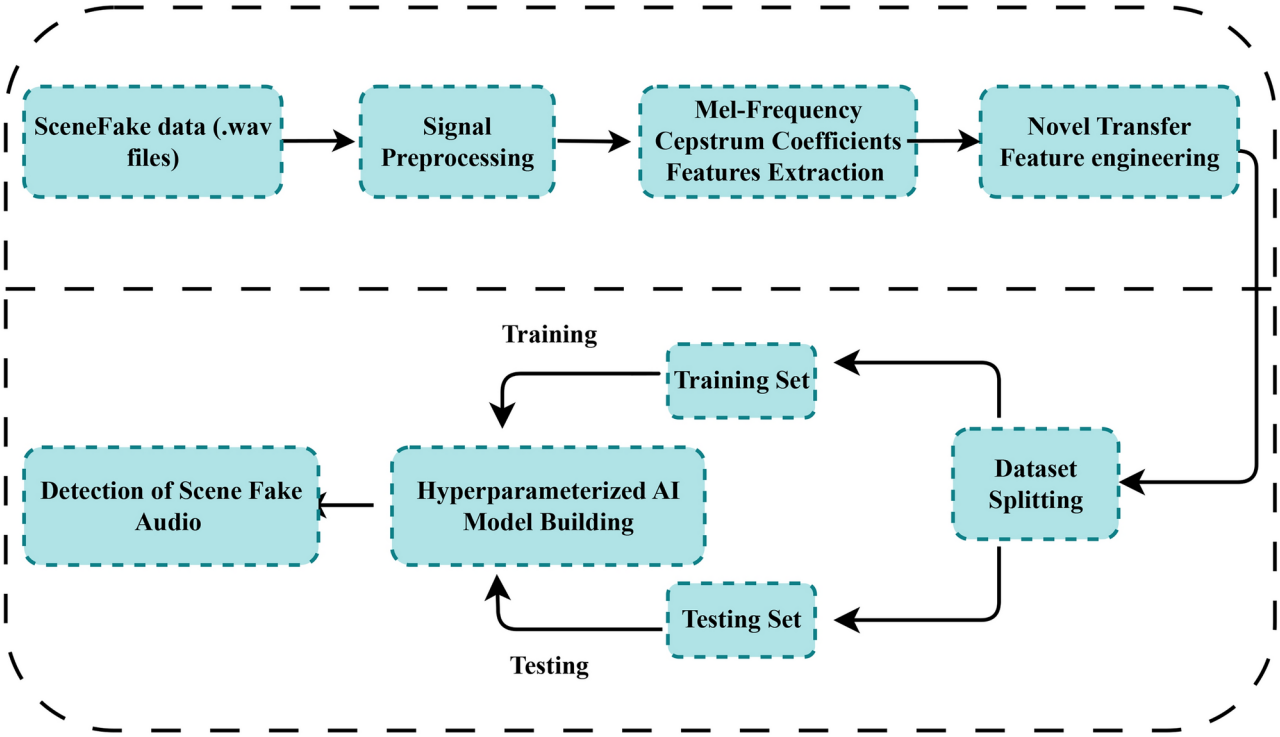


Figure 1. The architectural workflow of the proposed methodology for detecting fake audio.

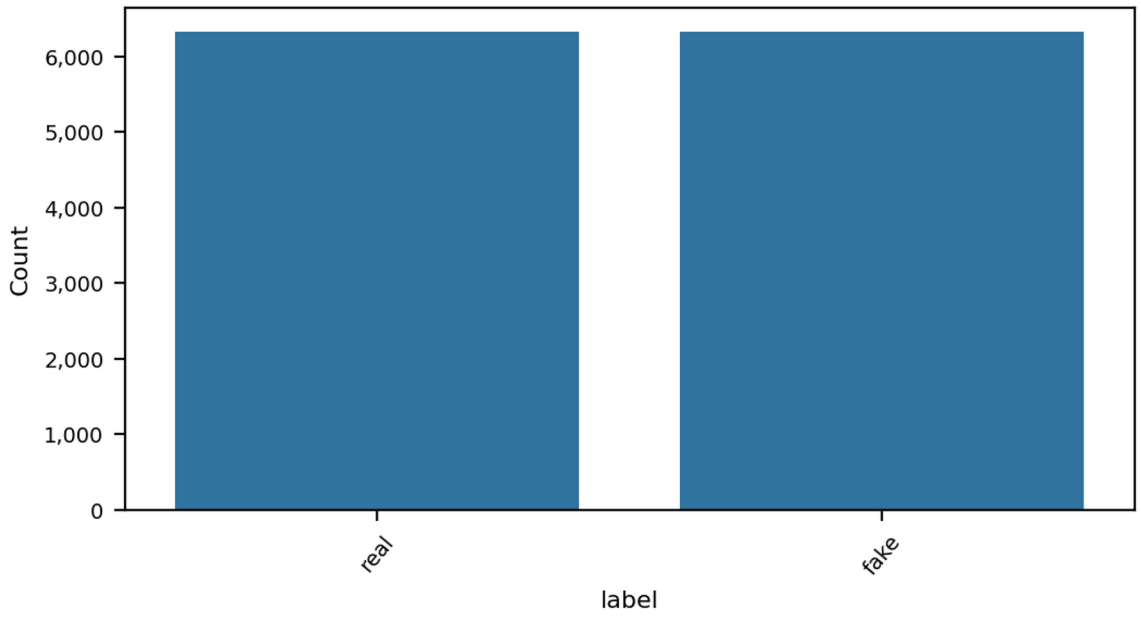


Figure 2. The target class distribution results.

