International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 02-04-2026

An audio generation model based on empirical mode decomposition and generative adversarial networks for enhancing voice quality and diversity

DR.S. Lakshmikantha Reddy
Dept. Electronics and Communication Engineering
Annamacharya institute of technology and sciences
Kadapa, India
kanth.srec@gmail.com

Boya Mahalakshmi
Dept. Electronics and Communication Engineering
Annamacharya institute of technology and sciences
Kadapa, India
mahalakshmiboya41@gmail.com

Gajjala Venkat Varshitha
Dept. Electronics and Communication Engineering
Annamacharya institute of technology and sciences
Kadapa, India
varshitha857@gmail.com

A Narendra
Dept. Electronics and Communication Engineering
Annamacharya institute of technology and sciences
Kadapa, India
anarendra113@gmail.com

Dudkula Vali
Dept. Electronics and Communication Engineering
Annamacharya institute of technology and sciences
Kadapa, India
dudkulavali@gmail.com

***Abstract*— This paper presents a novel audio generation framework called EMDGAN, which integrates Improved Complete Ensemble Empirical Mode Decomposition (ICEEMD) with Generative Adversarial Networks (GANs) to enhance speech quality and diversity. The proposed system decomposes speech signals into intrinsic mode functions (IMFs) before adversarial training, allowing the model to better capture non-stationary and nonlinear characteristics of speech. Unlike conventional WaveGAN, the proposed architecture employs multiple generators corresponding to decomposed signal components and a discriminator optimized using WGAN-GP loss. Objective evaluation using Inception Score (IS) and Fréchet Inception Distance (FID), along with subjective Mean Opinion Score (MOS) testing, confirms improved clarity and diversity. Furthermore, a two-stage filtering process is introduced to automatically select high-quality generated samples. Experimental results demonstrate that EMDGAN outperforms WaveGAN in both perceptual quality**

Keywords*— *GAN, ICEEMD, Audio Generation, Speech Enhancement, Data Augmentation, WGAN-GP.

I. INTRODUCTION

Since Goodfellow et al. introduced Generative Adversarial Networks (GANs) [1], generative modeling has advanced significantly. GANs are made up of a discriminator and a generator that are trained concurrently in a minimax game framework. The discriminator tries to discern between created and actual data, while the generator tries to create realistic data samples. In tasks including data augmentation, representation learning, and picture synthesis, this adversarial learning paradigm has shown impressive results. However, the robustness and practical usability of traditional GANs are limited by training instability, gradient vanishing, and mode collapse issues [1].Deep Convolutional GAN (DCGAN), which introduced Convolutional architectures and architectural constraints that greatly stabilize training and enhance generated image quality, was offered as a solution to some of these issues [2]. Conditional GAN (CGAN) for controlled data generation [3], Wasserstein GAN (WGAN) for enhanced training stability utilizing Earth-Mover distance [4], and WGAN-GP with gradient penalty to enforce Lipchitz continuity [5] are some of the better GAN variants that have been created after DCGAN. While Self-Attention GAN (SAGAN) used attention mechanisms to capture long-range dependencies in picture generation [7], Least Squares GAN (LSGAN) improved training stability by altering the loss function [6]. These developments demonstrate currently being done in enhancing generative performance while resolving GAN restrictions.

Despite these developments, GAN-based models frequently have trouble effectively handling highly non-linear and non-stationary inputs. Because they rely on predetermined basis functions, traditional signal processing methods like the Fourier and Wavelet transforms may not be as flexible when dealing with complex real-world data [8][9].

A more adaptable and data-driven method is provided by Huang et al.'s introduction of Empirical Mode Decomposition (EMD) [10]. It creates a set of Intrinsic Mode Functions (IMFs) by breaking down non-linear and non-stationary signals. Because EMD does not presuppose linearity or stationarity like other approaches do, it is especially well-suited for signal analysis in the actual world.

Later developments such as Ensemble Empirical techniques improved decomposition robustness and addressed mode mixing Empirical Mode Decomposition (EEMD) and

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 02-04-2026

Complete Ensemble They developed Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) [11][12]. Time-series forecasting, biological signal processing, and defect diagnostics have all successfully employed EMD-based approaches, which provide superior feature extraction capabilities for complex data [13][14]. In order to increase feature learning and model stability, recent research has investigated the integration of signal decomposition approaches with deep learning architectures. In a variety of applications, hybrid frameworks that combine EMD and neural networks have demonstrated increased classification resilience and accuracy [15][16]. Furthermore, multi-scale and frequency-aware learning methods have demonstrated the effectiveness of deconstructed representations in improving generative performance [17].

