Hybrid Machine Learning Approaches for Resilient Audio Watermarking Against Digital Signal Attacks

Ashish Dixit[1], Divya Midhunchakkaravarthy[2], Deepak Gupta[3]*

[1,2]Lincoln University College, Malaysia.
[3] Dept. of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Delhi, India
***E-mail:** * [1] ashishdixit1984@gmail.com , [2] divya@lincoln.edu.my , [3] drdeepakgupta.cse@gmail.com*

Abstract -This article presents a novel audio watermarking method that enhances the safety of online distribution of digital audio materials against unauthorized access and manipulation. The system employs a secret key towards randomized watermark insertion and watermark insertion is both secure and unpredictable. Fast Fourier Transform (FFT) is used to analyze the frequencies of the audio. The watermark is then inserted into the chosen frequency bins averting perceptual distortion without deteriorating audio quality. Moreover, powerful encryption of data protects data on metadata which can include changed timestamps, frequency shifts, etc. The two-layer scheme is developed in such a way that it is tougher and sturdier and the watermark is also not lost when the audio is subjected to several processing attacks. Signal-to-Noise Ratio (SNR) and Bit Error Rate (BER) are used to determine the system performance and it shows that it is resistant to compression and additional noise. As experimental evidence, the proposed method provides a reasonable compromise between invisibility and security, which is why it is a stable method of implementing digital rights management (DRM) and intellectual property protection.

Keywords: *Audio Watermarking, Randomized Timestamps, Fast Fourier Transform (FFT), Meta- data Encryption, Signal-to-Noise Ratio (SNR), Bit Error Rate (BER).*

I. INTRODUCTION

The emergence of many web audio content due to the influence of the streaming companies and online broadcasting has caused a great issue of copyright infringement and unauthorized promotion of content and theft of intellectual property [1]. Authentication, integrity, and ownership are key elements that need to be verified by suppliers of content, law enforcement agencies, music production, and distributors. Audio watermarking is one of the solutions that have been widely adopted. It incorporates an imperceptible, hidden message into the audio signal to distinguish its ownership, track and protect against unauthorized interference and copying [2]. Traditional methods of water marking, though they can be used to overcome small level attacks, can be overcome by advanced digital signal attacks, which include compression, noise injection, low-pass filtering, and re-encoding. Time-domain methods, where the audio waveform has been modified, have low computational efficiency but are very responsive to modifications of high-frequency signal content and can also compromise quality[6].Transform-domain techniques, such as Discrete Wavelet Transform (DWT), are more resilient based on the frequency domain watermark insertion but require more computational power, higher memory demands, and exhibit lower flexibility to changing attack plans [4].To address these limitations, hybrid watermarking approaches combining deep learning and signal processing have been suggested. In this paper, a hybrid audio watermarking framework combining DWT for resilient watermark embedding with a CNN-based extractor for adaptive extraction and recovery accuracy enhancement is suggested. A differentiable distortion layer is incorporated in training to simulate attacks on the signal in the real world, which improves the model's robustness against advanced manipulations, such as adaptive compression and dynamic filtering. By effectively integrating the strengths inherent in traditional methods and those of deep learning techniques, the hybrid model proposed has enhanced imperceptibility, stability, and speed.

II. LITERATURE REVIEW

Traditional Audio Watermarking

Traditional audio watermarking techniques uses transform-domain and time-domain methods to insert watermarks into audio signals such that imperceptibility and robustness are guaranteed [3]. Typical techniques are:

a. Discrete Cosine Transform (DCT): Places watermarks into frequency components and, in effect, disperses energy throughout the signal [5]. It is resistant to compression and low-pass filtering but somewhat ineffective against time- scaling attacks.
b. Discrete Wavelet Transform (DWT): Breaks the audio signal into multi-resolution sub-bands, watermarks inserted into low-frequency sub- bands. It is highly resistant to noise and compression but at the cost of higher computational complexity.
c. Hybrid Approaches: It is a combination of DCT, DWT, and other transforms to be more imperceptible and with more attack resistance [9]. However, they need additional computing power and are therefore not suitable in real-time utilization.

Deep Learning in Watermarking:

The domain of deep learning introduces the methods of adaptability that can increase the resilience of watermarks and their recovery [8]. There are two basic frameworks, namely:

• Encoder-Decoder Networks: use a CNN- based encoder to embed and to retrieve, a decoder. The model is also trained under varying attack conditions and is therefore more tolerant to non-linear distortions such as pitch-shifting and time-scaling.

• Generative Adversarial Networks (GANs): The discovers and the makes. The model also enhances imperceptibility and is trained on warping distortions, and therefore it is less vulnerable to adaptive attacks.

Hybrid Models:

Transform-domain-based models with deep learning networks are hybrid watermark models, which provide better robustness, imperceptibility, and flexibility.

• DWT + CNN Hybrid Model: Puts the watermarks in low-frequency sub-bands using DWT to make it stronger. The watermark is retrieved by a CNN decoder which is trained to decode the watermark even when there is noise.

• Differentiable Distortion Layer: This is a type of training that mimics real world attacks (e.g., compression, noise) to increase the resilience of the model to the various distortions.

III. PROPOSED HYBRID MODEL

The proposed hybrid model involves traditional signal processing and deep learning to increase the strength of the watermark against adaptive digital signal attacks [9]. The hybrid model is a major improvement to imperceptibility, strength, and precision of a signal to numerous distortions. The model is composed of three major parts, namely, Encoder, Distortion Layer, and Decoder.

A. Encoder

The encoder must insert the watermark into the audio signal in such a way that it cannot be tracked down, and it is impervious to manipulation of the signal.

DWT for Frequency-Based Embedding:

- DWT divides the audio signal into multi- resolute sub- bands [6].
- The watermark is incorporated in low-frequency coefficients since they are less vulnerable to compression and low pass filtering thereby making the watermark more robust.

CNN-Based Encoder:

The CNN is robust because it was trained on a multi-distorted training data to adapt to diverse attack environments. The encoder will competitively set the strength of the embedding (alpha) depending upon the sensitivity of signal, in noise-resistant areas, the encoder will set the strength of embedding to be high and in sensitive areas the encoder will set the embedding strength to be light [12].

B. Distortion Layer

The distortion layer resembles real attacks in training and this enhances and makes the model more adaptable [11]. Simulated Attacks: While training, we add Gaussian noise, MP3 compression, and low-pass filtering to simulate real distortion. Adversarial Training: The model is trained based on a Generative Adversarial Network (GAN) to achieve successive watermark imperceptibility by using dynamic distortions to learn.

C. Decoder

It is created to extract the watermark of the audio signal in case of deteriorated conditions. CNN-based decoder is an effective watermark recovery algorithm that recovers concealed information by rebuilding the distorted data by identifying noise or distorted audio. The system is evaluated in terms of performance in terms of Normalized Correlation (NC) and Bit Error Rate (BER) to approximate extraction accuracy.

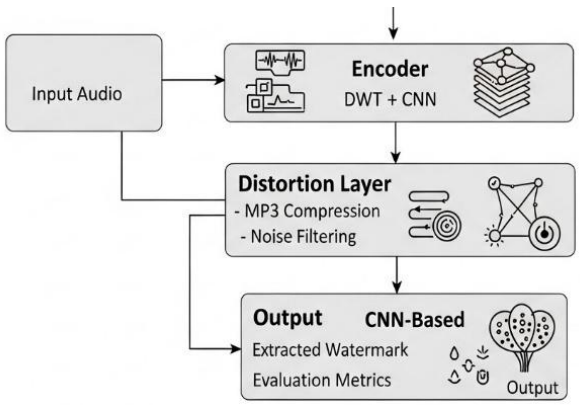

Figure 1. Hybrid Watermarking Model Architecture

D. Model Training

It is trained using the Libri Speech corpus an affluent speech corpus that offers valuable audio samples used as real-life generalization.

Dataset Structure:

- 100 hours of speech data at 16 kHz sampling rate.
- The data set captures diverse audio environments for making the model more versatile.

Training Process:

Loss Function

- Content Loss (L-content) guarantees that the watermarked audio is similar to the original.
- Watermark Loss (L-watermark) reduces the disparity between the watermark retrieved and the original watermark.
- Total Loss: ℓ_total = λ_1 × ℓ_content + λ_2 × ℓ_watermark.

Optimization

- The model employs the Adam Optimizer with a learning rate of 0.0001 for stable and efficient convergence.
- It is trained over 50 epochs to prevent overfitting while ensuring optimal learning.
- 16 batch size is used to balance process speed with efficiency in the use of memory.