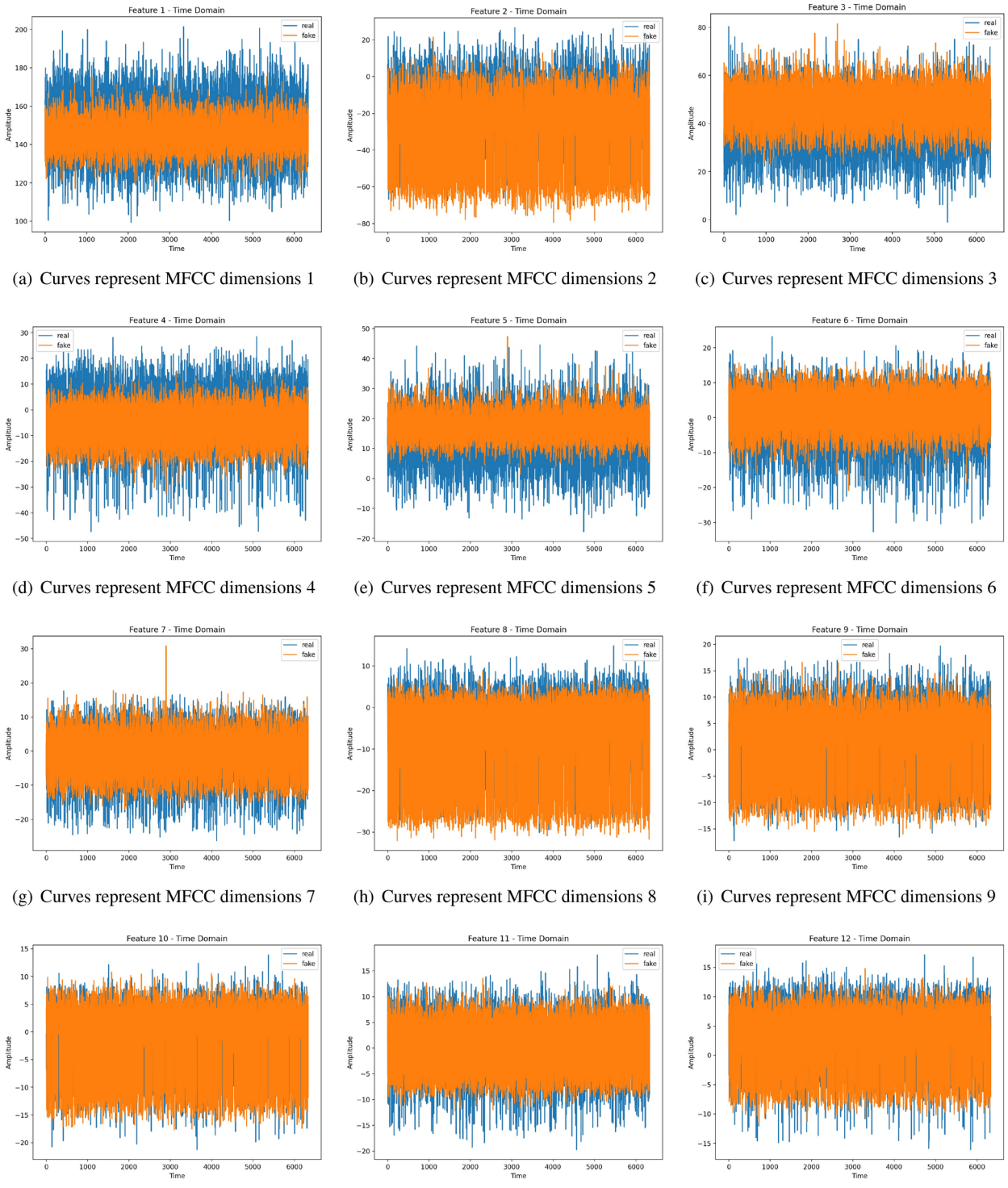
dataset is generated by tampering values with the acoustic scene of a real utterance using speech enhancement technologies. We have built several classifiers on this dataset and evaluated the results.

Audio preprocessing

We applied initial preprocessing steps to the dataset, which included selecting only audio files with the .wav extension and discarding the rest. The bar chart in Fig. 2 expressed the distribution of the dataset, with 6,334 files labelled as real and 6,334 files labelled as fake. We then encoded the target class using LabelEncoder, transforming the labels such that fake is encoded as 1 and real as 0.

