A Wavelet tour of Audio Deepfake Detection

Arth J. Shah*, Aniket Pandey*, Manav A. Gaikwad† and Hemant A. Patil*
* Speech Research Lab, Dhirubhai Ambani University (DAU)
E-mail: {202101154, 202411001, hemant_patil}@dau.ac.in
† MET BKC Institute of Engineering
E-mail: manavg.etc_ioe@bkc.met.edu

***Abstract*—With the rapid advancement of technology and Artificial Intelligence (AI), the misuse of AI-generated content has significantly increased. Since the onset of the COVID-19 pandemic, the proliferation of AI-generated fake (deepfake) videos, audio, and images has risen markedly. While considerable progress has been achieved in the domains of video and image deepfakes, the field of audio deepfakes remains relatively underexplored. Additionally, with the growing popularity of Machine Learning (ML) and Deep Learning (DL) methods, classical signal processing approaches—which emphasize capturing essential features from speech signals—are increasingly being overlooked in favor of large-scale data-driven models and transformer-based features. In this study, we address the Audio Deepfake Detection (ADD) task using classical wavelet-based signal processing techniques. Specifically, we employ six different types of wavelets—Bump, Morlet, Morse, Shannon, Mexican Hat, and Derivative of Gaussian (DoG)—to extract discriminative features for the ADD task. Among these, the Bump wavelet combined with a Convolutional Neural Network (CNN) classifier achieved the highest detection accuracy of 94.15 %. Furthermore, to assess the real-world applicability of the proposed approach, we also conducted a latency-based analysis.**

***Index Terms*—Wavelets, Bump Wavelet, CWT, ADD.**

I. Introduction

The COVID-19 pandemic (2019-2021) acted as a catalyst for accelerating the global shift toward Artificial Intelligence (AI) and Deep Learning (DL) technologies. As remote work, digital healthcare, and online education became necessities, there was a surge in demand for intelligent systems capable of automating tasks, analyzing large datasets, and enhancing decision-making. The servers experienced sudden exponential growth in data on internet. DL models, played a pivotal role in medical diagnostics, contact tracing, vaccine development, and virtual communication tools. As this period also witnessed increased investment in AI research and infrastructure, pushing industries and governments worldwide to integrate AI into critical services, marking a transformative step in the digital evolution of society. Alongside the rapid adoption of AI post-COVID, the proliferation of deepfake technology also saw a significant rise, bringing with it heightened security concerns. As AI models became more accessible and powerful, malicious actors began exploiting neural networks to create hyper-realistic fake audio, video, and images, often used for misinformation, fraud, and identity theft. The increase in digital communication and remote verification during the pandemic further exposed vulnerabilities, making it easier for attackers to deceive individuals and organizations. This boom in deepfakes led to a parallel growth in cybersecurity research, digital forensics, and AI-driven detection systems, as governments and tech companies scrambled to address the emerging threats in an increasingly virtual world.

In recent years, especially following the digital acceleration triggered by the COVID-19 pandemic, the world has gradually shifted from traditional signal processing techniques to AI, machine learning (ML), and DL approaches for both deepfake generation and detection. Classical methods, which relied heavily on handcrafted features and statistical models, began to fall short in handling the complexity and realism of modern synthetic media. AI and DL models, particularly convolutional and Recurrent Neural Networks (RNNs), offered a more powerful and scalable alternative by learning intricate patterns directly from large datasets. This shift marked a paradigm change in digital media forensics, where data-driven approaches now dominate both offensive and defensive strategies. While the shift to AI and DL has significantly advanced deepfake detection systems, it has also led to a growing restriction of classical signal processing methods, which still hold potential value. Many modern detection systems rely solely on data-driven models, often overlooking the rich insights that can be extracted from time-frequency analysis, phase inconsistencies, and other handcrafted signal features. Signal processing techniques offer interpretability, computational efficiency, and robustness in scenarios where data is limited or model generalization fails. Ignoring these methods entirely may result in systems that are more vulnerable to adversarial attacks or domain shifts. In this study, we explore the interaction between classical signal processing and modern machine learning by leveraging wavelet-based feature extraction for Audio Deepfake Detection (ADD). The wavelet based features have been explored partially before in [1]. We in this study, specifically employ 6 diverse set of wavelets (namely, bump, morlet, morse, Derivative of Gaussian (DoG), Mexican hat, and shannon) of wavelet families to capture both transient and stationary characteristics of speech signals across multiple resolutions. These wavelet transforms enable robust time-frequency localization, which is crucial for identifying subtle artifacts introduced by synthetic audio generation. The extracted features are then integrated into various DL models, including convolutional neural networks (CNNs), and GoogLeNET, to evaluate their effectiveness in detecting deepfake audio. Proposed hybrid approach demonstrates the continued relevance of wavelet theory in complementing data-

