iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75200

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Audio Deepfake Detection by Using Machine and Deep Learning

Prof. Dr. Anitha G[1], SV Murali[2], Sanath U[3], Srinivas KT[4], Yeshwanth HK[5]
Computer Science and Engineering Sapthagiri College of Engineering Karnataka, India

Abstract: ***Contemporarydevelopmentsinartificial intelligence have transformed the landscape of synthetic speech technology, facilitating the creation of exceptionally convincing audio that replicates human vocal characteristics including intonation, pitch, and speaking patterns. These synthetic audio productions, commonly referredto as audio deepfakes, represent a dual-natured phenomenon with both beneficial and harmful implications. While offering valuable applications in medical treatment, accessibility solutions, educational tools, and creative industries, they simultaneously introduce substantial security concerns including financial fraud, identity deception, propaganda distribution, and digital attacks. The increasing exploitation of synthetic audio technology highlights the critical necessity for developing dependable identification mechanisms.Thisresearchexaminescontemporary scholarly work in audio deepfake identification, emphasizingcomputationallearningandneural network methodologies. We present a comprehensive analysis of prevalent feature extraction techniques, examine different identification architectures, and evaluate their comparative effectiveness. Additionally, we address fundamental obstacles including insufficient training data, cross-linguistic and synthesis method compatibility, model transparency issues, and resistance to acoustic interference. We conclude by outlining future researchpathwaysthatprioritizesystemscalability, domain flexibility, and transparent artificial intelligence solutions.***

I. INTRODUCTION

The development of artificial speech synthesis throughcomputationalintelligencehasemergedas one of the most remarkable technological achievements in recent years. Advanced neural networksystemscannowproducevocaloutputthat accuratelyreproducesindividualspeech characteristics, including accent patterns, vocal inflection,andpersonalspeakingmannerismswith suchprecisionthat distinguishingitfromauthentic human speech becomes extremely challenging. Initial voice generation systems produced mechanical and uniform audio output, but recent progress in adversarial generative networks, sequential text-to-speech architectures, and vocal replication technologies has achieved near-perfect audio quality. While these innovations were originally intended for beneficial applications, including support for individuals with vocal disabilities and enhancement of digital assistant interactions, malicious users have exploited these capabilities for destructive purposes. Documented cases have already shown how replicated voices can deceive people into authorizing significant financial transactions or can disseminate false informationbycreatingfabricatedaudiorecordings attributed to prominent individuals. The risk of extensive misuse establishes synthetic audio identification as an essential research priority. Unlike visual deepfakes, where optical inconsistencies may offer detection clues, audio deepfakes present only minor irregularities in spectral patterns, vocal rhythm, or signal phase. Identifying such subtle anomalies demands sophisticated analytical techniques, leading researchers to investigate machine learning and deep learning algorithms as fundamental instruments in this field.

II. LITERATURE REVIEW

The identification of artificially modified speech has attracted considerable research interest recently, with multiple investigations presenting approaches to distinguish between synthetic and authentic audio content. Initial methodologies focused primarily on conventional signal analysis techniques that examined spectral distortions and temporal inconsistencies. Bispectral analysis, for instance,wasutilizedtoidentifycomplexstatistical relationships that differentiate between natural human speech and artificially generated audio. Nevertheless,thesetechniqueswerefrequently restricted to particular generation methods and demonstrated limited adaptability across different systems.

Machine learning approaches broadened the scope by utilizing constructed features including Mel- Frequency Cepstral Coefficients and Linear Frequency Cepstral Coefficients. Studies showed that these characteristics effectively highlight distinctions between natural and synthetic audio signals,particularlywhenpairedwithclassification algorithms such as support vector machines or random forests. As generation techniques progressed, however, manually designed features alone proved inadequate.

