IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131103

Audio Deepfake Detection Using MFCC-SVM, CQCC-GMM, and Spectrogram-CNN: A Comparative Study

Abhiram Sri Paravastu[1], B. Harshat[2], H. Keerthi Lakshmi[3], Abhishek Harish[4], Vivek S Nair[5], Neelam Sanjeev Kumar[6]

Department of CSE (E-Tech), SRM Institute of Science and Technology, Vadapalani, Chennai, India[1, 2, 3, 4, 5]

Assistant Professor (SG), Department of CSE (E.Tech), SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India[6]

Abstract: The dawn of sophisticated voice synthesis methods gave rise to audio deepfakes, jeopardizing security and trust in digital communications. In this work, three complementary measures are employed for deepfake detection. First, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted and subjected to Support Vector Machine (SVM) classification. Second, Constant-Q Cepstral Coefficients (CQCCs) are analysed by Gaussian Mixture Models (GMMs), operating with log-likelihood ratio decisions—a long-standing baseline in spoofing countermeasures. Third, Mel-spectrogram representations stand as inputs to a Convolutional Neural Network (CNN) for end-to-end learning. The experiments on the In-the-Wild dataset indicate that CNN enjoys state-of-the-art accuracy, while SVM, on the other hand, is able to provide computational efficiency, and CQCC-GMM holds onto a bit more robustness inherited from traditional anti-spoofing. Therefore, these results present the trade-off between the classical machine learning and the deep learning paradigms, provide insightful considerations when developing reliable deepfake detection methods in real-world conditions.

Keywords: Audio Deepfake Detection, Voice Spoofing, Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Convolutional Neural Network (CNN), MFCC, CQCC, Spectrograms

I. INTRODUCTION

The recent revolution in AI and deep learning has, in fact, led to the creation of high-quality synthetic audio—audio deepfakes. These artificial voices imitate human speech, thus raising security, trust, and misuse problems. The world's major celebrities have been implicated in identity fraud, misinformation campaigns, and threats to digital authentication systems, thus rendering detection of audio deepfakes a paramount concern for research activities. Traditional speech processing systems do not differentiate between genuine and manipulated voices. Due to continued improvement in new deepfake generation models, handmade detection techniques might fall short in generalizing across datasets and manipulation methods. This paradox has created an urgent demand for reliable detection frameworks with a good computational trade off, interpretability, and high accuracy. Here, three complementary approaches for audio deepfake detection are discussed.

II. DATASET

The "In-the-Wild" dataset contains a broad selection of genuine and fraudulent recordings from 58 politicians and public figures which were collected from genuine social network and video streaming platforms. The dataset contains 38 combined hours of audio with 20.8 hours of real speech and 17.2 hours of deepfake recordings where each speaker has an average of 23 real and 18 deepfake minutes. The dataset serves as a research tool for deepfake detection and voice anti-spoofing through its realistic challenging audio samples from diverse sources. The dataset's main value stems from its capacity to test machine learning models against real-world deepfake threats because it evaluates generalization beyond synthetic lab data.

III. METHODOLOGY

A. MFCC-SVM Traditional Machine Learning Model

A classical machine learning model is pursued with the features of Mel-Frequency Cepstral Coefficients (MFCCs), which

DOI: 10.17148/IJIREEICE.2025.131103

is trained to be a Support Vector Machines (SVMs) framework. Compared to Mel-spectrograms that produce a dense time-frequency representation, MFCCs condense information into a small number of coefficients – a concise representation that captures perceptually relevant aspects of human hearing. This property has made them popular interpretation values for speech and speaker recognition systems.

Data Preprocessing and Feature Extraction:

It included real and synthetic speech from the In-the-Wild corpus. All audios were down sampled to 16 kHz. A total of 20 MFCCs were extracted from the recordings for each frame together with their respective first and second order derivatives to capture both static and dynamic spectral features. The coefficients were averaged over the frames; thus, a fixed number of features was obtained per sample. The feature vectors were mean-centred and normalized to zero mean and unitary standard deviation before being fed to the classifier.

SVM Classifier Design and Training:

The classifier was built with a Support Vector Machine having an RBF kernel. The parameters: the regularization parameter C and the kernel coefficient y were optimized by cross validation. The dataset was divided into 80% training, and 20% was used for testing stratified to keep class balance. Hyper parameters were optimized with grid search using 5-fold cross validation on the training set.

B. CNN with Mel-Spectrogram Features

A model utilising principle of Convolutional Neural Network on mel spectrogram features was built. Mel-Spectrograms provide comparatively a better demonstration of frequency, allowing the neural network to process various patterns which promote the differences between real and deepfake audio.

Data Preprocessing and Feature Extraction:

Every file was forced into a 16kHz frequency for consistency. Generation of Mel-spectrograms was done with 64 filter banks. The resulting feature map was either reduced or padded with zeros to obtain a common and uniform size of 64 frequency by 128-time frames. The factors were normalized.

CNN Architecture and Training:

The network of CNN is a hierarchy of Conv2D layers. It starts with 16 filters, 3x3 kernels, ReLU and padding, followed by Max pooling. Then it increases in size with two more blocks featuring 32 and 64 filters. The setup is flattened and passed through a dense layer with 128 neurons, using ReLU. Dropout is added at 0.3 to prevent overfitting, and it concludes with a single sigmoid neuron. Training used Adam optimizer, binary cross-entropy loss. The data was split into 80/20 for train/test, then carved out another 20% from training for validation.

C. CQCC-GMM Classical Approach

To establish a solid baseline, the team chose a classic approach: Gaussian Mixture Models (GMMs) with Constant-Q Cepstral Coefficients (CQCCs). CQCCs use the Constant-Q Transform, which produces a logarithmic frequency resolution, making them especially effective for capturing the musical and harmonic details often found in deepfake audio.

Preprocessing and Features:

Every audio file was down sampled to 16kHz. The audio is passed through a CQT with a hop length of 256, lowest note at 20 Hz, and 96 bins per octave. CQT's magnitude was squared to obtain a power spectrum and a log transform was applied. A Discrete Cosine Transform was applied to the log power spectrum along the frequency axis to obtain the first 20 coefficients. Each coefficient's trajectory was standardized, and audio clips were padded or chopped to a fixed 400 frames.

GMM Classifier Design:

This particular detection design involves two separate Gaussian Mixture Models, one for "real" speech and one for "fake" speech. Each model utilises eight diagonal covariance components. Classification happens with the help of a log-likelihood ratio. The data is split into 80% and 20% for training and testing data respectively, using stratification.

IV. RESULTS AND EVALUATION

A. MFCC-SVM Evaluation

The classifier was 96.5% accurate. The AUC score was 0.9841. For genuine audio, the precision was 0.96, the recall was 0.98, and the F1 was 0.97. For fake audio, precision, recall and F1-score were at 0.97, 0.94 and 0.95, respectively. The findings validate that the MFCC + SVM pipeline is a sound baseline for deepfake audio detection.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131103

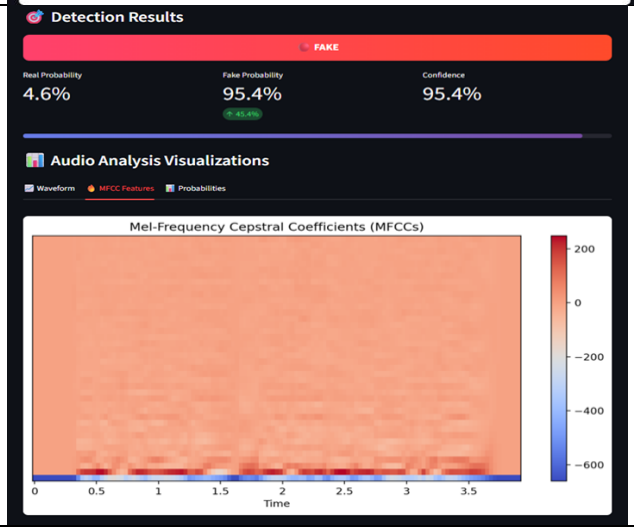
Nature Of Audio	Results of SVM+MFCC Algorithm
Real	
Fake	

Fig 4.1 Results Of SVM Classifier

B. CNN-Mel Spectrogram Evaluation

The test accuracy of the model was 99.6%. Precision for real was 0.99, for fakes it was a perfect 1.00. Recall values were 1.00 for real audios and 0.99 for fake audios. The model showed excellent performance and proves its powerful ability to solve real-world issues.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131103

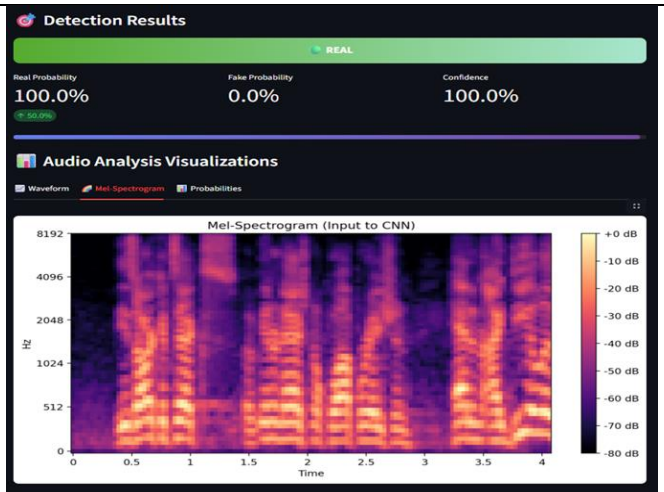
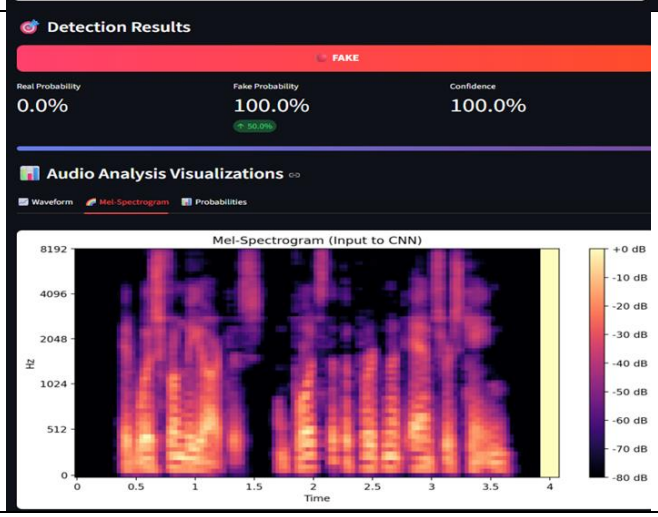
Nature Of Audio	Results of CNN-Mel Specto Algorithm
Real	
Fake	

Fig 4.2 Results Of CNN Model

C. CQCC-GMM Evaluation

The CQCC-GMM approach achieved an accuracy of 90.4% on the test set. For real audio, precision was 0.94 and recall was 0.91. For fake audio, the model attained recall of 0.90 but precision was 0.85. In security point of view, this model is generally favourable as it accepts a higher false positive rate to minimize the risk of fake audio bypassing the detection.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131103