979-8-3315-7206-8/25/$31.00

driven methods, offering improved generalization and interpretability for ADD task.

A. Related works

Many studies have been exploring wavelets for ADD task, however they fail to propose a system that employs both signal processing and DL models at a same time. In [2], authors employed multiresolution wavelet decomposition of Mel-spectrograms combined with computer vision architectures (Adversarially Robust WaveletCNN (ARWaveletCNN)) to obtain Equal Error Rate (EER) of 4.8 % on the ASVspoof2019 LA dataset. In [3], authors employed a Wavelet Prompt Tuning based method within a self-supervised learning framework, and combine them with XLSR-AASIST (SOTA method) to obtain EER of 3.58 %. Other studies, such as [4], operatos on frequency domain using Discrete Fourier Transform (DFT), followed by Discrete Wavelet Transform (DWT) and use Haar wavelets to obtain accuracy of 92.03 %. More recent study [5], employed one-level Harr wavelet decomposition, and fed them to VGG19 as a backbone classifier to obtain 79.76 % AUC. We in this study, employed the simmilar two stage method (feature extraction on stage 1 and feature classification on stage 2) inorder to analyze the difference between fake and real signals.

II. Proposed Methodology

Wavelet is a small wave that is localized in time-frequency domain, which is renounced to analyze non-stationary signals, such as speech, EEG, etc. In this study, we analyze different types of wavelets on basis of the properties they capture to detect the difference between fake and real audios. A wavelet is a mathematical condition which has finite energy ($\int |\psi(t)|^2 dt = 0$), and zero mean (i.e., positive and negative parts $\int_{-\infty}^{\infty} \psi(t)dt = 0$ cancel out each other), where $\psi(t)$ is defined as wavelet. In this work, we investigate different types of wavelets—both real-valued and complex (with real and imaginary components)—based on the signal characteristics they emphasize. Real wavelets are often used for energy localization, while complex wavelets offer improved directional selectivity and phase information. This analysis aims to identify the most suitable wavelet features for differentiating between genuine and manipulated audio signals. For a signal $x(t)$ *w.r.t.* chosen mother wavelet $\psi(t)$ (CWT) is given by:

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t).\psi^*_{a,b}(t)dt, \tag{1}$$

where,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{t-b}{a}\right). \tag{2}$$

Here $\psi_{a,b}(t)$ represents the scaling and shifting of the mother wavelet (a = scaling parameter, and b = shifting parameter), and $*$ represents the complex conjugate. Note: for real wavelets, the complex conjugate operator has no effect as the complex conjugate of a real-valued function is the function itself.

To extract features using specific types of wavelets, we computed scalograms, which visualize the energy distribution of a signal across the time-scale domain. A scalogram is obtained by taking the squared magnitude of the Continuous Wavelet Transform (CWT) coefficients. It effectively captures how signal energy varies over time and across different frequency scales, which makes it useful for identifying transient patterns and discriminative features in audio signals. The scalogram is formally defined as:

$$Scalogram(a,b) = |W_x(a,b)|^2. \tag{3}$$

For this particular study, we analyzed 6 different types of wavelets (among which 3 are real wavelets, and 3 are imaginary wavelets), which are briefed as follow.

A. Bump Wavelet

The bump wavelet is a smooth, infinitely differentiable wavelet with compact support in the frequency domain, which is used to analyze extreme smoothness of a signal. It appears as a single smooth rounded bump that smoothly rises and falls with no side lobes or oscillators [6]. Bump wavelet is employed due to its property to ideally analyze smooth signals (as it is infinitely differentiable, which leads to no sharp cutoff and prevents distortion), band limited, better frequency localization, and customization bandwidth. As a complex valued wavelet, bump wavelet captures the strength (magnitude), and timing (phase) of oscillations, and oscillatory patterns that detects slowly varying oscillations in the signals, such as harmonic components in audio, and frequency modes in data. Bump wavelet also captures the smooth frequency response ensuring clean detection of frequency content without ripple. Mathematically bump wavelet in frequency domain is defined as:

$$\Psi(\omega) = \begin{cases} \exp\left(-\frac{1}{1-\left(\frac{\omega-\mu}{\sigma}\right)^2}\right), & \text{if } |\omega-\mu| < \sigma \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where, $\Psi(\omega)$ is Fourier Transform (FT) of bump mother wavelet, μ is center frequency of bump, and σ is half-width of bump. Performing Inverse Fourier transform (IFT) on $\Psi(\omega)$ of Eq. (4), we get $\psi_{bw}(t)$,

$$\psi_{bw} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi(\omega) * e^{i\omega t} d\omega. \tag{5}$$

Performing scaling and shifting (as defined in Eq. (1), and Eq. (2)) on Eq. (5), to obtain the scalograms and writing CWT, we get:

$$W_x(a,b) = \frac{1}{2\pi}\sqrt{|a|}\int_{-\infty}^{\infty} x(t).\Psi(\omega)e^{i\omega(t-b)}d\omega. \tag{6}$$

B. Derivative of Gaussian (DoG) Wavelet

It is a wavelet obtained by differentiating a gaussian function, which enhances edges and changes in the signal by emphasizing rapid transitions. The order of derivation of the gaussian function also determines different properties and

gives various features. For DoG feature extraction, the order was chosen to be 1, i.e.,

$$\psi(u) = \frac{d^m}{du^m} e^{-\frac{u^2}{2}}, \tag{7}$$

where $u = \frac{t-b}{a}$, and for DoG features, m=1 (first order differentiation). Performing scaling and shifting, we obtain [7]:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}[\frac{d^m}{dt^m} e^{-\frac{(t-b)^2}{2a^2}}] = \frac{1}{\sqrt{a}}[\frac{1}{a^m}.\frac{d^m}{du^m}.e^{-\frac{u^2}{2}}], \tag{8}$$

where $\frac{1}{\sqrt{a}}$ is normalization to keep energy constant when scaling, and $\frac{1}{a^m}$ comes from the charm rule for changing variable when differentiating.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}.a^m}\frac{d^m}{du^m}.e^{-\frac{u^2}{2}}, u = \frac{t-b}{a}. \tag{9}$$

Performing CWT on Eq. (9), as defined in Eq. (1), we get:

$$W_{a,b} = \int_{\infty}^{\infty} x(t)[\frac{1}{\sqrt{a}.a^m}\frac{d^m}{du^m}.e^{-\frac{u^2}{2}}]. \tag{10}$$

DoG is a real wavelet, which captures the strength of signal, excellent time resolution (ideal for determining transitions and edges), zero-mean, and edge detection, which are useful in analyzing vision and signal properties.

C. Mexican Hat Wavelet

Mexican Hat Wavelet, also known as Ricker wavelet, was named after its shape which resembles the shape of mexican hat (typically features a wide, flat brim and a high, pointed crown). It is the second derivative of gaussian (m = 2 in Eq. (9)) function, and is widely used for detecting edges and singularities of signal. Mexican hat captures the second order features, such as peak, edges, and singularities in the signal. It has good time localization (as shape of the wavelet is central positive peak with symmetric negative side lobes, acting like a filter in time domain), which helps gunpoint abrupt changes. More-over it is simple and real valued wavelet, which further simplifies computation, interpretation, and visualization of the signal. Also two negetive lobes of mexican hat wavelet, cancel out slow trends in speech signal. The mexican hat wavelet can be defined as [7]:

$$\psi(t) = (1-t^2) * e^{-\frac{t^2}{2}}. \tag{11}$$

Scaling equation of a wavelet can be defined by,

$$\psi_a(t) = \frac{1}{\sqrt{a}}\psi(\frac{t}{a}), \tag{12}$$

and substituting value of Eq. (11) in Eq. (12), we obtain,

$$\psi_a(t) = \frac{1}{\sqrt{a}}\left(1-\left(\frac{t}{a}\right)^2\right).e^{-\frac{1}{2}.(\frac{t}{a})^2}, \tag{13}$$

after performing shifting on Eq. (13), we obtain,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\left(1-\left(\frac{t-b}{a}\right)^2\right).e^{-\frac{1}{2}(\frac{t-b}{a})^2}. \tag{14}$$

By substituting obtained Eq. (14) in Eq. (1) and solving it, we get:

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t)\cdot\frac{1}{\sqrt{a}}(1-u^2)e^{-\frac{u^2}{2}}\,dt, \quad \text{where } u = \frac{t-b}{a}. \tag{15}$$

D. Shannon Wavelet

The Shannon wavelet is defined using ideal bandpass filter. Its' frequency domain is a perfect rectangle, which is conceptually simple, but do not compactly support in time domain. In time domain, it looks like a *sinc* function , i.e., an infinite central peak with decaying slide ripple that oscillates forever. The central peak of the wavelet is responsible for capturing frequency band perfectly. The side ripple shows that the wavelet is ideally band limited but is localized poorly in time domain. This makes it perfect to extract frequency representation, however shannon wavelet is less precise for localizing sudden changes in time domain. The shannon wavelet come from an orthogonal wavelet basis, which is useful for reconstruction in wavelet transform, and helps to capture amplitude and phase of a complex valued wavelet. The mathematical representation of shannon wavelet can be described as [8]:

$$\psi(t) = sinc(t) - \frac{1}{2}sinc(\frac{t}{2}), \tag{16}$$

where $sinc(t) = \frac{sin(\pi t)}{\pi t}$. Applying scaling and shifting property on Eq. (16), and substuting value of $\psi(t)$ in Eq. (2), we get,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\left[sinc(u) - \frac{1}{2}sinc(\frac{u}{2})\right], u = \frac{t-b}{a}. \tag{17}$$

Substituting Eq. (17) in Eq. (1), gives us equation of CWT for shannon wavelet and can be calculated as,

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t)\left[\frac{1}{\sqrt{a}}\left(sinc\left(\frac{t-b}{a}\right) - \frac{1}{2}sinc\left(\frac{t-b}{2a}\right)\right)\right]. \tag{18}$$

E. Morlet Wavelet

The, most famous wavelet *w.r.t.* the historical development of wavelet research, is the morlet wavelet, which is modulated Gaussian, and it is defined as [6]:

$$\psi(t) = \frac{1}{\sqrt[4]{\pi}}e^{j\omega_0 t} * e^{-t^2/2}, \tag{19}$$

where ω_0 is taken as 5Hz for a standard morlet wavelet. The morlet wavelet is combination of gaussian envelop and sinusoidal wave, which are useful for analyzing oscillatory and frequency varying signal. The shape of morlet wavelet resembles a wave like oscillation enveloped smoothly by a bell shaped gaussian. The oscillatory wave is responsible for detecting periodic components in a signal, and the gaussian envelope localize the wave in time, ensuring that it analyzes a small time region without spreading too much. The excellent frequency resolution, which is ideal for capturing oscillations in the signals, and the bell shape envelope (gaussian) localize

wavelet in time domain, helps to capture time-varying content. Also the real and imaginary part allows the wavelet to measure both phase and amplitude of oscillations. Applying scaling and shifting on Eq. (19) gives us:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \left[\frac{1}{\sqrt[4]{\pi}} . e^{i\omega_0(\frac{t-b}{a})} . e^{-\frac{1}{2}(\frac{t-b}{a})^2} \right], \quad (20)$$

and by substituting Eq. (20) in Eq. (1), we get:

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t) . \frac{1}{\sqrt{a}} \left[\frac{1}{\sqrt[4]{\pi}} . e^{i\omega_0(\frac{t-b}{a})} . e^{-\frac{1}{2}(\frac{t-b}{a})^2} \right]. \quad (21)$$

F. Morse Wavelet

Morse wavelet, or more precisely the Generalized Morse Wavelet, is an exactly analytic wavelet, which means its Fourier transform is supported only on positive frequencies. This complex-valued wavelet is especially well-suited for analyzing modulated signals with varying amplitude and frequency, as well as detecting localized discontinuities in time. The generalized Morse wavelet is characterized by two main parameters: β, which controls the compactness (decay), and γ, which governs the symmetry of the wavelet. By tuning these parameters, a wide range of wavelet shapes can be obtained, making Morse wavelet a flexible and unified model that subsumes several other wavelets as special cases, such as the Cauchy wavelet ($\gamma = 1$) and Bessel-like wavelets ($\beta = 8, \gamma = 0.25$).

The generalized Morse wavelet in the frequency domain is defined as [6]:

$$\Psi_{\beta,\gamma}(\omega) = U(\omega) \cdot a_{\beta,\gamma} \cdot \omega^{\beta} e^{-\omega^{\gamma}}, \quad (22)$$

where $U(\omega)$ is the unit step function ensuring analyticity, and $a_{\beta,\gamma}$ is a normalization constant to ensure unit energy. Applying inverse Fourier transform and using the scaling and shifting properties, the time-domain representation becomes:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right), \quad (23)$$

and the Continuous Wavelet Transform (CWT) using Morse wavelet is given by:

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t) \cdot \frac{1}{\sqrt{a}} \psi^* \left(\frac{t-b}{a} \right) dt, \quad (24)$$

where $\psi^*(\cdot)$ denotes the complex conjugate of the Morse wavelet. Due to its analyticity and parameter flexibility, Morse wavelets can measure both instantaneous frequency and amplitude, making them suitable for applications in speech, biomedical, and deepfake signal analysis.

III. Experimental Setup

A. Classifier Used

We employed two architectures for classification: a simple Convolutional Neural Networks (CNN) and a custom Inception-based model (poppularly known as GoogLeNet). The CNN consists of two convolutional layers with ReLU activation and max-pooling, followed by a fully connected layer for final prediction. It provides a lightweight and efficient baseline for feature extraction. To improve feature diversity, an Inception-style network was designed. Each Inception block includes four parallel branches: 1×1, $1 \times 1 \rightarrow 3 \times 3$, $1 \times 1 \rightarrow 5 \times 5$ convolutions, and 3×3 max-pooling followed by 1×1 convolution. These branches capture multi-scale features effectively. The classifier uses two such Inception blocks after an initial convolution and pooling layer, followed by global average pooling and a final dense layer for classification.

B. Dataset Description

For this study, we utilized the *Fake or Real Norm (FoR-Norm)* dataset, a widely used benchmark for audio deepfake detection. The dataset consists of both genuine (real) speech and synthetic (fake) speech generated using various text-to-speech (TTS) and voice conversion (VC) techniques. All audio samples are standardized through normalization to ensure consistent amplitude and format across recordings, facilitating fair model evaluation. The dataset is balanced across the two classes and includes multiple speakers, which helps in training models that generalize well across different voice characteristics. This makes the Fake or Real Norm dataset particularly suitable for binary classification tasks involving the detection of spoofed or manipulated audio content.

IV. Experimental Results

To assess the role of different wavelet transforms in audio-based deepfake detection, we conducted a comprehensive evaluation using six wavelet types: bump, morlet, morse, derivative of Gaussian (DoG), Mexican Hat, and Shannon. The classification performance was assessed using several binary metrics, including accuracy, precision, recall, F1 score (macro), Hamming loss, AUC, EER, Matthews Correlation Coefficient (MCC), Cohen's Kappa, balanced accuracy, and Jaccard index.