The disciplinehas subsequentlytransitionedtoward deeplearningmethodologies.Convolutionalneural networks have gained widespread adoption for spectrogram analysis, as they can identify local patterns within the time-frequency representation. Recurrent neural networks, especially long short- termmemoryarchitectures,havebeenemployedto model sequential relationships in speech, which assists in detecting artificial transitions present in synthetic voices. Combined architectures merging CNNs and LSTMs have demonstrated enhanced performance byutilizingboth spatial andtemporal characteristics. Recent investigations have incorporated attention mechanisms and transformer-based architectures, drawing inspiration from natural language processing achievements,whichenablesystemstoconcentrate on essential segments of the audio sequence for classification purposes. While these models attain high precision in controlled settings, cross- generalization to unknown synthesis techniques and noisy environments continues to be an unsolved problem.

III. FEATUREEXTRACTIONINAUDIO DEEPFAKE DETECTION

A fundamental component in detection involves extracting characteristics that expose distinctions between genuine and synthetic audio. The Mel- FrequencyCepstralCoefficientstandsamongthe most commonly utilized approaches, converting audio into perceptually relevant features that correspond with human auditory frequency processing. MFCCs have demonstrated particular effectivenessbecausenumerousgenerativemodels struggle to accurately reproduce subtle vocal resonances and formant characteristics.

Spectrograms, created through short-time Fourier transformation, offer a time-frequency audio representation and are commonly employed as CNN inputs. They reveal distortions in harmonic contentandspeechtransitionsthattypicallyremain undetectabletohumanperception.TheConstant-Q Transform provides another approach, offering multi-resolution analysis and has been applied to identify inconsistencies in pitch fluctuations.

Bispectralanalysisextendsfurtherbyinvestigating correlations between frequency elements, facilitating the identification of subtle nonlinear distortions characteristic of synthesized speech. Linear Frequency Cepstral Coefficients have also found application in spoofing detection competitions, demonstrating competitive performance. These feature extraction methods collectively establish the groundwork for machine and deep learning algorithms, transforming raw audio into numerical formats suitable for effective analysis.

IV. MACHINELEARNINGANDDEEP LEARNING APPROACHES

Classical computational learning algorithms including support vector machines, decision trees, andcollectivemethodssuchasrandomforestsand gradient boosting have received extensive investigation in synthetic audio identification. These methods offer benefits in terms of computational simplicity, result interpretability, and modest processing demands. When implemented with appropriately designed input features, these algorithms can demonstrate effective performance against basic synthesis technologies.Nevertheless,theirdetection capabilities typically diminish when confronting more sophisticated generative systems.

Deep learning approaches, conversely, have becometheprevailingmethodology.Convolutional neural networks prove particularly efficient when processingspectrograminputs,astheycanidentify local spatial characteristics that expose synthetic irregularities.Recurrentneuralnetworks,including LSTMs and gated recurrent units, provide value in modeling extended temporal sequences in speech. Hybrid architectures combining CNN layers for feature identification with LSTM layers for sequence analysis have consistently achieved superior detection performance, occasionally exceeding88percentaccuracyincontrolledtesting environments.

Recent research has explored transformer-based frameworks,whichutilizeattentionmechanismsto recognize essential temporal relationships without the constraints of recurrent architectures. These models have demonstrated potential in enhancing generalization across various datasets. Ensemble learning approaches, which combine predictions from multiple algorithms, have also been implemented to improve reliability, though they may encounter difficulties when individual classifiers produce conflicting results.

V. METHODOLOGY

The approach employed in this survey relies on a comprehensive examination of current literature within the field of audio deepfake identification using machine learning and deep learning techniques. To guarantee thorough coverage, research publications were gathered from prominent digital repositories including IEEE Xplore, SpringerLink, Elsevier ScienceDirect, and ACM Digital Library, alongside open-access preprintsfromarXiv.

Theselectioncriteriafocused on relevance to the fundamental challenge of identifying synthetic speech and the application of computational methods for classification. Priority wasassignedtopublicationsfromthepreviousfive years,giventherapidadvancementindeepfake generation and detection technologies during this timeframe.