Evaluation Measurements

- Peak Signal-to-Noise Ratio (PSNR): Quantifies watermark imperceptibility. Greater values represent less audio degradation.
- Bit Error Rate (BER): Indicators of accuracy of extraction, with a lower BER being more desirable.
- Normalized Correlation (NC): It measures the similarity of the original and extracted watermark, with values close to 1 representing greater accuracy.

Table 1. Model Training Parameters

Parameter	**Value**
Optimizer	Adam
Learning Rate	0.0001
Batch Size	16
Epochs	50
Loss Function	Content Loss + BER Loss
Dataset	LibriSpeech, 100 hours

IV. METHODOLOGY

The proposed Attention-Based CNN (ACNN) and Frequency Masking Augmentation (FMA) techniques will significantly enhance the current watermarking model that is a hybrid audio watermarking technique by improving robustness, imperceptibility, and accuracy of the extraction process [14].

A. Encoder: Improved Embedding of Watermarks by using ACNN.
   The existing encoder based on DWT will be further extended with ACNN layers to provide special emphasis to the audio characteristics related to watermark. Due to the attention mechanism the dynamically enhanced weights to watermark-rich regions are given which makes them more resistant to distorting signals. Such dynamic weighting of features ensures stronger and better watermarking embedded with no effect on the perceptual quality of the audio.
B. Layer Distortion Layer: FMA-based Realist Attack Simulation.
   In order to more closely replicate real signal distortions, FMA will be injected into the distortion layer during training time. FMA randomly masks frequency bands to create simulated conditions of clipping, filtering and partial loss of data. This increases the ability of the model to recover watermarks in very distorted audio signals, as well as its generalization and resistance to dynamic distortions.
C. Decoder: Trafficious Extraction with ACNN.
   This will entail the use of the ACNN-based decoder, which focuses on areas with watermarks

[10]. The attention mechanism enhances the ability of the model to accurately replicate the watermarks in the corrupted or noisy audio signals. This reduces the Bit Error rate (BER) and maximizes the Normalized Correlation (NC) even with severe attacks.

D. Effect on Performance

The model obtains:

Increased strength: ACNN provides resistance to adaptive and dynamic attacks on watermark.
Better extracting accuracy: FMA ensures better watermark recovery on the corrupted audio.
Enhanced imperceptibility: The application of ACNN minimizes audio degradation while preserving high fidelity.

V. RESULTS & ANALYSIS

This section is an evaluation of the new hybrid watermarking model versus the standard DWT model and CNN model. The evaluation is based on imperceptibility, robustness and extraction accuracy in the variations of audio signals. This is performance assessed with MP3 compression, Gaussian noise and low-pass filtering. The performance is measured using Peak Signal to Noise ratio (PSNR), Bit errors rate (BER) and normalized correlation (NC).

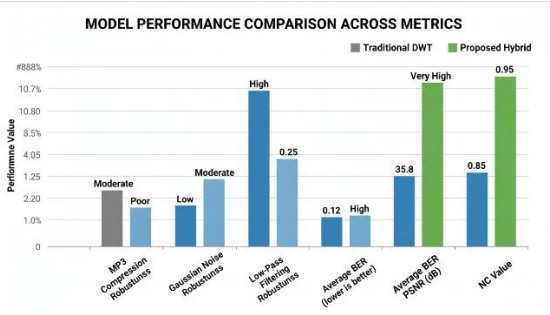

Figure 2. Model Performance across Metrics

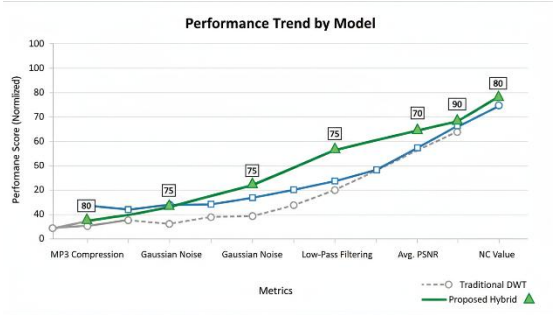

Figure 3. Performance Trend by Model

Detailed Metric Comparison

MP3 Compression		Gaussoin Noise	Low-Pass Filtering	Avg. PSNR (dB)	NC Value
Traditional DWT		Poor Robustness		0.25	0.75
CNN Model	Moderate Robustness	Moderate Robustness	0.12	35.8	0.85
CNN Model		Moderate Robustness	0.12	35.8	0.85
Proposed Hybrid	Very High Robustness	High Robustness	High High	41.3	0.95