Among all, the **bump wavelet** demonstrated the highest classification performance across all major metrics, achieving an accuracy of 94.15%, an AUC of 0.9865, and the lowest EER of 0.0586. This superior performance can be attributed to the excellent frequency localization properties of the bump wavelet. Its compact support and smoothness enable it to capture subtle high-frequency variations in the spectro-temporal representation of speech signals — characteristics often distorted in deepfakes due to synthesis artifacts or vocoder limitations. Additionally, the bump wavelet's strong time-frequency resolution allows it to preserve both formant structures and transient noise details, which are key discriminative cues in real vs. synthetic speech.

The **morlet** and **morse wavelets** also yielded reasonably strong results, with accuracies of 86.69% and 87.98% respectively, and F1 scores above 0.86. These analytic wavelets are known for their ability to provide good balance between time and frequency resolution. The morse wavelet, in particular, showed slightly better recall and AUC, indicating higher sensitivity to detecting fake samples. However, both wavelets underperformed compared to bump, possibly due to their fixed frequency profiles, which may limit their adaptability to the

TABLE I
RESULTS OBTAINED ON VARIOUS WAVELETS FOR CNN CLASSIFIER.

	Testing Accuracy	Precision	Recall	F1 Score (Macro)	Hamming Loss	AUC	EER	MCC	Cohen's Kappa	Balanced Accuracy	Jaccard Index
bump	**94.15**	0.9361	0.9448	0.9415	0.0585	0.9865	0.0586	0.883	0.883	0.9416	0.8876
morlet	86.69	0.8453	0.8905	0.8669	0.1331	0.941	0.1283	0.7349	0.7339	0.8674	0.7657
morse	87.98	0.8609	0.8993	0.8798	0.1202	0.9512	0.116	0.7604	0.7597	0.8802	0.7852
DoG	83.04	0.7989	0.8723	0.8303	0.1696	0.9178	0.1603	0.664	0.6613	0.8313	0.7153
Mex_h	73.28	0.7609	0.6608	0.7308	0.2672	0.8098	0.2608	0.4677	0.4638	0.7312	0.5472
shannon	75.6	0.8414	0.6167	0.7501	0.244	0.865	0.2203	0.5273	0.5088	0.7529	0.5525

varied spectral distortion patterns observed in deepfake audio generated by different synthesis models.

On the other hand, the **DoG** and **Mexican Hat** wavelets, which are real-valued and symmetric, performed significantly worse. The DoG wavelet yielded an accuracy of 83.04% and a notably lower MCC of 0.664. The Mexican Hat wavelet, with the lowest accuracy of 73.28%, exhibited poor recall (0.6608) and a high Hamming loss (0.2672), indicating a high rate of misclassifications. These wavelets, due to their poor frequency localization and lack of phase information, are less suited for capturing the nuanced spectral dynamics of speech signals. Furthermore, their inability to isolate non-stationary high-frequency components results in missed detection of synthesized perturbations typical in deepfakes.

The **Shannon wavelet**, which is inherently discontinuous and defined in the frequency domain, also performed sub-optimally, especially in terms of recall (0.6167). Despite its relatively high precision (0.8414), it exhibited a high EER (0.2203) and moderate F1 score, suggesting that the model trained on Shannon-based features tended to miss a substantial number of fake samples. This can be logically explained by the Shannon wavelet's poor temporal localization and inability to effectively track transient variations, which are often critical in detecting speech anomalies introduced by generative models.

Overall, wavelets with high smoothness, strong frequency localization, and analytic nature — such as bump and morse — provided superior discriminative features for deepfake detection. Their ability to retain phase information and emphasize subtle harmonic inconsistencies inherent in synthesized speech makes them more robust for binary classification in this domain. The results suggest that wavelet design plays a crucial role in determining the effectiveness of time-frequency representations for speech-based deepfake detection.