Following collection, the publications were organized into three main categories: research emphasizing feature extraction methods, studies highlighting machine learning classification techniques, and investigations examining deep learningframeworks.Thefirstcategoryfocusedon approaches utilizing Mel-Frequency Cepstral Coefficients, spectrograms, linear frequency cepstral coefficients, constant-Q transforms, and bispectral characteristics. The second category encompassed studies implementing traditional machine learning algorithms including support vectormachines,randomforests,gradientboosting, and ensemble methods. The third category concentrated on deep neural networks, specifically convolutional neural networks, recurrent neural networks, long short-term memory architectures, and hybrid CNN-LSTM systems.

Theapproachalsoinvolvedexaminingdatasetsand assessment protocols used throughout studies. Particular attention was given to datasets such as ASVspoof and proprietary collections providing both authentic and synthetic audio samples. Performance measures including accuracy, precision,recall,F1-score,andareaundertheROC curve were documented to enable fair comparison ofresultsacrossdifferentmodels.Additionally,the survey investigated how studies addressed multi- language detection, resistance to background interference, and adaptation to unknown synthesis methods.

Throughcategorizingreviewedworksaccordingto feature extraction, model architecture, and evaluation approach, the methodology ensured systematic comparison of various techniques. This facilitated identification of common advantages, constraints, and developing research directions in audio deepfake detection. The integration of findings from these categories provides the basis for the following sections of this publication.

VI. DISCUSSION

The examination of recent research demonstrates that audio deepfake detection has progressed substantially, yet multiple challenges remain. Traditionalmachinelearningapproachesincluding SVMs and decision trees perform adequately on smaller datasets but encounter difficulties with contemporary synthesis methods. Deep learning, especiallyCNNs,LSTMs,andhybridCNN-LSTM architectures, consistently provides superior accuracy by capturing both spectral and temporal characteristics. However, these models require substantialcomputationalresourcesandfrequently lacktransparency, constraining their application in real-time or forensic contexts.

A persistent issue is the restricted availability of extensive, varied datasets, which limits model adaptationacrosslanguages, accents, and evolving synthesis technologies. Another limitation is susceptibility to noise, reverberation, and compression, which reduces effectiveness in practical environments. Current research emphasizesthesignificanceofinterpretablemodels and efficient architectures for mobile implementation. While progress appears encouraging, practical deployment necessitates addressing dataset variety, efficiency, and transparency in detection frameworks.