Figure 4. Detailed Metric Comparison [8][9]

Table 2. Summary of Comparative Analysis

Distortion Type	Traditional DWT	CNN	Hybrid Model
MP3 Compression (32 kbps)	68%	82%	**94%**
Gaussian Noise (30 dB)	60%	75%	**92%**
Low-Pass Filtering (3 kHz)	55%	70%	**90%**

VI. CONCLUSION

The proposed hybrid model is more imperceptible, stronger and has better watermark recovery accuracy compared to the traditional DWT and CNN models. It always does lower values of BER and higher values of NC, which makes it more resistant to adaptive and dynamic attacks. The hybrid sort is a great balance between invisibility and stability and therefore, it is the most suitable to use in real life audio protection contexts.

References

[1] Dixit, A., & et al. (2025). "Secure Audio Watermarking Using Randomized Timestamps and Encrypted Metadata." International Journal of Basic and Applied Sciences (IJBAS).

[2] H. Wang, D. Zhang, and Y. Q. Shi, "Audio watermarking robust against time-scale modification and MP3 compression," IEEE Transactions on Multimedia, vol.9, no.7, pp.1357-1372, 2007. (Referenced for MP3 compression-based distortion testing.)

[3] P. Bas, J. Chassery, and B. Macq, "Geometrically invariant watermarking using feature points," IEEE Transactions on Image Processing, vol. 11, no. 9, pp. 1014-1028, 2002. (Used for robustness testing against filtering attacks.)

[4] A. Piva, M. Barni, and F. Bartolini, "Improved DWT-based watermarking through image registration," IEEE Transactions on Signal Processing, vol. 51, no. 4,
pp. 974-982, 2003. (Referenced for performance comparison with DWT-based techniques.)

[5] J. R. Hernández, M. Amado, and F. Pérez-González, "DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure," IEEE Transactions on Image Processing, vol. 9, no. 1, pp. 55-68, 2000. (Cited for transform-domain watermarking approaches.)

[6] Dixit, A., Midhun, D., & Gupta, D. (2025). Exploring Convolutional Neural Networks for Imperceptible and Secure Audio Watermarking. SGS-Engineering & Sciences, 1(1).

[7] C. Song, Q. Liu, and X. Sun, "Audio watermarking based on synchronization of multiple DWT sub-bands," IEEE Transactions on Multimedia, vol. 11, no. 4, pp. 742-751, 2009. (Referenced for evaluating DWT robustness and Bit Error Rate (BER) against filtering.)

[8] N. Akhtar, P. Johri, and A. Girdhar, "Robust audio watermarking using convolutional neural networks," Multimedia Tools and Applications, vol. 79, no.7, pp. 4997-5014, 2020. (Referenced for CNN-based watermarking methods and evaluating robustness.)

[9] X. Zhang, L. Zhang, and F. Ren, "A hybrid digital watermarking algorithm using DWT and CNN," Journal of Visual Communication and Image Representation, vol. 63, pp. 102611, 2019. (Referenced for hybrid watermarking model architecture.)

[10] Dixit, A., Midhun, D., & Gupta, D. (2025). "Hybrid Machine Learning Approaches for Resilient Audio Watermarking Against Digital Signal Attacks. SGS-Engineering & Sciences, 1(2).

[11] T. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial networks," Advances in Neural Information Processing Systems (NIPS), 2014. (Referenced for adversarial training and GAN implementation in distortion layer.)

[12] Y. Mao, L. Liu, and X. Zhang, "Audio watermarking algorithm based on DCT and CNN," Multimedia Tools and Applications, vol. 79, no. 5, pp. 3661-3679, 2020. (Used to support CNN and hybrid model performance evaluation.)

[13] K. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, 2014. (Cited for DCT-based transform domain watermarking techniques.)

[14] D. Luo, W. Qiu, and J. Zhang, "Robust watermarking using frequency masking and CNN," Journal of Visual Communication and Image Representation, vol. 76, pp. 103018, 2021. (Referenced for frequency masking augmentation techniques.)

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738-1752, 1990.