A. Analysis of Latency Period

To investigate how varying the input signal size—interpreted here as latency period—affects the performance of deepfake detection, we evaluated classification accuracy across five frame sizes: 512, 256, 128, 64, and 32. Table II shows the resulting accuracies for each wavelet at different input sizes.

TABLE II
ACCURACY (IN %) OF MODEL ON CNN ON VARIOUS FEATURE SIZE

Size	Shannon	Morse	Morlet	Mex_h	DoG	Bump
512 x 512	78.43	80.41	79.63	77.99	85.46	85.09
256 x 256	75.89	85.33	81.66	84.51	85.30	88.30
128 x 128	75.60	87.98	86.69	73.28	83.04	94.15
64 x 64	76.49	90.25	87.59	77.91	81.94	92.77
32 x 32	68.98	85.48	84.92	71.01	85.13	93.41

The results highlight that detection performance generally improves as the input size is reduced from 512 to 128 samples, reaching peak performance at or near 128 for most wavelets, particularly **bump** and **morse**. The bump wavelet reaches its highest accuracy (94.15%) at size 128, indicating that medium-sized windows strike a balance between temporal and spectral resolution. This suggests that such input lengths effectively capture discriminative artifacts introduced by deepfake generation, such as inconsistent pitch modulations or synthetic noise bursts. Notably, the bump wavelet maintains high accuracy even at lower sizes (92.77% at 64 and 93.41% at 32), demonstrating its robustness in time-localized feature extraction. The morse and morlet wavelets also follow a similar trend, showing improved accuracy from 512 to 64. This indicates that analytic wavelets benefit from shorter latency periods, likely due to better alignment with the transient nature of speech signal alterations in deepfake audio. In contrast, wavelets such as Shannon and Mexican Hat (Mex_h) show more instability across sizes, with Shannon dropping significantly to 68.98% at size 32. This is likely due to their discontinuous or poor frequency localization properties, which become more pronounced as window size reduces, resulting in loss of meaningful spectral information. Interestingly, the **DoG** wavelet exhibits relatively stable performance across all input sizes, consistently staying within the 81–85% range. This may indicate that while it is not the most discriminative wavelet, it is less sensitive to changes in window size, perhaps due to its symmetric and scale-invariant nature. Also the results obtained on GoogLeNet classifier can be obtained on [1]. Input sizes around 128–64 samples are particularly effective when used with smooth,

TABLE III
COMPARISON WITH EXISTING WORKS ON DIFFERENT DATASETS

Study	Dataset	Feature Set	Classifier	Accuracy (in %)	EER (in %)
[9]	Baidu Silicon Valley AI Lab dataset	mfcc	VGG16	85.906	-
[10]	FoR	Scatter plot	CNN	88.9	11
[11]	H-voices	handcrafted features	deep4snet	98.5	-
[12]	ASVSpoof 2019	ELTP-LFCC	BDBiLSTM	-	33.28
[13]	FoR	Melspectrogram	TCN & STN	92,80	-
[14]	ASVSpoof 2019	LFCC	Resnet34	-	5.32
[15]	H-voices	histograms	Deep4SNET	98.50%	-
[16]	In-the-Wild	raw audio	XLSR	-	7.46
[17]	four benchmark	raw audio	SafeEar's	-	2.02
[18]	ASVSpoof 2019 & 2021	raw audio	BTS-E	-	8.11
[19]	In-the-Wild	handcrafted features	XLS-R, WavLM, and Hubert	-	24.27
Proposed	**FoR**	**Bump Wavelet**	CNN	**94.15**	**5.86**

[1]"https://acesse.one/2D6JE"

high-resolution wavelets like bump and morse. This insight is valuable for designing real-time deepfake detection systems where latency and accuracy must be optimally balanced. Table III compares the proposed methodology, with other existing works on different datasets.