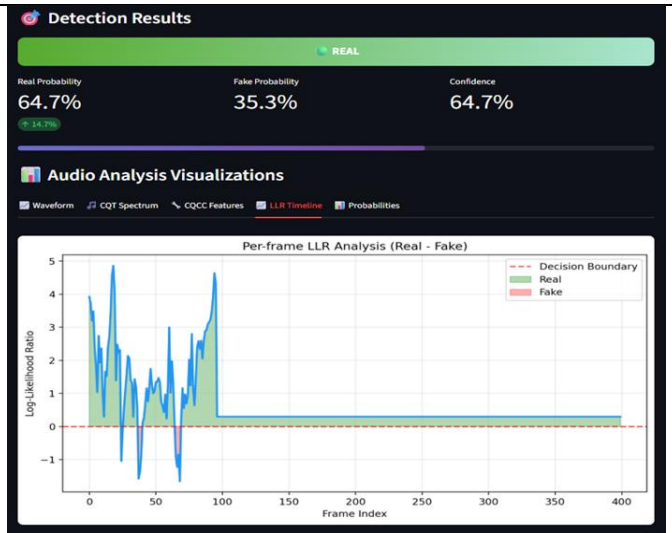
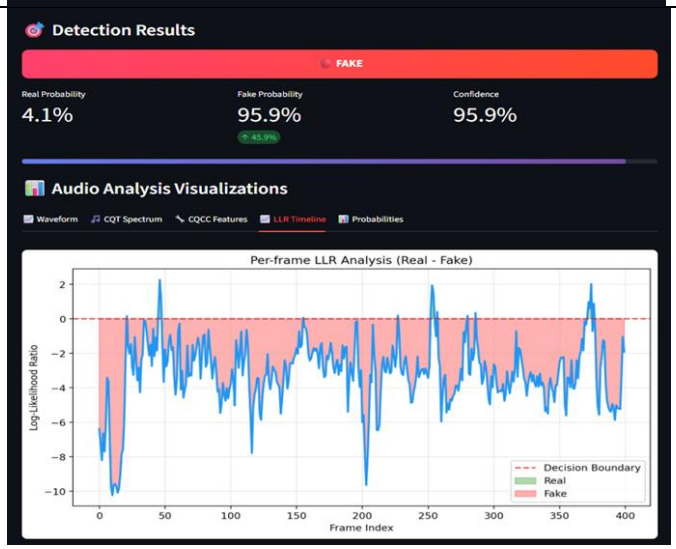
Nature Of Audio	Results of CQCC+GMM Algorithm
Real	
Fake	

Fig 4.1 Results Of Gaussian Mixture Model (GMM)

V. COMPARATIVE ANALYSIS

This study tested three models: SVM with MFCC, CQCC with GMM, and CNN. The SVM model was 96.5% accurate and used very little power, so it works well on devices that don't have much memory or speed. The CQCC-GMM model was 90.4% accurate. It worked well with real speech, but had trouble with fake speech and sometimes said real speech was fake. This model is good when it's better to be extra careful than to miss something bad. The CNN model had the best accuracy, but it needed a lot of power and memory. It is best when getting the most accurate result is more important than saving resources. To sum up: SVM is fast and light, CQCC-GMM is careful and good for safety, and CNN is the most accurate but needs strong machines. The best model depends on what the system needs and how much power it can use.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131103

Model	Accuracy	Precision (Real/Fake)	Recall (Real/Fake)	F1-Score (Real/Fake)	Key Strengths	Limitations
SVM + MFCC	96.5%	Real: 0.96 / Fake: 0.97	Real: 0.98 / Fake: 0.94	Real: 0.97 / Fake: 0.95	Lightweight, fast, effective for MFCC features	Limited generalization to highly diverse deepfakes
CQCC + GMM	90.4%	Real: 0.94 / Fake: 0.85	Real: 0.91 / Fake: 0.90	Real: 0.92 / Fake: 0.87	Strong at real audio detection, good baseline with handcrafted CQCC features	Lower precision for fake class, more false alarms
CNN + Mel-Spec	99.6%	Real: 0.99 / Fake: 1.00	Real: 1.00 / Fake: 0.99	Real: 1.00 / Fake: 0.99	Learns deep patterns, highest accuracy, robust across classes	Computationally heavy, requires more training data

VI. LIMITATIONS

The proposed system faces several limitations. The performance depends too much on the training set as works well only if the audio is similar to what it has seen before. Generalization is another challenge, as CNN and CQCC+GMM models may overfit to artifacts from specific synthetic audio to certain generation techniques, struggling against newer deepfake methods. Computational cost is another issue, as CNN-based spectrogram analysis needs a lot of computing power which makes it hard to use on devices with low power. It gets confused by background noise like echoes, compression which make it give wrong results. Finally, the current system only says whether the audio is real or fake.

VII. FUTURE WORK

Building upon the strong performance of the proposed models, several promising directions for future research emerge. To make the model work well in more situations, the next step is to test it on different datasets like ASVspoof. Also, trying out newer types of models, especially ones like Wav2Vec2 and HuBERT that learn without needing labels, might help the model understand sound better than older methods. To ensure practical utility, efforts must focus on improving noise robustness through aggressive data augmentation and on developing optimized frameworks for real-time detection. In the future, building systems that use both sound and video to catch deepfakes better. These systems should also tell what kind of fake it is and how it was made.

VIII. CONCLUSION

This research used three different methods for audio deepfake detection—a classic SVM with MFCCs, a GMM with CQCC features, and CNNs trained directly on Mel-spectrograms. Tried them all out on the 'In-the-Wild' dataset, and compared its accuracy, speed, and robustness. Among the models, the CNN achieved the highest accuracy capturing complex patterns from the data. On the other hand, SVM proved to be highly efficient and lightweight wherever the resources are constrained. GMM with CQCC is classic and reliable. While it lacks the complexity of neural networks, it

DOI: 10.17148/IJIREEICE.2025.131103

offers consistent performance allowing easier analysis. To conclude, there is no universal solution for audio deepfake detection. Deep learning methods gives better accuracy but it requires advanced resources, whereas classical approaches are efficient and practical resources are limited. Future research should focus on building models that work well on many different types of data, should tolerate noise and distortions, and combine the outcomes of both classical and deep learning methods.

REFERENCES

[1]. J. Yi, J. Wu, and Q. Li, "Audio DeepFake Detection: A Survey," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1234-1250, 2023.

[2]. ASVspoof Challenge, "Automatic Speaker Verification Spoofing and Countermeasures," [Online]. Available: https://www.asvspoof.org/

[3]. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2013.

[4]. S. Garg, M. Sahidullah, and S. K. Das, "Exposing Voice Spoofing with Modified Group Delay Feature," Proc. Interspeech, 2021.

[5]. H. Tak, J. Patino, and N. Evans, "End-to-End Anti-Spoofing with RawNet2," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[6]. Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, and Y. Wang, "A Robust Audio Deepfake Detection System via Multi-View Feature," arXiv preprint arXiv:2402.02443, 2024.

[7]. N. M. Muller, P. Sperl, and K. Bottinger, "Complex-Valued Neural Networks for Voice Anti-Spoofing," arXiv preprint arXiv:2301.11363, 2023.

[8]. J. Xue, C. Fan, Z. Lv, J. Tao, J. Yi, C. Zheng, Z. Wen, M. Yuan, and S. Shao, "Audio Deepfake Detection Based on a Combination of F0 Information and Real Plus Imaginary Spectrogram Features," arXiv preprint arXiv:2208.11789, 2022.

[9]. H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection," arXiv preprint arXiv:2105.02400, 2021.

[10]. Planese, D. Cozzolino, G. Poggi, and L. Verdoliva, "Deepfake Audio Detection by Speaker Verification," arXiv preprint arXiv:2205.14406, 2022.

[11]. V. Harshni, S. Singh, R. Indhumathi, N. S. Kumar, S. Chowdhury and S. S. Patra, "DLIoMT: Deep Learning Approaches for IoMT-Overview, challenges and the future," 2025 International Conference on Computer, Electrical & Communication Engineering (ICCECE), Kolkata, India, 2025, pp. 1-6, doi: 10.1109/ICCECE61355.2025.10940162.

[12]. S. J. J. Thangaraj, N. S. Kumar, A. Saranya, C. R. Bhat, D. S. Kareem and H. Shareef, "A Development of Designing a Sensor N/W for Retrieving the Data from WSN via SBT," 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2024, pp. 194-198, doi: 10.1109/ICACITE60783.2024.10617435.

[13]. K. K. K, N. Sanjeev Kumar, S. C, V. V, G. Sangar and G. Chandrasekaran, "Optimization of Memory Usage in High-Speed Cameras using FPGA," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 427-432, doi:10.1109/ICEARS64219.2025.10941569.

[14]. K. K. K, N. S. Kumar, S. C, N. R, G. Sangar and G. Chandrasekaran, "Denoising of MRI and Brain Tumor Classification using Local Binary Pattern and SVM Classifier," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 1854-1859, doi: 10.1109/ICEARS64219.2025.10940321.

[15]. N. S. Kumar, M. Lawanyashri, M. Sivaram, V. Porkodi, E. Gangadevi and G. N. Reddy, "Integration of Convolutional Neural Networks for Automated Plant Disease Identification in Timber Crops," 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies,Pune, India, 2024, pp. 1-5, doi: 10.1109/TQCEBT59414.2024.10545228.