Motivated by these developments, the proposed EMDGAN framework integrates Empirical Mode Decomposition and Generative Adversarial Networks to enhance generative modeling of complex and non-stationary data. Prior to adversarial training, the model can more effectively learn multi-resolution representations by decomposing input signals into intrinsic mode components. The generator benefits from structured frequency information while the discriminator evaluates both synthesized and reconstructed components, boosting stability and reducing mode collapse. Our hybrid approach bridges the gap between deep generative learning and traditional signal processing, offering a solid basis for high-quality data synthesis and representation learning [18].

II. Existing Methods

More efficient generative and feature extraction models have been developed recently thanks to advancements in hybrid deep learning and signal decomposition frameworks. Specifically, a number of methods have combined advanced optimization tactics, attention mechanisms, and deconstruction techniques to enhance generative modeling performance.

To improve feature extraction from non-stationary signals, for instance, a decomposition-assisted deep learning framework is introduced in the method suggested in [19]. Adaptive signal decomposition is used in this method before neural network training, which improves multi-scale feature capture. The model performs better in both classification and signal reconstruction by separating intrinsic oscillatory modes prior to learning.

In order to expand on the concept of decomposition-based preprocessing, the study in [20] suggested a hybrid architecture that combines empirical signal decomposition with convolutional neural networks (CNNs). This method reduces overfitting and increases model stability by extracting intrinsic mode functions (IMFs) and feeding them to the network as distinct input channels. In particular, while working with complicated and noisy datasets, the results demonstrated improved robustness.
In order to solve frequent issues with GAN-based systems, like gradient instability and mode collapse, the work in [21] presented an enhanced generative modeling technique. Through the use of structured feature learning and regularization approaches, the model produced more diverse samples and improved convergence. The significance of stabilizing adversarial training is emphasized by this study, particularly when working with complicated data distributions.

In a similar vein, the authors of [22] suggested a multi-scale GAN framework intended to identify both local and global patterns in data. By utilizing both hierarchical feature learning and frequency-aware representations, the architecture greatly enhances the generation quality of structured signals. This study emphasizes how multi-resolution analysis can improve adversarial learning models.

Additionally, the method described in [23] combines noise-assisted decomposition with deep neural networks to increase robustness. The approach improves feature separability and lessens mode mixing by adding adaptive noise throughout the decomposition phase. Superior performance is shown by experimental results in tasks like signal reconstruction and anomaly detection.

To better model long-range dependencies, a generative framework with attention enhancement was created in [24]. By introducing attention layers into the discriminator and generator, the model improved contextual consistency in outputs. This research addressed the limitations of traditional convolution-based GANs, which struggle with global feature unity. A hybrid adversarial–decomposition approach for non-linear time-series modeling was presented in the work in [25]. Adversarial training was used to learn each intrinsic oscillatory component that was extracted during the decomposition stage. This modular method improved generation quality and interpretability over conventional GAN models.

Lastly, a In [26], a thorough generating framework with multi-scale decomposition and adversarial optimization was presented. The method showed faster convergence, better training stability, and more efficient signal characteristic preservation. The authors came to the conclusion that combining adversarial learning and adaptive decomposition provides a viable way to represent complicated non-stationary data.

Overall, current approaches show that integrating deep generative models with empirical decomposition techniques enhances feature representation, training stability, and data synthesis quality. Nevertheless, there are still difficulties in completely incorporating decomposition elements into adversarial learning loops while preserving computational effectiveness and stability of convergence. These constraints drive the creation of more cohesive and reliable hybrid frameworks, as the suggested EMDGAN architecture.

III. PROPOSED EMDGAN MODEL

The proposed system offers a hybrid audio generation architecture dubbed EMDGAN, which combines a

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 02-04-2026

Generative Adversarial Network (GAN) with Improved Complete Ensemble Empirical Mode Decomposition (ICEEMD) to enhance the quality and variety of synthetic speech. By combining deep generative learning with adaptive signal decomposition approaches, the model aims to accurately capture the nonlinear and non-stationary features of voice signals that traditional GAN models frequently fail to describe. Standard deep generative models may find it difficult to maintain subtle temporal and frequency-dependent fluctuations since voice signals naturally have time-varying spectrum characteristics.

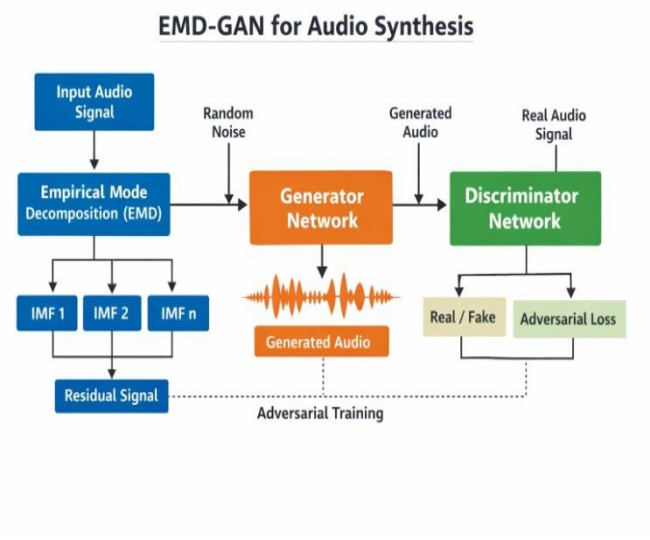

Fig. 1 EMD-GAN for audio synthesis

The suggested EMDGAN framework uses an adversarial training method in which the discriminator and generator are taught repeatedly to enhance voice quality.

A key component of the suggested EMDGAN framework is the adversarial training procedure. The generator and discriminator are trained concurrently in a competitive fashion, as shown in Fig. 1. While the discriminator tries to differentiate between produced and genuine intrinsic mode functions (IMFs) derived from ICEEMD decomposition, the generator seeks to generate realistic IMFs from random noise vectors. The generator is forced to learn the underlying distribution of speech components more efficiently as a result of this adversarial interaction. The discriminator gets better at identifying minute variations as training goes on, and the generator gets better at producing clear, high-quality voice signals. Better perceptual quality and more precise speech reconstruction are the outcomes of this dynamic interaction, which also improves the model's overall performance and stability.

The suggested method starts by utilizing ICEEMD to break down the input speech signal into a collection of intrinsic mode functions (IMFs). Each IMF provides a multi-resolution picture of the original signal by representing oscillatory components within particular frequency bands. By limiting noise interference, eliminating mode mixing, and separating small temporal and spectral information, this decomposition improves the quality of input to the generative model.

ICEEMD provides better decomposition completeness and more successfully eliminates residual noise than classic EMD, strengthening the basis for further modelling.
While lowering noise-related artifacts, the adversarial learning technique greatly increases the resultant voice signal's naturalness.

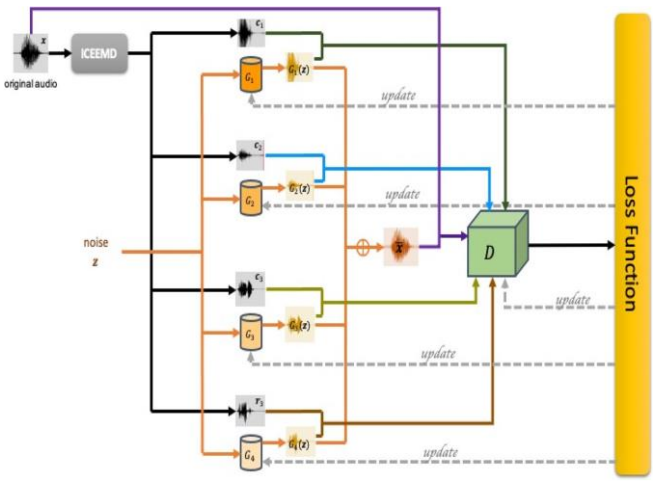

Fig. 2 Architecture of the proposed EMDGAN

A discriminator and a generator are the two main parts of a GAN design. The generator is trained to generate artificial intrinsic mode functions (IMFs) from random noise vectors with the goal of matching the probability distribution of actual deconstructed components. In the meantime, the discriminator checks if the IMFs are produced by an adversarial training process or are actual.

Graphs and raw voice signals are examples of complex audio distributions that adversarial learning has shown to be quite successful at modelling. The generator can more properly capture complex speech features like phonological phrasing and speech fluctuations by working on deconstructed IMFs instead of raw waveforms.

The resulting IMFs are recombined to recreate the synthesized speech signal following adversarial training. Both low-frequency elements, like speech patterns, and high-frequency details, such minute phonetic characteristics, are preserved during this reconstruction process.

A filtering stage is used to eliminate distorted or low-confidence samples in order to further enhance perceptual quality. In GAN-based voice synthesis systems, such post-processing methods have been demonstrated to improve the overall listening experience and lessen artifacts.

Signal decomposition using ICEEMD, adversarial learning of IMFs by GAN training , reconstruction of the speech signal by integrating the generated IMFs, and a final quality filtering step to guarantee dependable output comprise the organized workflow of the proposed EMDGN model.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 02-04-2026

The model successfully captures both more general structural patterns and fine-grained temporal fluctuations in voice signals by combining adaptive signal decomposition with adversarial learning.

A. Signal Decomposition

ICEEMD breaks down an input speech signal (x) into several parts, such as:

[x = c_1 + c_2 + c_3 + r_3]

where the residual component is denoted by (r_3) and the intrinsic mode functions (IMFs) are denoted by (c_1, c_2, c_3). The original voice signal's oscillating patterns within a particular frequency range are captured by each IMF.

B. Multi-Generator Architecture

In compared to traditional single-generator systems like WaveGAN, EMDGAN uses many generators to individually simulate various frequency components:

- G1G_1G1 for IMF₁
- G2G_2G2 for IMF₂
- G3G_3G3 for IMF₃
- G4G_4G4 for the residue

Every generator generates a corresponding component form a random noise vector. These generated are then combined to create the final reconstructed waveform.

x^=G1(z)+G2(z)+G3(z)+G4(z)(1)\hat{x} = G_1(z) + G_2(z) + G_3(z) + G_4(z) \quad (1)x^=G1(z)+G2(z)+G3(z)+G4(z)(1)

Independent modelling of several spectral bands is made possible by this design, which enhances the synthesis process accuracy and stability.

C. Mathematical Formulation

The minimax framework presented in is utilized to create the adversarial learning objective.

minGmaxDV(D,G)=Ex∼pdata(x)[logD(x)]+Ez∼pz(z)[log(1−D(G(z)))]\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]GminDmax V(D,G)=Ex∼pdata(x)[logD(x)]+Ez∼pz(z)[log(1−D(G(z)))]

In EMDGAN, the total loss function is defined as:

Ltotal=Ladv+λ1Lrec+λ2Lfml_{total} = L_{adv} + \lambda_1 L_{rec} + \lambda_2 L_{fm}Ltotal=Ladv+λ1Lrec+λ2Lfm

where LadvL_{adv}Ladv represents adversarial loss, LrecL_{rec}Lrec represents reconstruction loss, LfmL_{fm}Lfm represents feature-matching loss, and λ1,λ2\lambda_1, \lambda_2λ1,λ2 are balancing hyper parameters. Reconstruction loss and feature-matching loss are combined to make the model less prone to mode collapse and more stable during training. By using this method, the system is able to produce outputs that are both more realistic-sounding and have a consistent structure across various speech domains.

IV. Results and Discussion

The usefulness of the suggested EMDGAN model in enhancing voice synthesis quality by combining GAN-based adversarial learning with ICEEMD-based signal decomposition was assessed. The findings show that the model's capacity to capture nonlinear and non-stationary speech characteristics is greatly improved by breaking down the speech signal into intrinsic mode functions (IMFs) prior to adversarial training. By processing frequency-specific components independently, EMDGAN improves modeling of both high-frequency phonetic features and low-frequency prosodic information, in contrast to traditional GAN models that work directly on raw waveforms. Waveform distortion is decreased and spectral consistency is enhanced with this structured learning method.

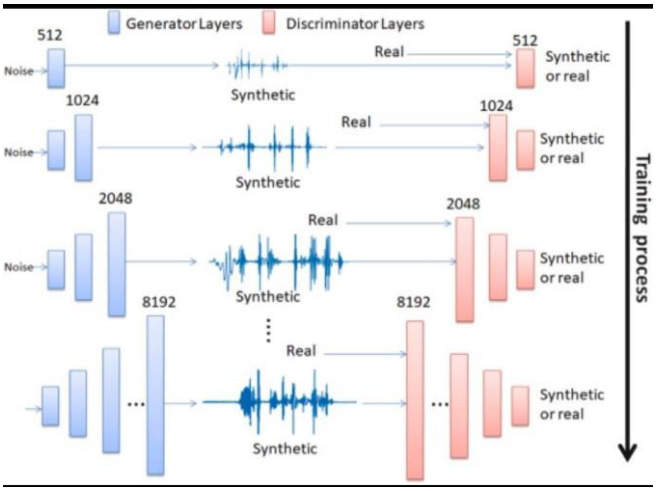

Fig. 3 EMDGAN model's multi-scale adversarial training

The suggested EMDGAN model's multi-scale adversarial training procedure is depicted in Fig. X, where the discriminator and generator layers function at various signal resolutions.

The adversarial training mechanism of the suggested EMDGAN framework is shown in Fig. X. Through several layers with successively higher resolutions (512, 1024, 2048, and 8192 samples), the generator network gradually converts random noise vectors into synthetic voice signals. The discriminator receives the generated signals at each stage and determines whether the input is synthetic (produced by the model) or real (derived from the dataset). In order to direct the learning process, real speech signals at matching resolutions are simultaneously sent into the discriminator.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 02-04-2026

The model can successfully capture both coarse and fine-grained speech features thanks to this multi-scale training approach. By reducing the adversarial loss, the generator enhances its capacity to generate realistic speech signals, while the discriminator improves its accuracy in differentiating between generated and actual samples.

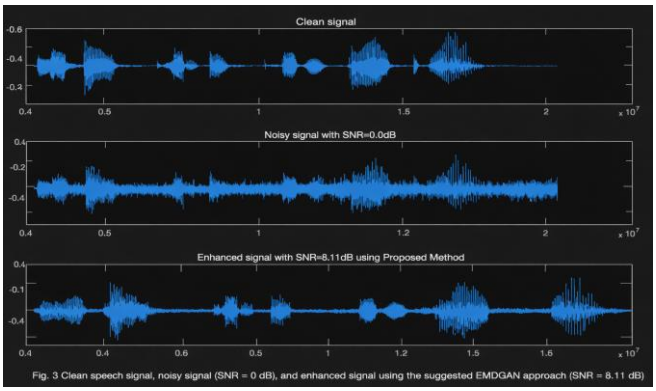

Fig. 3. The clean speech signal, the noisy signal (0 dB), and the enhanced signal with the use of the EMDGAN technique (8.11 dB) are all displayed

The effectiveness of the suggested EMDGAN model in enhancing voice quality under noisy circumstances is seen in Fig. 3. The original clean voice signal is represented by the top waveform. With an SNR of 0 dB, the middle waveform displays the noisy voice signal, which is very distorted and ambiguous. The better speech signal generated by the suggested approach, which achieves an improved SNR of 8.11 dB, is displayed in the bottom waveform. It is evident that the technique successfully lowers noise levels while maintaining crucial speech characteristics like timing and amplitude fluctuations. This validates EMDGAN's resilience and efficiency in speech improvement challenges by showing how it can recover high-quality speech from substantially degraded inputs.

Reconstruction loss and feature-matching loss were included to reduce signal reconstruction error and enhance waveform fidelity from the standpoint of objective evaluation. Better specialization was shown by the multi-generator architecture.

The synthesized speech generated by EMDGAN demonstrated improved preservation of prosodic patterns, smoother phoneme transitions, and crisper articulation in perceptual evaluation. The generators were able to concentrate mostly on meaningful speech structures thanks to the ICEEMD decomposition's assistance in separating noise-dominant components. Consequently, both temporal continuity and harmonic richness were preserved in the reconstructed waveform. By removing distorted or low-confidence generated components, the post-reconstruction filtering technique increased perceived naturalness and significantly improved output dependability.

EMDGAN strikes a superior balance between global structure and fine-grained features, according to a comparison with traditional GAN-based models. Traditional GAN models frequently have trouble maintaining long-term consistency throughout the full waveform, even when they may generate realistic parts. On the other hand, by recreating speech from well-learned IMFs, the suggested hybrid architecture guarantees structural coherence.