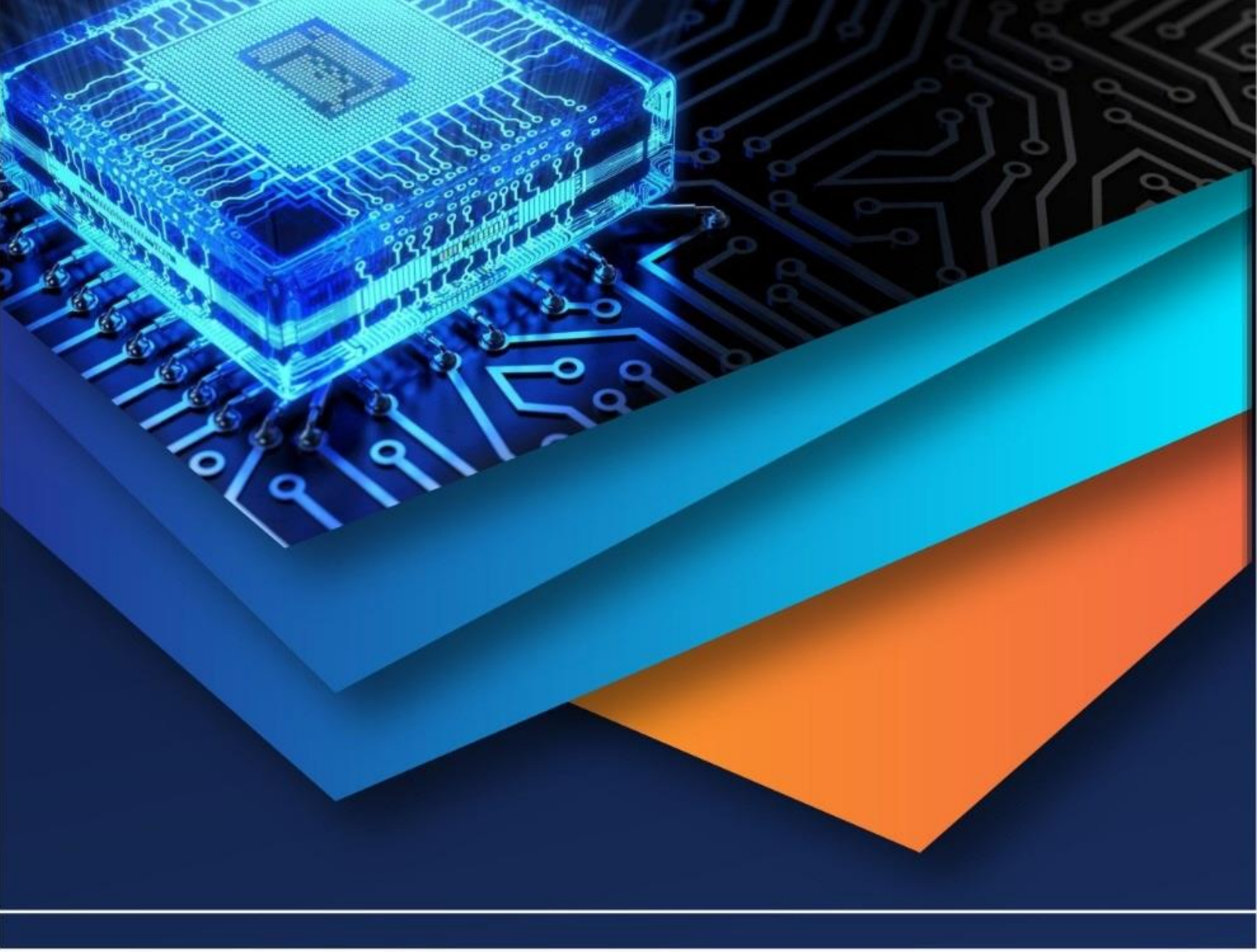
VII. CONCLUSION

The expansion of audio deepfakes constitutes both an advancement and a challenge in contemporary digital communication. While synthetic speech technologiesofferclearadvantagesinaccessibility and human-computer interaction, their malicious application in fraud, impersonation, and misinformation creates risks for individuals and communities. Investigation into audio deepfake detection has achieved significant advancement, with MFCC-based feature extraction and deep learning models reaching high detection precision in controlled conditions. Nevertheless, obstacles persistregardingdatasetdiversity,generalization capability, interpretability, and practicalrobustness.

The future development of this field depends on creating scalable, transparent, andflexible solutions that can maintain pace with the continuous evolutionofgenerativetechnologies.Achievement inthis domain will not only enhancecybersecurity and forensics but will also serve a crucial function inpreservingconfidenceindigitalcommunication. As deepfakes become progressively more advanced, the creation of dependable detection systems is no longer discretionarybut essential for protecting digital authenticity.

REFERENCES

[1] A. K. Singh and P. Singh, "Detection of AI- synthesized speech using cepstral bispectral statistics," IEEE MIPR, 2021.
[2] T. Arif et al., "Voice spoofing countermeasure forlogicalaccess attacks detection," IEEEAccess, 2021.
[3] D.Marietal.,"Thesoundofsilence:Efficiency of first digit features in synthetic audio detection," Elsevier Signal Processing, 2022.
[4] L. Cuccovillo et al., "Open challenges in synthetic speech detection," ACM Workshop, 2022.
[5] D. Salvi et al., "Synthetic speech detection through audio folding," ACM, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY