IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

Detection of Synthetic Audio Using MFCC Features and Machine Learning Techniques

S. Sinduja[1], N. Narmadhavarshini[2], S. Yasotha[3]
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3]
Vivekanandha College of Engineering for Women (Autonomous), Tiruchengode, India
sindujacse@vcew.ac.in[1], narmadhavarshini6@gmail.com[2], yasothaseerangan24@gmail.com[3]

Abstract: *Deepfake content, generated or modified using advanced AI to mimic authentic media, spans across audio, video, images, and text, presenting escalating challenges in detection due to its increasing realism. Recent research has focused on addressing this issue using the Fake-or-Real dataset, a comprehensive benchmark for detecting deepfake media. By leveraging machine learning algorithms, researchers have demonstrated promising advancements in identifying deepfake audio, with the VGG-16 model achieving notable accuracy in feature extraction and classification tasks. Furthermore, support vector machines (SVM) and gradient boosting models have shown exceptional performance on specific subsets of the dataset, effectively distinguishing between real and synthetic audio. These findings highlight the potential of combining robust datasets and advanced algorithms to counter the growing threat of deepfake media in diverse applications.*

Keywords: Deepfakes, deepfake audio, synthetic audio, machine learning, acoustic data

I. INTRODUCTION

Deepfake is a combination of "deep learning" and "fake." It refers to digital content where human faces in photos, videos, or recordings are replaced with computer-generated ones. The concept first appeared on Reddit in 2017 when a user named "deepfakes" posted a manipulated video featuring a different actor's face. This technology raises various legal issues, such as violating portrait rights, harming reputations, and infringing on copyrights, which can cause both economic and reputational damage to individuals and businesses. Additionally, fake videos of politicians or governments could lead to media crises, social unrest, and national instability.Audio deepfakes, where artificial intelligence creates or alters audio to sound real, have been used in criminal activities, making their detection crucial. Identifying deepfakes in audio, video, and text is an active area of research. Between 2018 and 2019, the number of articles about deepfakes increased significantly, and by 2020, over 730 articles were expected to be published.Deepfakes pose serious risks to privacy, social security, and authenticity. While most studies focus on detecting video-based deepfakes with high accuracy, audio deepfake detection remains underexplored. Malicious audio calls generated by deepfake technology require specialized models for detection. Unlike methods combining audio and video information, audio-only classifiers are vital for detecting manipulated audio. To address this, a machine-learning-based approach using Random Forest, Decision Tree, and SVM algorithms is proposed, with comparative results analyzed using the Fake-or-Real dataset.The ASVspoof datasets, such as ASVspoof2015, ASVspoof2017, and ASVspoof2019, have significantly advanced research in speaker verification and spoofing detection. Techniques like convolutional neural networks (CNN) have been applied to detect synthetic speech using audio spectrograms, though these models may lose time-based information.

Temporal convolutional networks are better at preserving this data, improving the accuracy of detection.This research focuses on identifying deepfake audio from real audio and provides the following key contributions:

1. A transfer learning approach for better detection.
2. Detailed experiments using machine learning and deep learning models on the Fake-or-Real dataset.
3. Superior feature extraction using MFCC (Mel-Frequency Cepstral Coefficients).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

4. Results show the SVM model performed best for most subsets, while the VGG-16 model excelled on the original dataset.

The paper is structured as follows: Section II reviews previous studies, Section III explains the proposed methods, Section IV discusses experimental results, Section V presents a discussion, and Section VI concludes the research.

II. LITERATURE REVIEW

Audio deepfakes are artificially generated, edited, or synthesized audio that mimics real voices. Detecting them is crucial because they have been used in crimes like fraud in banking, customer service, and call centers. To identify audio deepfakes, it's essential to understand how they are created. These methods are categorized into replay attacks, speech synthesis, and voice conversion. Below is a summary of these types and related detection techniques. Audio forensics is a field used to verify, improve, and analyze audio evidence in criminal investigations. Before using audio as evidence in court, it must be validated to confirm its authenticity and ensure it hasn't been altered. AI and machine learning have been applied to detect audio manipulations over the last decade. For example, Long Short-Term Memory (LSTM) networks have been used to identify patterns in audio signals, while subsampling techniques simplify audio complexity for easier analysis. Replay attacks involve playing a recorded voice to impersonate someone. These attacks are detected using methods like deep convolutional networks, which analyze specific audio features. For example, these networks achieved a zero percent error rate on the ASVspoof2017 dataset, demonstrating their effectiveness**.** Speech synthesis (SS) digitally replicates human speech using computer software. Text-to-speech (TTS) is a common form of SS that converts written text into spoken language. TTS is used in applications like virtual assistants and can mimic different voices and accents. Companies like Lyrebird use deep learning models to synthesize thousands of sentences in seconds. However, creating high-quality speech databases for TTS systems is expensive. Advanced systems like Tacotron 2 and WaveNet improve speech synthesis by generating natural-sounding audio using mel spectrograms and neural networks. Despite progress, current methods for detecting audio deepfakes often face challenges such as high computational costs and limited validation. Deep learning techniques show better accuracy but require extensive training time. To overcome these issues, machine learning models can be enhanced using feature-based methods, and transfer learning-based approaches can address the complexity of larger datasets. These advancements hold promise for more efficient and accurate detection of audio deepfakes.

III. PROPOSED METHODOLOGY

In machine learning, training a model requires balancing overfitting and underfitting, which can negatively affect its ability to perform well in real-world scenarios. Overfitting occurs when the model memorizes the training data, including noise, and struggles to generalize to new data. Underfitting happens when the model fails to capture important patterns in the data. Striking this balance is challenging. A major challenge in deepfake detection is the high false-positive rate, where models incorrectly classify unseen patterns as fake because they weren't part of the training data. This issue arises from the limited size and variety of datasets, which makes it impractical to include all possible real and fake patterns. The Fake-or-Real (FoR) dataset addresses this by providing four subsets: for-original, for-norm, for-2-sec, and for-rerec. These subsets include synthetic and real audio samples, with variations in processing and duration. This research aims to create a method for identifying deepfake audio under different background noises and durations. The proposed framework includes data preprocessing, feature extraction, and classification using machine learning algorithms.

- **Data Preprocessing:** The Fake-or-Real dataset contains over 195,000 human and synthetic audio samples from sources like Deep Voice 3 and Google Wavenet. To improve data quality, preprocessing removes duplicates and invalid files, standardizes bit rates, and normalizes audio data for better model training.
- **Feature Extraction:** Since deepfake audio often mimics real signals, it's difficult to differentiate them. Features extracted from the frequency domain of audio signals can help detect deepfakes. These features improve the model's ability to classify and identify synthetic audio accurately.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

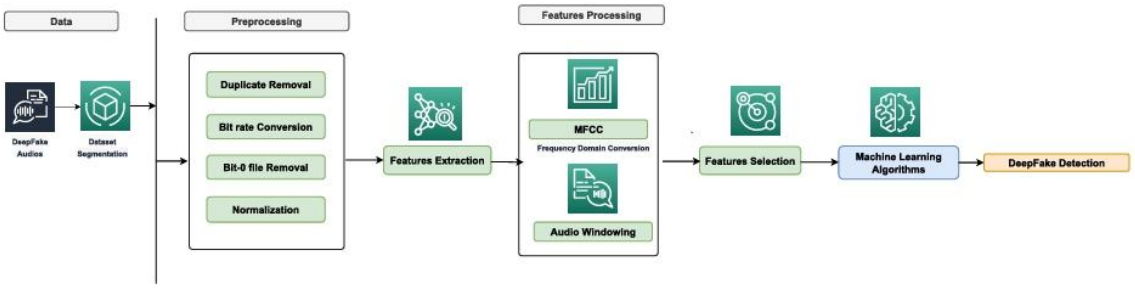


Figure 1 Graphical representation of proposed approach for detection of deepfake audios.

In this study, we utilize the Fake or Real Audio dataset alongside various features such as MFCC, cepstral, spectral, raw signal, and signal energy, with a primary emphasis on MFCC due to its effective simulation of human hearing through the use of logarithmic functions and Mel-filters.

MFCC employs triangular band-pass filters to replicate human auditory perception, and Figure 2 illustrates the MFCC audio series alongside amplitudes in decibels, as this study utilizes MFCC for deepfake audio detection by generating a vector group from each sound waveform frame with Mel-frequency cepstral coefficients and short-time Fourier transform to convert time-domain signals into the time-frequency domain.

Figure 3 illustrates the comparison of fake and genuine audio signals through their spectrograms, highlighting key auditory features that distinguish them, while detailing the feature extraction and selection process, using a sampling rate of 44100 to reduce the initial 270 features to 65 crucial characteristics for deepfake detection, with an explained variance ratio of 97% confirming the relevance of the selected data.

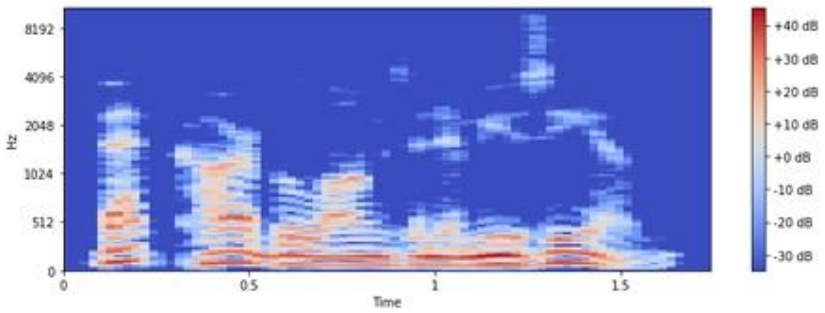


Figure 2 Melspectrogram representation of audio signal where the amplitude is depicted in terms of decibel.

Classification Models :

1. Random Forest : Random Forest is an ensemble learning method that constructs multiple decision trees from various sub-samples of the dataset and averages their results to reduce overfitting, while also calculating feature importance based on the gain and sample distribution across nodes**.** The importance of a feature can be represented as in equation 1:

$$X_j = \sum k : jY_k G_K \qquad (1)$$

The importance of each feature X_j is determined by normalizing the values for each tree in the random forest and then summing these normalized values across all trees, where the nodes k split based on feature j .

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

$$X_{j'} = \frac{X_{j'}}{\sum_z X_{j_z}} \quad (2)$$

$$RFX_{jj'} = \frac{\sum_z Xj'_{jz}}{\sum_{z,t} Xj'_{jzt}} \quad (3)$$

In equations 2 and 3, z represents all features, t represents all trees in a random forest, and Xj signifies the importance. The model uses the important features, as illustrated in Figure 3, where Xj represents the normalized feature importance for node j, RFXjj denotes the feature importance across all trees in the random forest, and Xjz indicates the normalized importance of feature j relative to tree t.