MFCC feature extraction

In our research, we applied a signal features extraction approach to the dataset using the MFCC feature extraction approach. We utilized the librosa module to load and extract features from the audio signals. Specifically, we extracted 13 MFCC features for each audio signal, which are then used for further research experiments. The sample frequency-time domain graph of MFCC features is illustrated in Fig. 3.



(a) Curves represent MFCC dimensions 1 (b) Curves represent MFCC dimensions 2 (c) Curves represent MFCC dimensions 3

(d) Curves represent MFCC dimensions 4 (e) Curves represent MFCC dimensions 5 (f) Curves represent MFCC dimensions 6

(g) Curves represent MFCC dimensions 7 (h) Curves represent MFCC dimensions 8 (i) Curves represent MFCC dimensions 9

(j) Curves represent MFCC dimensions 10 (k) Curves represent MFCC dimensions 11 (l) Curves represent MFCC dimensions 12

Figure 3. The plotted curves represent MFCC dimensions 1–12.

Novel transfer feature generation

The novel proposed transfer feature generation approach MfC-RF is analyzed in this section. The workflow architecture of feature generation is illustrated in Fig. 4. The following steps are performed during the transfer learning mechanism:

1. **Step-1**: Initially, the benchmark SceneFake data contains audio signal files for both real and fake scenes and is prepared for input.
2. **Step-2**: The signal dataset is then input to the Librosa module for MFCC features extraction.
3. **Step-3**: The extracted MFCC features with dimension (12668 rows $\times$ 13 columns) are then formed and input into a random forest model.
4. **Step-4**: Then the class prediction probability features with dimension (12668 rows $\times$ 2 columns) are generated by the random forest model from MFCC features.
5. **Step-5**: Finally, the transfer features are used to build the applied machine learning methods.

This proposed approach helps to achieve high-performance results for detecting the scene's fake audio.

Class prediction probability features overview

- Let n be the total number of samples in the dataset.
- Let $\mathrm{X} = \{x_1, x_2, \ldots, x_n\}$ represent the feature set, where $x_i \in \mathbb{R}^d$ (each sample has d features).
- Let $\mathrm{Y} = \{y_1, y_2, \ldots, y_n\}$ represent the corresponding class labels, where $y_i \in \{1, 2, \ldots, C\}$ for C classes.

The Random Forest model $\mathcal{F}$ is an ensemble of T decision trees:

$$\mathcal{F} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\},$$

where each $\mathcal{T}_t$ is a decision tree.

For a given sample x:

1. Each decision tree $\mathcal{T}_t$ produces a probability distribution over the classes:

$$\mathrm{P}_t(x) = \{P_t(y = c \mid x) \;\mid\; c \in \{1, 2, \ldots, C\}\},$$

where $P_t(y = c \mid x)$ is the fraction of samples in the leaf node of $\mathcal{T}_t$ that belong to class c.

2. The final probability for class c is the average of the probabilities across all trees:

$$P(y = c \mid x) = \frac{1}{T} \sum_{t=1}^{T} P_t(y = c \mid x).$$

3. This ensures that the predicted probabilities are normalized:

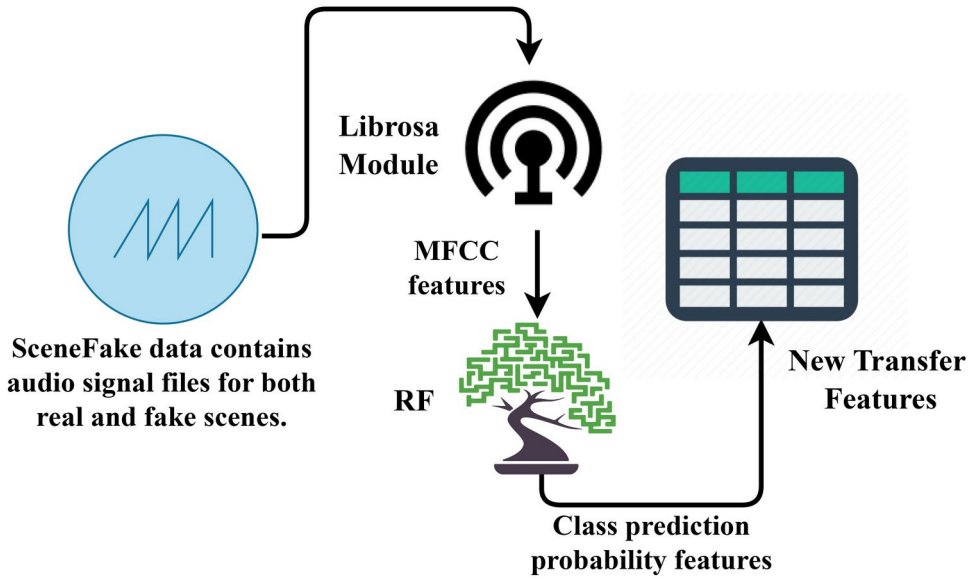


Figure 4. The novel transfer feature generation mechanism.

$$\sum_{c=1}^{C} P(y = c \mid x) = 1.$$

In this research, for the first time, we have extracted high-level class prediction probability features derived from the MFCC features. These novel features capture richer and more discriminative information about class likelihoods, effectively leveraging the strengths of MFCC while incorporating probabilistic insights from the classification model. This enhanced representation improves the model's ability to differentiate between classes, leading to better overall performance.

Artificial intelligence approaches used for detection

This section explores artificial intelligence (AI) approaches[24–28] for scene fake audio detection using audio acoustic signal data. Leveraging advanced AI techniques, we aim to identify and distinguish between real and manipulated audio scenes. By employing various machine learning algorithms and deep learning classifiers, we analyze the acoustic signals to detect anomalies indicative of fake audio. Our approach utilizes transfer learning features extracted from the audio data to train classifiers. The results demonstrate the potential of AI in enhancing the accuracy and reliability of detecting scene-manipulated fake audio.

- *Random Forest (RF)*: method for scene fake audio detection involves creating an ensemble of decision trees, each trained on numerous subsets of the audio acoustic signal data. Each tree in the forest independently classifies the audio as real or fake based on extracted features, such as MFCCs. The final prediction is obtained by aggregating the individual tree outputs through majority voting, enhancing the robustness and accuracy of the detection model.
- *K-Neighbors Classifier (KNC)*: method for scene fake audio detection utilizes the k-nearest neighbors algorithm to classify audio acoustic signal data. This approach involves calculating the distance between a given audio sample and all other samples in the dataset. The method then identifies the 'k' closest samples (neighbors) and assigns the most common class among these neighbors to the given audio sample, effectively distinguishing between real and fake audio scenes based on their acoustic characteristics.
- *Logistic Regression (LR)*: method for scene fake audio detection involves modelling the probability of an audio signal being fake or real. It utilizes the logistic function to map the linear combination of extracted audio features, such as MFCCs, into a probability value between 0 and 1. The model is trained on labelled audio acoustic signal data, optimizing the weights of the features to minimize the classification error, allowing for the effective detection of fake audio scenes.
- *Gaussian Naive Bayes (GNB)*: method for scene fake audio detection operates on the principle of Bayes' theorem, assuming independence among features[29,30]. Each audio acoustic signal's feature is modelled using a Gaussian (normal) distribution. The method calculates the probability of an audio signal belonging to either the real or fake class based on the extracted features and classifies the signal by selecting the class with the highest posterior probability.
- *Long Short Term Memory (LSTM)*: networks are employed for scene fake audio detection by analyzing sequential audio acoustic signal data. The mathematical workings of LSTM involve the use of gates (input, forget, and output) to manage the flow of information and maintain long-term dependencies within the data. By learning patterns and temporal relationships in the acoustic signals, LSTM models effectively distinguish between real and manipulated audio scenes, enhancing detection accuracy.

Hyperparameter tuning

The tuning parameters of applied classification models and neural network approaches are described in Table 1. The best-fit parameters of applied models are determined through recursive testing, training mechanisms, and k-fold cross-validation splits. The analysis shows that models improve performance accuracy for fake audio detection.

Results and discussions

This section offers a thorough analysis of the evaluation metrics, experimental setup, and the outcomes achieved using the proposed approach. In this study, we evaluated several machine and deep neural network learning models for detecting deep fake audio using a dataset of authentic and synthesized speech samples. The models are assessed based on key metrics such as accuracy, precision, recall, and F1 score.

Method	Hyperparameters tuning
RF	max_depth value=300, criterion value="gini", n_estimators value=300, splitter value="best"
LR	penalty value='l2', to value=1e-4, max_iter value=100, solver='lbfgs'
KNC	weights value='uniform', n_neighbors value=2, leaf_size value=30, metric value='minkowski'
GNB	var_smoothing value =1e-09
LSTM	activation= 'sigmoid', optimizer='adam', loss value='binary_crossentropy', metrics value=['accuracy']

Table 1. Parameters tuning analysis of applied classification models and neural network models for fake audio analysis.

Experimental setup

The experimental setup for detecting deep fake audio utilized a dataset of authentic and synthesized speech samples. The equipment included a high-performance computing cloud of Google Colab Jupyter Notebook, specifically with 90 GB of disk space and 13 GB RAM. The frameworks employed are TensorFlow and Python programming language. For signal data pre-processing, we leveraged the LibROSA library.

Evaluation measures

Evaluation metrics play a crucial role in assessing the performance of machine learning models. We have concentrated on several important assessment criteria in this study to determine how well our proposed model works:

- *Accuracy*: gauges how well the model predicts things generally. It determines the proportion of accurately identified samples to all samples. When evaluating something, accuracy alone isn't always enough, particularly when working with datasets that are imbalanced or when different kinds of errors have varied outcomes;

$$\text{Accuracy} = \frac{TPvalue + TNvalue}{TPvalue + FPvalue + TNvalue + FNvalue} \tag{1}$$

- *Precision* measures how well the model can distinguish the true positive samples from the anticipated positives. The proportion of actual positives to the sum of both false positives and true positives is calculated. The accuracy of positive forecasts is the main emphasis of precision;

$$P = \frac{TP}{TP + FP} \tag{2}$$

- *Recall*: also known as sensitivity or the true positive rate, evaluates the model's ability to correctly identify positive samples from all actual positives. It is calculated by dividing the number of true positives by the sum of true positives and false negatives. The completeness of good predictions is the subject of recall;

$$R = \frac{TP}{TP + FN} \tag{3}$$

- *F1 Score*: is the harmonic mean of precision and recall. It is particularly useful in scenarios with imbalanced class distributions or when it is important to equally prioritize both types of errors, as it provides a single metric that balances precision and recall. The F1 score is a number between 0 and 1, where 1 represents good performance.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

Performance analysis with MFCC features

Table 2 presents the performance measures (accuracy, precision, recall, and F1 score) of various classifiers on a voice signal dataset. RF achieved an accuracy of 92%, with precision, recall, and F1 scores of approximately 0.93, 0.91, and 0.92, respectively, indicating robust performance across metrics. KNC showed an accuracy of 90%, with precision, recall, and F1 scores around 0.88, 0.92, and 0.90, respectively, demonstrating good recall but slightly lower precision and F1 score. LR and GNB both yielded an accuracy of 79%. LR had precision, recall, and F1 scores of approximately 0.82, 0.74, and 0.78, respectively, while GNB showed precision, recall, and F1 scores around 0.86, 0.71, and 0.78, highlighting higher precision for GNB but lower recall compared to LR. The deep learning model LSTM also shows moderate performance scores, as shown in Fig. 5. In summary, the RF method

Method	Accuracy	Target class	Precision	Recall	F1
RF	0.92	Real	0.93	0.91	0.92
		Fake	0.91	0.93	0.92
KNC	0.90	Real	0.88	0.92	0.90
		Fake	0.92	0.87	0.87
LR	0.79	Real	0.82	0.74	0.78
		Fake	0.76	0.83	0.80
GNB	0.79	Real	0.86	0.71	0.78
		Fake	0.75	0.88	0.81
LSTM	0.94	Real	0.95	0.94	0.94
		Fake	0.94	0.95	0.94

Table 2. Classification accuracies using 13-D MFCC features for 5 different classifiers.

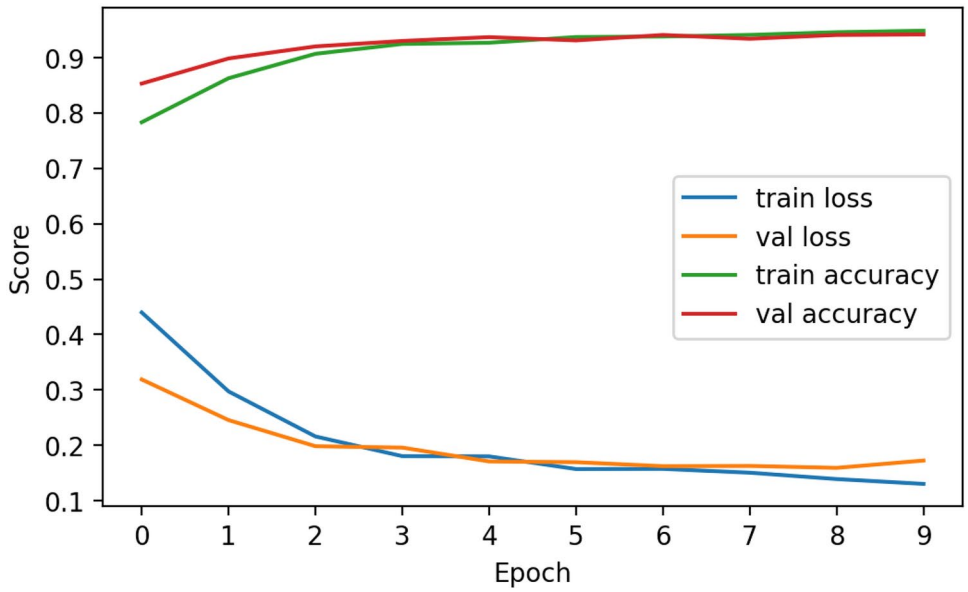


Figure 5. During the training of the LSTM model, the time series is analyzed.

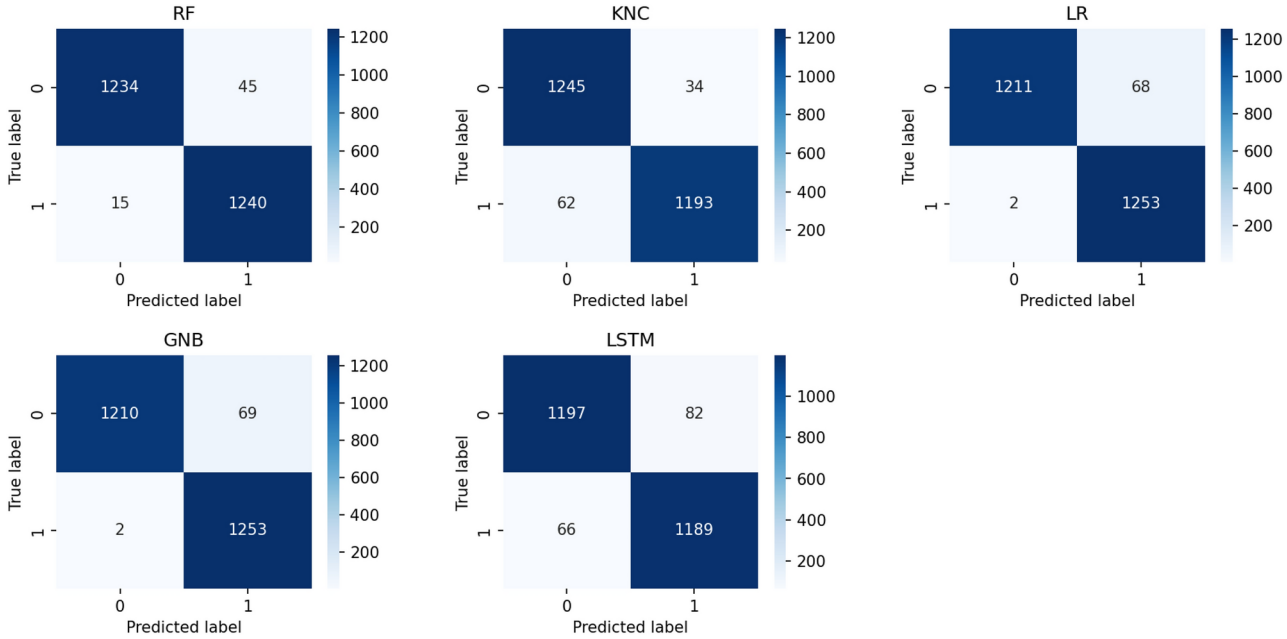


Figure 6. The confusion matrix validations analysis.

performed the best across these classifiers in terms of accuracy and balanced performance in precision and recall, making it a strong candidate for the classification of fake and real audio signals. In addition, the confusion matrix-based validation is illustrated in Fig. 6.

Performance analysis with novel transfer features

After the performance analysis with simple MFCC features, we applied a transfer learning approach for further performance enhancement in real and fake scene audio detection. Table 3 summarizes the classification performance metrics of the applied classifiers on a voice signal dataset. The analysis reveals that RF achieved the highest accuracy of 98%, with precision, recall, and F1 scores of 0.99, 0.96, and 0.98, respectively, indicating strong overall performance. KNC follows with an accuracy of 96%, showing a good balance between precision (0.95), recall (0.97), and F1 score (0.96). LR and GNB both achieved an accuracy of 97%. Logistic Regression showed perfect precision (1.00), while GNB had slightly lower precision (0.95). However, both classifiers had identical recall (0.97) and F1 scores (0.97). In summary, the RF method performed the best overall across these

Method	Accuracy	Target class	Precision	Recall	F1
KNC	0.96	Real	0.95	0.97	0.96
		Fake	0.97	0.95	0.96
LR	0.97	Real	1.00	0.95	0.97
		Fake	0.95	1.00	0.97
GNB	0.97	Real	1.00	0.95	0.97
		Fake	0.95	1.00	0.97
RF	**0.98**	**Real**	**0.99**	**0.96**	**0.98**
		Fake	**0.96**	**0.99**	**0.98**

Table 3. Classification accuracies using novel transfer features for 4 different classifiers. Significant values are in bold.

Method	K fold	Accuracy	Standard deviation (+/−)
RF	**10**	**0.9803**	**0.0026**
KNC	10	0.9653	0.0046
LR	10	0.9795	0.0048
GNB	10	0.9792	0.0030

Table 4. Performance validation and analysis of the proposed method using the K-fold cross-validation technique. Significant values are in bold.

Methods	Runtime computation (s)
RF	0.1349797248840332
KNC	0.03293132781982422
LR	0.1820213794708252
GNB	0.010112762451171875

Table 5. Comparative computations performance analysis for 4 different classifiers.

metrics, demonstrating high accuracy and balanced precision and recall, making it potentially suitable for the classification of real and fake audio signals.

Kfold validations analysis

The K-fold cross-validation results demonstrate in Table 4 that the RF classifier outperformed the other methods, achieving the highest accuracy of 98% with a standard deviation of ±0.0026, indicating both high performance and stability. The KNC algorithm followed with an accuracy of 96% and a lower standard deviation of ±0.0046, reflecting consistent results. LR and GNB had lower accuracies of 97% and 97%, respectively, with higher standard deviations, suggesting that these models are less robust in comparison. The overall analysis highlights the effectiveness and reliability of RF in this context.