V. Summary and Conclusions

This study investigates the application of classical signal processing via wavelet-based feature extraction for the ADD task. These time-frequency representations were then classified using a CNN and a lightweight Inception-based architecture. The FoR-Norm dataset was used for benchmarking, offering a balanced mix of genuine and fake speech generated using various TTS and VC systems. Among all wavelets, the Bump wavelet demonstrated the highest performance, achieving 94.15 % accuracy, 0.9865 AUC, and the lowest EER of 0.0586. A latency-based analysis showed that input sizes of 128 or 64 samples optimize performance for wavelets like Bump and Morse. This work highlights the interpretability and computational efficiency of wavelet-based features, especially in resource-constrained or real-time scenarios. The limitations of the proposed methodology includes the comparison of the obtained results with State-Of-The-Art (SOTA) models such as AASIST, which also remains as future work along with feature fusion based approaches.

References

[1] A. J. Shah and H. A. Patil, "Significance of lower frequency regions for audio deepfake detection," in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024, Macau, China, pp. 1–6.

[2] A. Fathan, J. Alam, and W. Kang, "Multiresolution decomposition analysis via wavelet transforms for audio deepfake detection," in *International Conference on Speech and Computer (SPECOM)*, 2022, Gurugram, India, pp. 188–200.

[3] Y. Xie, R. Fu, Z. Wang, *et al.*, "Detect all-type deepfake audio: Wavelet prompt tuning for enhanced auditory perception," *arXiv preprint arXiv:2504.06753*, 2025, {Last Accessed: 21^{st} June, 2025}.

[4] A. Dutta, A. K. Das, R. Naskar, and R. S. Chakraborty, "WaveDIF: Wavelet sub-band based deepfake identification in frequency domain," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, Nashville , USA, pp. 6312–6321.

[5] S. Sadhya and X. Qi, "Detection of deepfakes using wavelet-based convolutional neural network," in *International Conference on Pattern Recognition (ICPR)*, 2024, Kolkata, India, pp. 170–184.

[6] P. Gupta and H. A. Patil, "Morse wavelet transform-based features for voice liveness detection," *Computer Speech & Language*, vol. 84, p. 101 571, 2024.

[7] T. Le-Tien, "Some issues of wavelet functions for instantaneous frequency extraction in speech signals," in *IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications*, vol. 1, 1997, pp. 31–34.

[8] K. Paliwal, B. Shannon, J. Lyons, and K. Wójcicki, "Speech-signal-based frequency warping," *IEEE signal processing letters*, vol. 16, no. 4, pp. 319–322, 2009.

[9] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," *Procedia Computer Science*, vol. 219, pp. 211–219, 2023.

[10] S. Camacho, D. M. Ballesteros, and D. Renza, "Fake speech recognition using deep learning," in *Applied Computer Sciences in Engineering: 8^{th} Workshop on Engineering Applications, WEA 2021, Medellín, Colombia, October 6–8, 2021, Proceedings 8*, Springer, pp. 38–48.

[11] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4SNet: Deep learning for fake speech classification," *Expert Systems with Applications*, vol. 184, p. 115 465, 2021.

[12] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice spoofing countermeasure for logical access attacks detection," *IEEE Access*, vol. 9, pp. 162 857–162 868, 2021.

[13] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering*, vol. 47, no. 3, pp. 1–12, 2021.

[14] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual learning for fake audio detection," *INTERSPEECH*, pp. 886–890, 2021, Brno, Czechia.

[15] M. Rabhi, S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," *Expert Systems with Applications*, vol. 250, p. 123 941, 2024.

[16] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised xls-r and sls classifier," in *ACM Multimedia*, 2024, Melbourne Australia, pp. 6765–6773.

[17] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "Safeear: Content privacy-preserving audio deepfake detection," *Proceedings on ACM SIGSAC Conference on Computer and Communications Security (CCS), Salt Lake City, UT, USA*, pp. 3585–3599, 2024.

[18] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "Bts-e: Audio deepfake detection using breathing-talking-silence encoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, Rhodes Island, pp. 1–5.

[19] Y. Yang, H. Qin, H. Zhou, *et al.*, "A robust audio deepfake detection system via multi-view feature," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, Seoul, Korea, pp. 13 131–13 135.