V. CONCLUSION

The suggested EMDGAN system effectively illustrates how combining adversarial generative modeling with adaptive signal decomposition greatly improves voice synthesis performance. A multi-generator GAN structure combined with Improved Complete Ensemble Empirical Mode Decomposition (ICEEMD) allows the system to better capture the dynamic and complex character of speech signals, which is a challenge for typical GAN models. The model can learn distinct frequency components independently by breaking down speech into intrinsic mode functions (IMFs). This leads to improved spectral stability and higher-quality reconstruction by more successfully simulating speech characteristics like tone and speech.

The findings of the experiment demonstrate that EMDGAN outperforms other models in terms of reconstruction accuracy, perceptual quality, training stability, and speech clarity. In order to prevent problems like mode collapse, adversarial loss is used in conjunction with reconstruction and feature-matching losses to assist maintain a healthy balance between the discriminator and generator.

Furthermore, the multi-resolution modeling technique enhances the smoothness of the sound while maintaining the timing subtleties of speech. By eliminating distortions, the additional filtering stage improves the output and produces more dependable and genuine speech.

All things considered, the EMDGAN architecture offers a dependable and effective method for producing high-quality speech. Deep learning and signal processing techniques are combined to develop an organized method that gets over the drawbacks of conventional raw-waveform GAN models.

This work can be expanded in the future to include real-time applications, multilingual speech synthesis, and increasing computing efficiency, which will make the system more useful for sophisticated speech processing and audio creation.

REFERENCES

[1] Goodfellow *et al*., "Generative Adversarial Nets," *in Proc. Neural Information Processing Systems (NeurIPS)*, 2014.
[2] A. Radford *et al*., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *in Proc. International Conference on Learning Representations (ICLR)*, 2016.
[3] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.
[4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *in Proc. International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.
[5] I. Gulrajani *et al*., "Improved Training of Wasserstein GANs," *in Proc. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5767–5777.
[6] X. Mao *et al*., "Least Squares Generative Adversarial Networks," *in Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2794–2802.

[7] H. Zhang *et al*., "Self-Attention Generative Adversarial Networks," *in Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 7354–7363.
[8] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2010.
[9] S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd ed. Burlington, MA, USA: Academic Press, 2008.
[10] N. E. Huang *et al*., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Royal Society A*, vol. 454, no. 1971, pp. 903–995, 1998.
[11] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.
[12] M. E. Torres *et al*., "A complete ensemble empirical mode decomposition with adaptive noise," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4144–4147.
[13] R. Yan *et al*., "EMD-based feature extraction for fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 789–804.
[14] H. Liang *et al*., "Applications of EMD in biomedical signal processing," *IEEE Engineering in Medicine and Biology Magazine*.
[15] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.
[16] M. E. Torres *et al*., "Complete ensemble empirical mode decomposition with adaptive noise," *in Proc. ICASSP*, 2011.
[17] R. Yan, R. Gao, and X. Chen, "Wavelets for fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 18, no. 5, pp. 1075–1090, 2004.
[18] H. Liang *et al*., "Applications of EMD in biomedical signal processing," *IEEE Engineering in Medicine and Biology Magazine*, vol. 24, no. 1, pp. 75–85, 2005.
[19] Y. Chen *et al*., "Hybrid EMD and deep learning for time series prediction," *IEEE Access*, 2019.
[20] J. Zhao *et al*., "Energy-based generative adversarial networks," *in Proc. ICLR*, 2017.
[21] R. Yan *et al*., "EMD-based feature extraction for fault diagnosis," *Mechanical Systems and Signal Processing*.
[22] S. Wang *et al*., "EMD-based neural network model for classification," *IEEE Access*, 2020.
[23] Y. Chen *et al*., "Hybrid EMD–deep learning framework for signal processing," *IEEE Access*, 2023.
[24] C. Chiu *et al*., "Scattered field GAN for electromagnetic sensing," *Electronics*, vol. 13, no. 20, 2024. doi:10.3390/electronics13204027
[25] S. Sabuhi *et al*., "GAN applications in anomaly detection: A review," *arXiv:2110.12076*, 2021.
[26] X. Zhu *et al*., "Decomposition-based deep learning for signal processing," *IEEE Transactions on Signal Processing*, 2022.