Complexity computational analysis

The computational complexity analysis reveals significant differences in runtime among the methods in Table 5. The RF classifier, while offering the highest accuracy, has a moderate runtime at approximately 0.13 seconds, reflecting its complexity due to multiple decision trees and the ensemble method. KNC is much faster, with a runtime of just 0.033 seconds, attributed to its straightforward distance-based approach. LR takes 0.182 seconds, balancing between complexity and speed, while GNB is the fastest at just 0.01 seconds due to its simple probabilistic model. This trade-off between accuracy and computation time is crucial when selecting models for real-time applications.

Comparative analysis of state-of-the-art approaches

The state-of-the-art approaches comparison is analyzed in Table 6. Recent studies have demonstrated the efficacy of various deep learning architectures in achieving high accuracy levels. For instance, Hochare et al.[31] utilized Temporal Convolutional Networks and achieved an accuracy of 92%, while Camacho et al.[32] applied CNNs and reported an accuracy of 88.9%. Another study by authors in 2023[33] employed CNNs and reached an impressive 94% accuracy. In contrast, the proposed approach, leveraging Transfer Learning with the Novel MfC-RF technique, surpasses these results with a notable 98% accuracy. This comparison underscores the potential of transfer learning techniques to enhance performance beyond traditional CNN architectures, suggesting incorporating advanced methods for fake audio detection.

Ref.	Learning type	Proposed technique	Accuracy (%)
[31]	Deep learning	Temporal convolutional network	92
[32]	Deep learning	Convolutional neural network	88.9
[33]	Deep learning	Convolutional neural network	94
Proposed	**Transfer learning**	**Novel MfC-RF**	**98**

Table 6. Existing studies performance comparisons utilizing the same dataset. Significant values are in bold.

Discussions and limitations

This Study introduces a new transfer learning technique, MfC-RF, for audio fake detection that is applied on Scene Fake dataset. Using the probabilities of class prediction and MFCC features, our technique succeeded in achieving an accuracy of classification as high as 0.98. We fine-tuned the hyperparameters and cross-validation to ensure the robustness of our results. Moreover, computational complexity analysis proved our approach to be effective for practical applications.

Unlike previous studies, which have mainly relied on traditional deep learning models or handcrafted feature extraction methods, our method improves feature representation using transfer learning, which further enhances performance and generalization. Unlike existing methods, which usually suffer from high computational costs or limited adaptability, MfC-RF balances accuracy and efficiency well, making it a practical solution for fake audio detection. Our work advances the art by providing a better, more computationally efficient method to find manipulated audio content. In doing so, with reduced complexity compared to achieving classification accuracy, developing reliable and scalable AI-driven audio authentication will be significantly accelerated.

Future research directions may include studying further into integrating other feature extraction techniques as well as other deep learning models with a goal of improving the performance further. Increasing the size of the database for more diverse manipulation techniques in audio as well as more representative real-world situations would improve the adaptability and robustness of the proposed model. Real-time implementation and evaluation of this proposed method in forensics and security appears quite promising as a future research direction.

Conclusion and future work

This study introduced a novel transfer learning approach for fake audio detection. We utilized a benchmark dataset, SceneFake, that contains 12,668 audio signal files for both real and fake scenes. We propose a novel transfer learning method, which initially extracts MFCC and then class prediction probability features. The newly generated transfer features set by the proposed MfC-RF are utilized for further experiments. Results expressed that using the MfC-RF features random forest method surpasses state-of-the-art methods with a high-performance accuracy of 0.98. We have tuned hyperparameters of applied machine learning approaches, and cross-validation is applied to validate performance results. In addition, the complexity of the computation is measured. The proposed research aims to enhance the accuracy and efficiency of identifying manipulated audio content, thereby contributing to the integrity and reliability of digital communications.

Future directions

In future work, we will develop a framework-based graphical interface for real-time detection of fake audio detection. The proposed model will be deployed in the backend of the application.

Data availability

All data generated or analysed during this study is available at https://www.kaggle.com/datasets/mohammedabdeldayem/scenefake.

Received: 3 December 2024; Accepted: 4 March 2025
Published online: 08 March 2025

References

1. Zakariah, M., Khan, M. K. & Malik, H. Digital multimedia audio forensics: Past, present and future. *Multimed. Tools Appl.* **77**, 1009–1040 (2018).
2. Malik, H. Securing voice-driven interfaces against fake (cloned) audio attacks. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* 512–517 (IEEE, 2019).
3. Sahidullah, M. *et al.* Introduction to voice presentation attack detection and recent advances. In *Handbook of Biometric Anti-spoofing: Presentation Attack Detection and Vulnerability Assessment* 339–385 (2023).
4. Liz-Lopez, H. et al. Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Inf. Fusion* **103**, 102103 (2024).
5. Mubarak, R. *et al.* A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access* (2023).
6. Walter, M. R. et al. A situationally aware voice-commandable robotic forklift working alongside people in unstructured outdoor environments. *J. Field Robot.* **32**, 590–628 (2015).
7. Narayanan, S. & Georgiou, P. G. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* **101**, 1203–1233 (2013).
8. Jahangir, R. et al. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Syst. Appl.* **171**, 114591 (2021).
9. Hamza, A. et al. Deepfake audio detection via mfcc features using machine learning. *IEEE Access* **10**, 134018–134028 (2022).