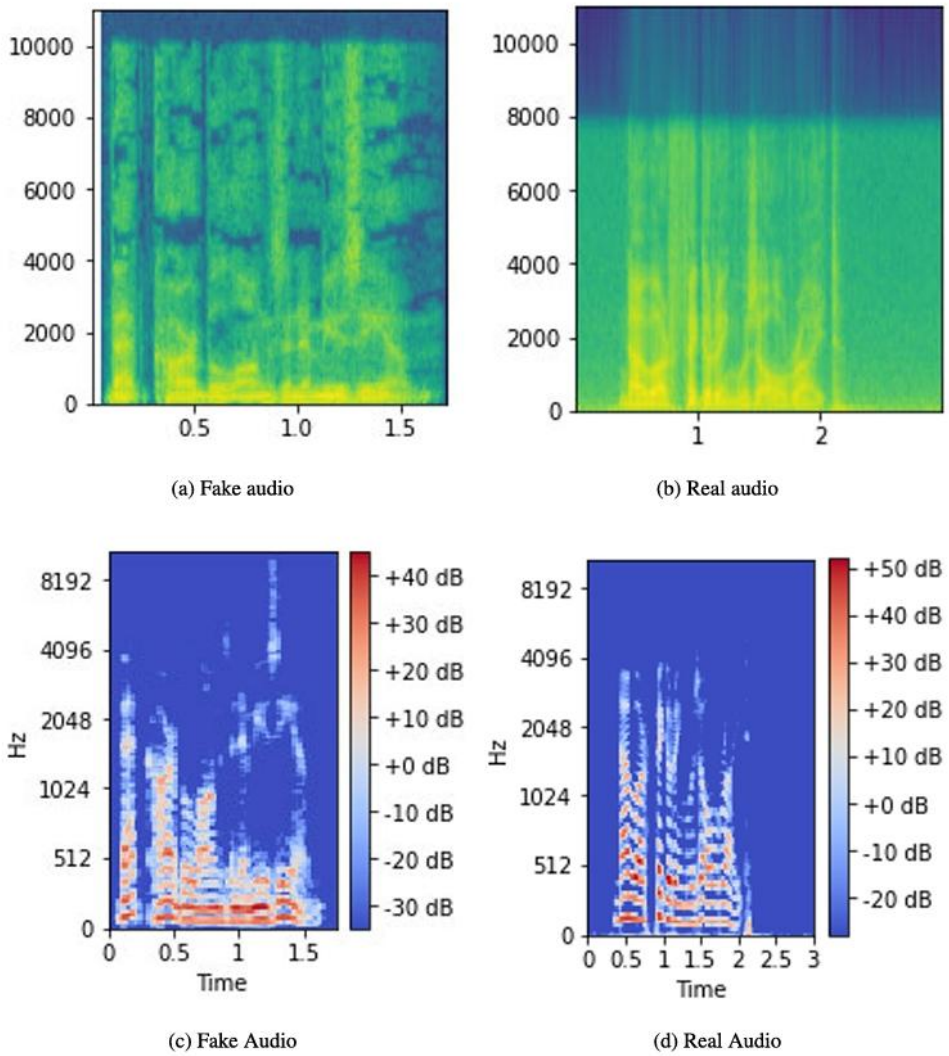


Figure 3 In (a) and (b), the comparison is shown between the deepfake and real audio signal in spectrogram where the difference in amplitude is apparent. In (c) and (d), the amplitude is shown in terms of decibels (db) for understanding the auditory parts of the audio signal.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

2) Support Vector Machine (Svm) : Support Vector Machines (SVM) is a supervised learning technique that operates under two main principles: transforming data into a high-dimensional space simplifies complex classification challenges by making them linearly separable, and it focuses on training patterns that are closest to the decision boundary, which are crucial for effective classification, such as in the binary classification of deepfake detection using the hyperplane as the decision surface**.** If x is a random vector n R, we define

$$f_{(x)} = w.x + b \qquad (4)$$

The dot product, denoted as d() in equation 4, pertains to the collection of all x-vectors that fulfill the condition f(x) = 0, represented by H0, while considering two hyperplanes, H1 and H2. the distance between them is referred to as their margin which can be represented as follows

$$\frac{2}{\|w\|} \qquad (5)$$

The optimal hyperplane H0 is determined by support vectors which are the closest to two parallel hyperplanes, ensuring maximum margin to accurately classify data points without error.

$$f_{x_i} > +1 \; for \; y_i = +1 \qquad (6)$$

$$f_{x_i} > -1 \; for \; y_i = -1 \qquad (7)$$

Hence, ndingtheSVMclassifyingfunctionH0 can be stated as follows :

$$minimize \; \frac{1}{2}\|w\|^2 \qquad (8)$$

$$y_i f_{(x_i)} \geq 1, \quad \forall_i \qquad (9)$$

The SVM was selected for its effective classification of deepfake audios due to its ability to handle high-dimensional data and maintain efficiency through the use of a training subset, although it struggles with the for-original dataset due to longer training times and noise, but performs better with clean datasets when utilizing a radial basis function kernel and the Scikit-learn library.

3) Multi-Layer Perceptron (Mlp) : The MLP model is suitable for classification tasks as it includes an input layer, a hidden layer for processing, and an output layer, and for this study, we utilize specific hyperparameters such as a hidden-layer size of 100, solvers Adam and RMSprop, with RMSprop tailored for smaller datasets, shuffling enabled, and the ReLU activation function.

4) Extreme Gradient Boosting (Xgb) : XGB is an efficient and resourceful parallel variant of gradient boosting that iteratively builds stronger models from weaker ones, uses a learning rate of 0.1 with 10000 estimators, but can struggle with outliers due to its reliance on previous predictions for accuracy correction, complicating the streamlining process.

IV. EXPERIMENTS AND RESULTS

The Fake-or-Real (FoR) dataset, compiled from approximately 195,000 human and synthetic speech samples, includes data from various Text-to-speech programs and multiple recorded human voice sources, with four public versions available: for-original, for-norm, for-2sec, and for-rerec, including raw data in the for-original folder.

The for-norm dataset contains some duplicate files but is generally well-balanced in terms of demographics (gender and socioeconomic status) and technical aspects (sample rate, volume, and multiple channels); the for-2sec variant truncates files after 2 seconds, while for-here is a re-recording of for-2sec designed to simulate an attack via a vocal channel, and we present our binary classification analysis results in Table 2, which details the findings for detecting deepfakes.

The experiments incorporated synthetic noise into the audio signals from the three datasets (for-2sec, for-norm, and for-rerec), resulting in samples increasing from 17870 to 35740 while preserving both original and noisy audio.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

A. For-Rerec Dataset : Table 2 presents the for-rerec dataset results, highlighting that among the tested machine learning models, the Support Vector Machine achieved the highest accuracy of 98.83%, with further classification results for noisy audio signals provided in Table 3.

Datasets	Size	Description
FOR-REREC DATASET	1.5 GB	It is a re-recorded version of the for-2-second dataset to simulate a scenario where an attacker sends an utterance through a voice channel (i.e., a phone call or a voice message).
FOR-2SEC DATASET	1 GB	Contains audios based on FOR-NORM dataset, but with the files truncated at 2 seconds
FOR-NORM DATASET	5.8 GB	the same files as FOR-ORIGINAL dataset, but balanced according to gender and the same sampling rate, volume, and channels for each.
FOR-ORIGINAL DATASET	7.7 GB	The audios are collected from various sources, without any modification

Table 1 Dataset description.

The results show that the MLP and SVM models achieved the highest accuracy scores of 98.66% and 98.43%, while other models like DT, LR, and XGB recorded 82.12%, 88%, and 88.92% respectively.

B. For-2sec Dataset : The for-2sec dataset, which contains complex audio at two-second intervals, facilitates easier processing by machine learning algorithms, resulting in better performance as shown in Table 2, where the MLP classifier achieved an accuracy of 94.69, surpassing other models such as Random Forest (94.44), SVM (97.57), gradient boosting (94.30), and AdaBoost (90.23).

Table 3 presents the classification results for the noisy audio signal in the for-2sec dataset, showcasing that the SVM model achieved the highest accuracy at 99.59%, followed by MLP at 99.49% and DT at 87.52%.

C. For-Norm Dataset : The dataset recorded audio at 12-second intervals, with Table 2 displaying classifier results showing Gradient Boosting achieved the highest accuracy at 92.63, while QDA and KNN underperformed with 61.36 and 64.21, respectively.

Table 3 shows that while the results from the for-norm dataset with noisy audio were lower than those of the other two datasets, the XGB model performed best among all ML models, which fared reasonably well but lacked the same level of impressiveness.

D. For-Original Dataset : The for-original datasets, which include audio samples of varying lengths, bit-rates, and noise levels, proved too complex for the machine learning models, leading us to adopt a transfer learning-based deep learning approach.

Models	for-2sec	for-norm	for-rerec
SVM	97.57	71.54	98.83
MLP Classifier	94.69	86.82	98.79
Decision Tree	87.13	62.16	88.28
Extra Tree Classifier	94.61	91.46	96.87
Gaussian Naive Bayes	88.20	81.81	81.91
Ada Boost	90.23	88.40	87.67
Gradient Boosting	94.30	92.63	93.51
XGBoost	94.52	92.60	93.40
Linear Discriminant Analysis	89.50	91.35	87.56
Quadratic Discriminant Analysis	96.13	61.36	96.91

Table 2 Accuracy comparison for machine learning models

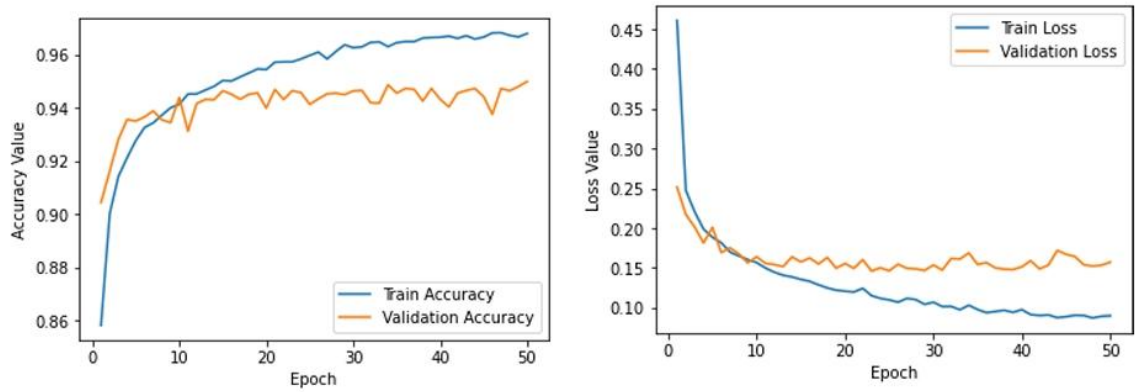
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

Models	for-2sec	for-norm	for-rerec
SVM	99.59	75.21	98.43
MLP Classifier	99.49	89.22	98.66
Decision Tree	87.52	65.10	82.12
Extra Tree Classifier	97.91	90.19	96.25
Logistic Regression	87.53	86.28	88.00
Gaussian Naive Bayes	79.77	80.16	82.14
Ada Boost	85.28	91.35	83.89
Gradient Boosting	92.29	93.50	88.86
XGBoost	92.22	94.25	88.92
Linear Discriminant Analysis	87.05	90.52	85.88
Quadratic Discriminant Analysis	96.22	65.15	95.59

Table 3 Accuracy comparison for noisy audio signals using machine learning models.

We extracted visual features of MFCC from the audio data to train VGG-16 and LSTM models for classifying deepfake and real audio, with the VGG-16 model achieving a superior testing accuracy of 93% compared to the LSTM's 91%, while also attaining a validation accuracy of 0.94 and loss of 0.14, as illustrated in Figures 4a and 4b.

E. Model Comparison : This research evaluates our proposed model's accuracy against the baseline from paper [36], ensuring that the experimental conditions, including the dataset and samples, remain identical for a valid comparison.



(a) Model accuracy (b) Model loss

Figure 4 Comparison between the validation and training (accuracy and loss).

The dataset used in this study has only been utilized once before, making comparisons with other research challenging, yet our method demonstrates promising classification accuracy, particularly with the XGBoost algorithm outperforming the baseline model, as detailed in Tables 2 and 3.

Tables 2 and 3 illustrate that our feature-based approach for training machine learning models achieved promising results on three datasets, though it struggled with the high-dimensional FOR-NORM dataset using a simple SVM algorithm, indicating that a combination of windowing techniques and MFCC could enhance performance, while our method outperformed existing approaches with a notable testing score of 93%, surpassing the best score from prior work by 26%.

The dataset in this study has only been utilized in one prior research, with similar experimental settings to existing methods, while Table 4 presents a comparative analysis of the proposed method against leading feature extraction techniques that optimally combines features from multiple approaches for classification.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

This research employs VGG16 and LSTM deep learning models with a feature ensemble of MFCC-40, roll-off point, centroid, contrast, and bandwidth features, achieving an accuracy of 93% with VGG16 and 91% with LSTM, outperforming previous studies that utilized various machine learning models and feature extraction methods.

V. DISCUSSION

This study built upon previous research on deepfake audio by expanding the Fake-or-Real dataset.

Approaches	Features	Models	Accuracy (%)
Existing Approach [36]	MFCC-20	SVM	67
		RF	62
		KNN	62
		XGB	59
Existing Approach [28]	Timbre Model Analysis (Brightness, Hardness, Depth, Roughness)	NB	67.27
		SVM	73.46
		DT (J48)	70.26
		RF	71.47
	STFT, Mel-Spectrograms, MFCC and CQT	VGG19	89.79
Proposed Approach	MFCC-40, Roll-off point, centroid, contrast, bandwidth	LSTM	91
		VGG16	93

Table 4 Comparison between results of the proposed approach and existing approach.

This advanced audio detection and classification dataset shows significant improvements with MFCC features over traditional methods, achieving a 10-20% accuracy increase, while our experiments with various machine learning algorithms, including statistical, tree-based, and boosting models, aim to further enhance performance, culminating in a VGG-16 deep learning model yielding 93% accuracy using only half of the original dataset.

VI. CONCLUSION

The detection of audio data is crucial for improving security against scams, as deepfake audios pose significant risks; this study enhances the Fake-or-Real dataset for deepfake classification by extracting MFCC features and applying various machine learning algorithms, achieving impressive accuracy rates, and plans future enhancements through exploring different techniques and testing in varied conditions.

REFERENCES

[1]. A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics, IEEE Access, vol. 10, pp. 38885–38894, 2022.
[2]. A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, Authorship identification using ensemble learning, Sci. Rep., vol. 12, no. 1, pp. 1–16, Jun. 2022.
[3]. A. Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, Privacy of web browsers: A challenge in digital forensics, in Proc. Int. Conf. Genetic Evol. Comput., Springer, 2021, pp. 493–504.
[4]. S. Ö. Ark, H. Jun, and G. Diamos, Fast spectrogram inversion using multi-head convolutional neural networks, IEEE Signal Process. Lett., vol. 26, no. 1, pp. 94–98, Jan. 2019.
[5]. Y. Chen, Y. Kang, Y. Chen, and Z. Wang, Probabilistic forecasting with temporal convolutional neural network, Neurocomputing, vol. 399, pp. 491–501, Jul. 2020.
[6]. Z. Khanjani, G. Watson, and V. P. Janeja, How deep are the fakes? Focusing on audio deepfake: A survey, 2021, arXiv:2111.14203.
[7]. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, The ASVSPOOF 2017 challenge: Assessing the limits of replay spoofing attack detection, in Proc. 18th Annu. Conf. Int. Speech Commun. Assoc., 2017, pp. 2–6.
[8]. H. J. Landau, Sampling, data transmission, and the Nyquist rate, Proc. IEEE, vol. 55, no. 10, pp. 1701–1706, Oct. 1967.
[9]. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, Deep learning for deepfakes creation and detection: A survey, 2019, arXiv:1909.11573.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

[10]. J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. Lee, V. Vestman, and A. Nautsch, ASVSPOOF 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. [Online]. Available: http://www.asvspoof.org.

[11]. Y. Kawaguchi, Anomaly detection based on feature reconstruction from subsampled audio signals, in Proc. 26th Eur. Signal Process. Conf. (EUSIPCO), Sep. 2018, pp. 2524–2528.

[12]. Y. Kawaguchi and T. Endo, How can we detect anomalies from subsampled audio signals? in Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP), Sep. 2017, pp. 1–6.

[13]. A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions, IEEE Access, vol. 10, pp. 11065–11089, 2022.

[14]. A. R. Javed, Z. Jalil, W. Zehra, T. R. Gadekallu, D. Y. Suh, and M. J. Piran, A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions, Eng. Appl. Artif. Intell., vol. 106, Nov. 2021, Art. no. 104456.

[15]. A. R. Javed, F. Shahzad, S. U. Rehman, Y. B. Zikria, I. Razzak, Z. Jalil, and G. Xu, Future smart cities: Requirements, emerging technologies, applications, challenges, and future aspects, Cities, vol. 129, Oct. 2022, Art. no. 103794.

[16]. S. Anwar, M. O. Beg, K. Saleem, Z. Ahmed, A. R. Javed, and U. Tariq, Social relationship analysis using state-of-the-art embeddings, ACM Trans. Asian Low-Resource Lang. Inf. Process., Jun. 2022.

[17]. C. Stupp, Fraudsters used AI to mimic CEO's voice in unusual cybercrime case, Wall Street J., vol. 30, no. 8, pp. 1–2, 2019.

[18]. Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, ASVSPOOF 2015: The first automatic speaker verification spoofing and countermeasures challenge, in Proc. Interspeech, Sep. 2015.