10. Abdul, Z. K. & Al-Talabani, A. K. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access* **10**, 122136–122158 (2022).
11. Liu, W., Zhang, W.-Q., Li, Z. & Liu, J. Parallel absolute-relative feature based phonotactic language recognition. In *INTERSPEECH* 59–63 (2013).
12. Huang, L. & Pun, C.-M. *Self-attention and hybrid features for replay and deep-fake audio detection*. arXiv preprint arXiv:2401.05614 (2024).
13. Zhang, Y., Lin, W. & Xu, J. Joint audio-visual attention with contrastive learning for more general deepfake detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **20**, 1–23 (2024).
14. Jędrasiak, K. Audio stream analysis for deep fake threat identification. *Civitas et Lex* **41**, 21–35 (2024).
15. Li, M., Ahmadiadli, Y. & Zhang, X.-P. A comparative study on physical and perceptual features for deepfake audio detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia* 35–41 (2022).
16. Yan, X. *et al.* An initial investigation for detecting vocoder fingerprints of fake audio. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia* 61–68 (2022).
17. Yang, W. et al. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Trans. Inf. Forensics Secur.* **18**, 2015–2029 (2023).
18. Park, K. & Kwakm, I.-Y. Comparative study of data augmentation methods for fake audio detection. *Korean J. Appl. Stat.* **36**, 101–114 (2023).
19. Igarashi, A. *et al.* How information on acoustic scenes and sound events mutually benefits event detection and scene classification tasks. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* 7–11 (IEEE, 2022).
20. Liu, T. & Yan, D. Identification of fake stereo audio. arXiv preprint arXiv:2104.09832 (2021).
21. Yi, J. et al. Scenefake: An initial dataset and benchmarks for scene fake audio detection. *Pattern Recognit.* **152**, 110468 (2024).
22. Komatsu, T., Imoto, K. & Togami, M. Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 646–650 (IEEE, 2020).
23. Lv, Z., Zhang, S., Tang, K. & Hu, P. Fake audio detection based on unsupervised pretraining models. In *ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 9231–9235 (IEEE, 2022).
24. Raza, A., Munir, K., Almutairi, M. S. & Sehar, R. Novel transfer learning based deep features for diagnosis of down syndrome in children using facial images. *IEEE Access* (2024).
25. Rustam, F., Raza, A., Qasim, M., Posa, S. K. & Jurcut, A. D. A novel approach for real-time server-based attack detection using meta-learning. *IEEE Access* (2024).
26. Younas, F. et al. An efficient artificial intelligence approach for early detection of cross-site scripting attacks. *Decis. Anal. J.* **11**, 100466 (2024).
27. Raza, A. et al. Optimized virtual reality design through user immersion level detection with novel feature fusion and explainable artificial intelligence. *PeerJ Comput. Sci.* **10**, e2150 (2024).
28. Raza, A. *et al.* An improved deep convolutional neural network-based youtube video classification using textual features. *Heliyon* (2024).
29. Abu-Shareha, A. A., Qutaishat, H. & Al-Khayat, A. A framework for diabetes detection using machine learning and data preprocessing. *J. Appl. Data Sci.* **5**, 1654–1667 (2024).
30. Abu-Shareha, A., Abualhaj, M., Alshahrani, A. & Al-Kasasbeh, B. A four-state Markov model for modelling bursty traffic and benchmarking of random early detection. *Int. J. Data Netw. Sci.* **8**, 1151–1160 (2024).
31. Khochare, J., Joshi, C., Yenarkar, B., Suratkar, S. & Kazi, F. A deep learning framework for audio deepfake detection. *Arab. J. Sci. Eng.* **66**, 1–12 (2021).
32. Camacho, S., Ballesteros, D. M. & Renza, D. Fake speech recognition using deep learning. In *Applied Computer Sciences in Engineering: 8th Workshop on Engineering Applications, WEA 2021, Medellín, Colombia, October 6–8, 2021, Proceedings 8* 38–48 (Springer, 2021).
33. Wijethunga, R. *et al.* Deepfake audio detection: A deep learning based solution for group conversations. In *2020 2nd International Conference on Advancements in Computing (ICAC)*, vol. 1 192–197. https://doi.org/10.1109/ICAC51239.2020.9357161 (2020).

Acknowledgements

The authors would like to acknowledge the support of Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R234), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

ASA conceptualization, formal analysis, writing—original draft. HR conceptualization, data curation, writing—original draft. ASA methodology, formal analysis, data curation. AR software, methodology, project administration. SAE funding acquisition, investigation, visualization. DSA investigation, software, visualization. AR validation, supervision, writing—review & edit. All authors reviewed the manuscript.

Funding

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number(PNURSP2025R234), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S.A. or A.R.